THE UNIVERSITY OF SYDNEY

PH.D. THESIS

# Deep Generative Modeling for Chest X-ray Interpretation and Synthesis


*Author:*
Ling YANG

*Suprvisor:*
A/Prof. Luping ZHOU
*Co-Supervisor:*
A/Prof. Wanli OUYANG


*A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy*

*in the*


School of Electrical and Information Engineering
Faculty of Engineering


January 12, 2026

# Authorship Attribution Statement

Chapter 3 of this thesis is published as,

**Ling Yang**, Zhenghao Chen, Kaisiyuan Wang, Luping Zhou*, “Improving CXR Bone Suppression by Exploiting Domain-level and Instance-level Information", *IEEE Transactions on Medical Imaging*.

Chapter 4 of this thesis is published as,

**Ling Yang**, Zhanyu Wang, Zhenghao Chen, Xinyu Liang, Luping Zhou*, “Medxchat: A Unified Multimodal Large Language Model Framework Towards CXRS Understanding and Generation", *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*.

Chapter 5 of this thesis is published as, **Ling Yang**, Xinyu Liang, Zhanyu Wang, Ziyu Diao, Xuan Huang, Die Shen, Xin Tan, Haifeng Li, Zhenghao Chen, Shijun Qiu*, Luping Zhou*, “Constructing A Unified Vision-Language Model for Chest X-Ray Diagnosis, Medical Education, and Data Augmentation",*Radiology: Cardiothoracic Imaging*.

Chapter 6 of this thesis is published as,

**Ling Yang**, Zhanyu Wang, Luping Zhou*, “Medvisiochat: A Multimodal Large Language Model Framework for Interpretable Diagnosis with Visual Grounding in CXRs", *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*.

In all these publications, I proposed the methods and wrote the paper, and my primary supervisor, Associate Professor Luping Zhou, is the co-corresponding author.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Student Name: Ling Yang

Signed: _____

Date: January 12, 2026

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Luping Zhou

Signed: _____

Date: January 12, 2026

Abstract of thesis entitled

# Deep Generative Modeling for Chest X-ray Interpretation and Synthesis

Submitted by

**Ling YANG**

for the degree of Doctor of Philosophy

at The University of Sydney

in January, 2026

Multimodal Generative Modeling has significantly advanced chest x-rays (CXRs) interpretation and synthesis, but key challenges remain—such as bone suppression task for CXRs diagnosis, unifying interpretive and generative tasks, expert radiologists' evaluation for unified medical model and grounding visual features effectively for diagnosis. To tackle these issues, we leverage generative models such as Vector-Quantized Generative Adversarial Network (VQGAN) and Stable Diffusion, along with large language models like mPLUG-Owl and Qwen-VL, to present four novel contributions: (i) a bone suppression framework that improves disease diagnosis by reducing the visual interference of ribs in CXRs; (ii) a unified large language model (LLM) that supports report generation, visual question answering (VQA), and image synthesis—offering an end-to-end solution for comprehensive CXRs understanding; (iii) a comprehensive evaluation combining computational metrics and radiologists' assessments for medical LLMs. and (iv) an instruction-tuned multimodal model for accurate disease classification and visual grounding.

**Bone Suppression for Enhanced Disease Diagnosiss:** CXRs often suffer from overlapping bony structures such as ribs and clavicles, which can obscure important soft tissue abnormalities like pulmonary nodules. To enhance disease visibility and improve diagnostic accuracy, we formulate bone suppression as a crucial image synthesis task. We propose

a novel dual-domain translation framework that leverages both domain-level and instance-level information. Specifically, we introduce a Multi-head Codebook Attention (MCA) module and a Cross-Covariance Attention Block (CAB) network to generate bone-suppressed images while preserving anatomical fidelity. To validate the effectiveness of our approach, we conduct downstream classification and segmentation tasks, demonstrating consistent performance gains. Furthermore, Grad-CAM-based visualizations highlight improved disease localization, showcasing the practical benefits of bone suppression for interpretability and clinical relevance.

**Unified Large Language Model for CXR Understanding and Generation:** Recent LLMs such as LLaMA and Qwen-VL typically support only text and image inputs while generating text-only outputs, lacking the ability to synthesize images. To enable unified CXR report generation, visual question answering (VQA), and image synthesis within a single framework, we propose MedXChat, a multimodal large language model (LLM) framework that not only analyzes medical images but also generates corresponding CXR images based on user prompts. MedXChat leverages pre-trained LLMs (e.g., mPLUG-Owl) and Stable Diffusion to support text-to-image generation, guided by special tokens <Xray> and </Xray>. We utilize GPT-4 to construct instruction data by generating dialogues with embedded special tokens for each medical image. These instructions are used to fine-tune the LLM, while the text spans between special tokens and their corresponding images are used to fine-tune the Stable Diffusion model. To preserve the original capabilities of the pre-trained diffusion model without degrading its performance, we selectively fine-tune only the encoder and zero-convolution layers. Our model achieves new state-of-the-art performance on both interpretive and generative tasks in MIMIC-CXR. Furthermore, we validate the diagnostic quality of the synthesized images through comparative experiments on downstream classification tasks.

**Comprehensive Radiologists Evaluation for Medical LLMs:** Recent advancements in medical LLMs have demonstrated promising capabilities in radiology tasks such as report generation, VQA, and image synthesis. However, existing evaluation protocols primarily rely on computational metrics, which often fail to capture the clinical relevance, accuracy, and interpretability of model outputs. To address this gap, we introduce a comprehensive evaluation framework that incorporates both objective computational benchmarks and subjective assessments from expert radiologists. This dual-pronged approach provides a more holistic understanding of model performance, capturing nuances such as diagnostic correctness, visual plausibility, and linguistic clarity. We apply this framework to MedXChat, a unified multimodal LLM designed for chest X-ray interpretation and generation. Our results demonstrate that while quantitative metrics reflect general performance trends, radiologists' evaluations are essential for identifying clinically meaningful strengths and weaknesses. This study highlights the importance of integrating domain expertise into the benchmarking process and sets a precedent for more clinically grounded evaluation standards for future medical LLMs.

**Multimodal Diagnosis with Classification and Visual Grounding:** Recent advances in LLMs have demonstrated their potential to integrate image and text understanding for more interpretable medical diagnosis. However, most existing models struggle with multi-label classification and visual grounding in CXRs. To address this gap, we propose MedVisioChat, an instruction-tuned LLM specifically designed for interactive diagnosis with support for both multi-label classification and visual grounding. We present the model with a list of fourteen diseases and prompt it to label each as "Positive" or "Negative" and design a reward function to generate Chain of Thought (CoT) for classification. Additionally, we leverage GPT-4 Turbo to generate instruction-based dialogues containing special tokens for disease names and bounding box predictions to guide the visual grounding process. Through a two-stage training pipeline— classification pretraining followed by GPT-4 Turbo-guided instruction tuning— our model achieves accurate disease identification and localized region highlighting. This enables interpretable,

multi-turn, dialogue-based diagnosis, outperforming existing LLMs on both classification and grounding benchmarks for CXR datasets.

Extensive experiments and ablation studies demonstrate strong robustness and adaptability, confirming the effectiveness of the proposed methods and their potential for real-world applications in medical image analysis and synthesis.

# Deep Generative Modeling for Chest X-ray Interpretation and Synthesis

by

**Ling YANG**

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of

**Doctor of Philosophy**

at

University of Sydney
January, 2026

# Statement of Originality

I, Ling YANG, declare that this thesis titled, "Deep Generative Modeling for Chest X-ray Interpretation and Synthesis", which is submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy, represents my own work except where due acknowledgement have been made. I further declared that it has not been previously included in a thesis, dissertation, or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed: _____

Date: January 12, 2026

# Use of Generative AI Statement

I, Ling YANG, declare that during the preparation of this thesis, ChatGPT was used as an integral part of the research design and is described in the methodology sections of Chapters 4, 5, and 6. In text citation should be included for any section of text and/or result figures that were generated by a generative AI tool. The generative AI tool was not used to enhance or change text. The author takes full responsibility for the submitted thesis and confirms the work is their own and has used generative AI in accordance with University guidelines and policies.

Signed: _____

Date: January 12, 2026

*For My Family*

# *Acknowledgements*

As I near the completion of my Ph.D. journey, I find myself reflecting on nearly a decade of university life across three different countries, each chapter devoted to a different discipline. This path has been filled with both remarkable challenges and deeply personal growth. There were moments of solitude, uncertainty, and emotional struggle—but I never gave up. The unwavering support of my supervisor, family, friends, and those who believed in me helped me stay grounded and move forward. Their encouragement gave me the strength to persist, and it is because of them that I've made it this far.

I would like to express my deepest gratitude to my supervisor, Professor Luping Zhou. Her pioneering work in computer vision has not only shaped my academic direction but also inspired me every step of the way. She has offered meticulous feedback on my papers, motivated me through countless days of research, and provided honest and valuable career advice. She has been more than an academic mentor—she has been a wise and caring guide in life, and I am sincerely thankful for that.

I am also grateful to my lab mates—Zhanyu Wang, Zhenghao Chen, and Xinyu Liang—for their insightful discussions and generous help with experiments and manuscripts. Our collaborations have been both productive and enjoyable, and I cherish the lasting friendships we've built beyond the academic setting.

Words cannot fully capture the gratitude I feel for my family. My father Xiaohua Yang and my mother Xiaomei Leng have carried the financial and emotional weight of my education without ever asking for anything in return. Despite the burden, they never once complained—they simply encouraged me to follow my dreams and become the person I aspire to be. Their sacrifices and love are the foundation on which all my

accomplishments rest. I am forever indebted to them for their unconditional support, even from afar.

Though this doctoral journey is coming to an end, the perseverance I've developed, the rigor I've learned, and the relationships I've built will accompany me into the future. This is not a conclusion, but the beginning of a new chapter—one written with the wisdom of struggle, the warmth of connection, and the hope that I can give back as much as I have received.

Ling YANG
University of Sydney
January 12, 2026

# List of Publications

**Ling Yang**, Zhenghao Chen, Kaisiyuan Wang, Luping Zhou*, “Improving CXR Bone Suppression by Exploiting Domain-level and Instance-level Information", *IEEE Transactions on Medical Imaging (TMI)*, 2025, accepted.

**Ling Yang**, Zhanyu Wang, Zhenghao Chen, Xinyu Liang, Luping Zhou*, “Medxchat: A Unified Multimodal Large Language Model Framework Towards CXRS Understanding and Generation", *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, accepted as Oral Paper.

**Ling Yang**, Xinyu Liang, Zhanyu Wang, Ziyu Diao, Xuan Huang, Die Shen, Xin Tan, Haifeng Li, Zhenghao Chen, Shijun Qiu*, Luping Zhou*, “Constructing A Unified Vision-Language Model for Chest X-Ray Diagnosis, Medical Education, and Data Augmentation",*Radiology: Cardiothoracic Imaging*, 2025, accepted.

**Ling Yang**, Zhanyu Wang, Luping Zhou*, “Medvisiochat: A Multimodal Large Language Model Framework for Interpretable Diagnosis with Visual Grounding in CXRs", *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, accepted.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background and Problem Statement

Chest X-rays (CXRs) remain one of the most widely used imaging modalities in clinical diagnosis due to their low cost, accessibility, and diagnostic breadth. In recent years, the rapid evolution of deep learning and multimodal generative models has significantly advanced the interpretation and synthesis of CXRs. However, despite these developments, several fundamental limitations persist that hinder the reliability, generalizability, and clinical adaptability of current approaches. This section provides a structured overview of the key background concepts that motivate the challenges addressed in this thesis.

**Background 1: Bone Suppression task and Cross-domain Conversion.**

Bone suppression models are developed to enhance the visibility of soft-tissue structures in chest X-rays (CXRs) by reducing or eliminating overlying bone shadows such as ribs and clavicles. These bony structures often obscure critical pathological features—ranging from subtle pulmonary opacities to small nodules—thereby hindering accurate diagnosis of conditions such as pneumonia, pulmonary edema, or pneumothorax. To address this issue, two major families of methods have been established: traditional dual-energy subtraction (DES) and deep learning–based bone suppression models.

Traditional DES techniques [100] as illustrated in Figure 1.1, exploit the differential attenuation characteristics of bone and soft tissue at high- and low-energy X-ray spectra. By acquiring a rapid pair of exposures,

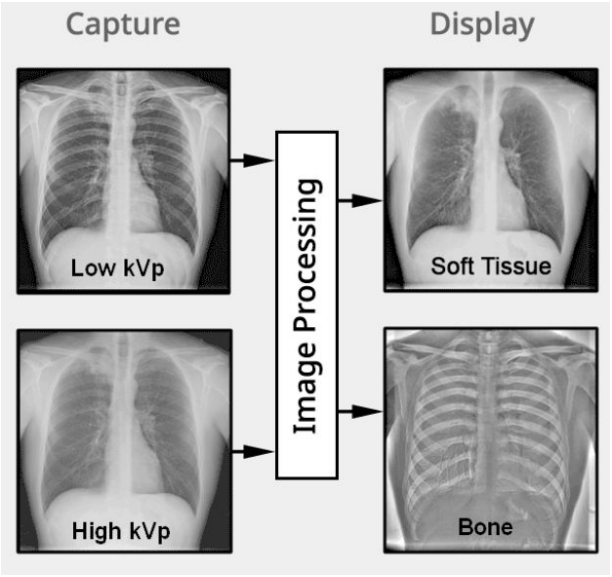



**Figure 1.1:** Process of Dual-Energy Subtraction (DES). High- and low-energy X-ray images are acquired in rapid succession (left) and processed to generate two separate outputs: a soft tissue–only image and a corresponding bone-only image (right). From [100].

DES produces soft-tissue–only and bone-only projections through weighted subtraction. Although effective, DES requires specialized dual-energy hardware and leads to increased radiation dose, limiting its feasibility for routine radiographic practice.

With the rise of deep learning, numerous image-based bone suppression methods have been proposed [29, 58, 124]. Encoder–decoder architectures, including U-Nets, GANs, and more recently Vision Transformers, have been employed to learn direct mappings from bone-included CXRs to bone-suppressed counterparts. These models are attractive due to their computational efficiency, reduced hardware requirements, and integration into existing imaging workflows. Furthermore, bone-suppressed CXRs generated by such methods have been shown to benefit downstream tasks such as disease classification, segmentation, and anomaly detection.

However, despite these advantages, current deep learning–based bone suppression approaches face several limitations. Most models focus heavily on instance-specific features extracted from individual CXRs

and fail to fully leverage domain-level anatomical priors that are consistent across patient populations. CXRs typically exhibit stable structural regularities—such as rib geometry, lung boundaries, and mediastinal contours—which form a low-dimensional anatomical manifold. Existing models seldom incorporate these shared priors explicitly, resulting in limited generalizability when applied to images from heterogeneous scanners, institutions, or patient demographics. Additionally, many generative frameworks exhibit rigid architectural designs that struggle to adapt to complex or subtle structural variations in real-world clinical data.

These challenges highlight the need for more robust cross-domain conversion strategies that integrate both domain-level (shared structural priors) and instance-level (patient-specific) information. Such integration is essential for producing reliable bone-suppressed CXRs that preserve clinically relevant details, maintain anatomical consistency, and remain suitable for downstream diagnostic tasks.

**Background 2: Unified Vision–Language Models for CXR Understanding and Generation.**

The rapid progress of deep learning has led to the development of numerous automated systems for CXR interpretation, including disease classification, radiology report generation, visual question answering (VQA), and medical image synthesis. However, these tasks are typically addressed using separate task-specific models, which creates fragmented workflows, increases system complexity, and limits scalability across diverse clinical environments. Moreover, the lack of interoperability among these models prevents effective cross-task knowledge sharing, making it difficult to deploy them cohesively in real-world clinical settings.

Recent advances in LLMs offer a promising solution by enabling unified vision–language reasoning within a single framework. Unlike conventional CNN- or Transformer-based models that are trained on narrow, task-specific inputs, LLMs are pretrained on large-scale heterogeneous datasets spanning both visual and textual modalities. This broad

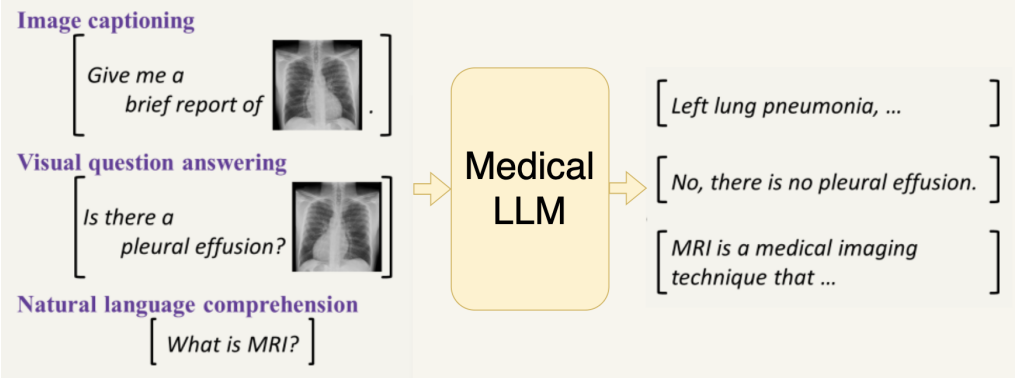



**Figure 1.2:** An overview of the capabilities of a Medical Large Language Model (Medical LLM). The model takes as input various forms of multimodal queries—including image captioning, visual question answering (VQA), and natural language comprehension—and generates corresponding clinical outputs, such as diagnostic reports, visual answers, and medical knowledge explanations. From [54].

multimodal pretraining allows them to generalize more effectively across diagnostic tasks, providing a cohesive interface for clinical decision support. As illustrated in Figure 1.2, medical LLMs can accept multimodal inputs—such as CXRs, textual instructions, and clinical queries—and generate a wide range of outputs, including image captions, diagnostic reports, visual explanations, and domain-specific knowledge.

LLMs also possess the unique capability to jointly process and generate visual and textual information, enabling more interpretable, interactive, and clinically meaningful applications. For example, a unified LLM can answer context-dependent clinical questions, highlight abnormalities through grounded explanations, and generate coherent radiology reports that reflect both image context and medical reasoning. This integrated modeling paradigm enhances transparency and interpretability, fosters better engagement between clinicians and AI systems, and ultimately improves diagnostic consistency and workflow efficiency.

Despite these advantages, current medical LLMs still exhibit important shortcomings. Most existing models accept multimodal inputs but remain limited to text-only outputs, lacking the capability to generate medical images directly from language instructions. To address this gap, recent studies have introduced vector-quantized (VQ) encoders [25]

to discretize images into codebook indices, enabling image synthesis within an LLM-driven token generation pipeline. While conceptually appealing, VQ-based approaches suffer from quantization errors, reduced spatial fidelity, and rigid fixed-length codebooks that constrain the model's capacity to represent complex anatomical variations. Furthermore, static VQ codebooks adapt poorly to heterogeneous clinical data sources, requiring retraining or redesign when encountering new imaging domains, acquisition protocols, or disease distributions.

**Background 3: Clinical Validation and Radiologist-driven Evaluation.**

Although medical LLMs have exhibited promising performance on benchmark datasets, their true clinical applicability remains largely uncertain. Existing studies predominantly rely on automated metrics—such as BLEU, METEOR, or BERTScore for report generation, CheXbert or RadGraph F1 for clinical entity extraction, and FID for image synthesis—to quantify model performance. While these metrics provide convenient numerical assessments, they capture only surface-level similarity or distributional properties and fail to reflect whether the outputs are clinically meaningful or diagnostically reliable.

Radiological interpretation, however, requires nuanced reasoning, contextual understanding, and years of domain-specific experience. Automated metrics cannot evaluate whether subtle abnormalities are correctly identified, whether key findings are omitted, or whether generated descriptions follow radiological standards. Similarly, metrics such as FID cannot determine if synthesized CXRs maintain anatomically accurate structures or faithfully depict pathological patterns of diagnostic importance. As a result, computational benchmarks often overestimate a model's practical utility and may overlook clinically critical errors.

For these reasons, radiologist-driven evaluation is essential when assessing multimodal medical LLMs. Expert radiologists are uniquely qualified to judge factual correctness, diagnostic completeness, and clinical interpretability of generated reports, VQA responses, and synthetic images. Their professional insight enables the identification of issues

that automated metrics cannot detect, such as ambiguous phrasing, inappropriate terminology, inconsistent reasoning, and clinically implausible visual artifacts.

Despite this necessity, systematic expert evaluation remains scarce in current research. Many studies only involve limited case reviews, single-task assessments, or informal clinician feedback. Few works implement structured evaluation protocols spanning multiple CXR tasks—such as report generation, visual question answering, and image synthesis—using radiologists with different experience levels. This lack of rigorous human-centered validation leaves a significant gap in demonstrating clinical readiness.

**Background 3: Multi-label Classification and Visual Grounding for Interpretable Diagnostics.**

In clinical practice, CXRs frequently exhibit multiple coexisting abnormalities, each associated with distinct anatomical regions and often presenting with subtle or overlapping radiographic patterns. As a result, multi-label classification [72] and visual grounding [65] are indispensable components of interpretable and clinically actionable medical AI. These tasks enable models not only to identify the presence of several diseases simultaneously but also to associate each condition with its corresponding anatomical location, thereby mirroring the reasoning process used by radiologists.

Yet, most current systems fall short of this goal. Many models focus exclusively on image-level multi-label classification, generating only a set of predicted diseases without providing spatial cues or visual explanations. Others incorporate textual reasoning but still lack the capability to localize pathological findings within the CXR. This separation between classification and localization limits the interpretability and clinical trustworthiness of AI systems, as radiologists and trainees rely heavily on the spatial context of abnormalities to make diagnostic decisions.

For radiologists, understanding where a pathological finding appears is often as important as understanding what the finding represents. Precise visual grounding—typically formulated as bounding boxes

or regions of interest—is essential for validating the relevance and correctness of model predictions. It also plays a critical role in medical education, where trainees learn to relate textual descriptions of abnormalities to their actual radiographic manifestations.

However, existing LLMs and multimodal architectures still struggle to achieve such fine-grained grounding due to several inherent limitations. First, the ambiguous or diffuse boundaries of many thoracic lesions make precise localization challenging, even for human experts. Second, current vision–language models possess limited spatial reasoning capabilities, which restrict their ability to accurately associate textual descriptions with specific anatomical regions in the image. Third, most existing systems lack unified frameworks that can simultaneously handle multi-label classification, visual grounding, and explanatory reasoning, often resulting in outputs that are incomplete or misaligned across modalities. Finally, inconsistencies between vision and language embeddings further exacerbate this issue, frequently leading to mismatches between predicted disease labels and their corresponding spatial regions.

## 1.2 Challenges and Motivations

### 1.2.1 Challenges

Analyzing and synthesizing medical images using traditional deep learning models and LLMs remains a challenging task. Although deep learning–based bone suppression models are more accessible than traditional methods, they often lack generalizability and flexibility. Meanwhile, most multimodal LLMs exhibit strong capabilities in visual-text understanding but are still limited to text-only outputs and struggle with multi-class classification and lesion localization in fine-grained medical images. Our research focuses on three key tasks in medical image analysis and synthesis. First, we propose a bone suppression model for CXRs that leverages a continuous multi-head attention codebook combined with cross-covariance attention to simultaneously capture domain-level and instance-level information. Second, we build a unified large model for medical report generation, VQA, and CXR synthesis by integrating an

LLM with Stable Diffusion. Third, we leverage prompt engineering and instruction tuning using GPT- generated data for classification and visual grounding. Through these efforts, we aim to develop a more robust and generalizable solution for bone suppression and multimodal medical tasks. Our method not only improves image and text generation quality but also enables more accurate downstream tasks such as classification and segmentation. Furthermore, we collaborate with expert radiologists to systematically evaluate the performance of our unified model, marking a significant advancement in the field of medical image analysis and synthesis.

**Challeng 1: Cross-domain Image Conversion between Bone CXR and Bone-suppressed CXR.**

Bone suppression is a critical preprocessing step in CXR analysis and synthesis, aimed at enhancing the visibility of soft-tissue structures by reducing or eliminating overlying bony anatomy such as ribs and clavicles. Recent research has focused on deep learning–based bone suppression models that learn a mapping from bone-included CXRs to bone-suppressed counterparts. Early approaches to bone suppression in CXRs, spanning both supervised [89, 124, 29, 9, 58] and unsupervised methods [24, 56, 31], directly incorporated classical architectures like U-Net [38] and dilated convolutions [124]. These models typically remove bone structures by identifying their locations and generating bone-suppressed images accordingly. Compared to DES, such vision-based techniques offer substantial advantages in terms of computational efficiency and clinical practicality.

More recently, generative modeling techniques have gained traction in this domain, with frameworks such as Auto-Encoders [29, 52] and GANs [16, 125, 124, 76] being applied to bone suppression tasks. Drawing on developments from image-to-image translation [52, 16, 125, 76], these methods aim to learn cross-domain mappings that transform bone CXRs into bone-free images.

Despite these advancements, current deep learning methods primarily focus on extracting instance-specific features from CXRs, overlooking a key insight: CXRs exhibit high structural similarity and typically lie on a low-dimensional manifold. This shared anatomical structure across patient CXRs, representing domain-level consistency, has yet to be fully exploited. Without leveraging these shared structural priors, models may struggle with generalization and robustness across diverse clinical settings.

This cross-domain image conversion refers to the inherent distributional, structural, and semantic differences between the two image types. Bone-suppressed CXRs exhibit distinct pixel intensities, anatomical emphasis, and visual priors compared to their bone-included counterparts. When models are trained solely on synthetic bone suppression pairs, they often overfit to idealized transformations that fail to reflect the noise, anatomical variability, and degradation commonly found in real clinical images. Consequently, the learned representations may not transfer well to real-world settings. On the other hand, unpaired training strategies, such as CycleGAN [125], tend to prioritize global appearance translation and may sacrifice important local structural details that are essential for clinical interpretation. The result of this mismatch is a decline in the performance of downstream tasks, such as disease classification and disease segmentation, where accurate soft-tissue representation is critical.

Cross-domain image conversion between bone and bone-suppressed CXRs remains an open challenge. It calls for the development of domain-adaptive learning strategies that can effectively align features across domains while preserving clinical fidelity. Promising solutions lie in learning domain-invariant representations, applying attention-based feature alignment techniques, and adopting self-supervised contrastive learning methods that simultaneously capture global structural coherence and localized anatomical detail. Addressing this cross-domain image conversion is essential to ensure the robustness and generalizability of bone suppression models and their downstream tasks in practical clinical environments.

**Challeng 2: Lack of Adaptive Unified Multimodal for CXR Understanding and Generation.**

Recent breakthroughs in LLMs [66, 4] have introduced exciting opportunities for unified vision-language learning in the medical domain. These models, by design, aim to bridge the gap between visual and textual information, enabling integrated reasoning across multiple modalities. However, most existing medical LLMs are limited in their output capabilities [93, 4]: while they can accept both CXRs and textual prompts as input, they are typically confined to producing only textual outputs. This constraint significantly limits their utility in real-world clinical workflows that require more comprehensive, bidirectional multimodal capabilities—such as synthesizing diagnostic chest X-rays from clinical instructions, conducting visual-textual round-trip inference, or validating generated reports against image content.

To address this limitation, recent efforts [53, 54] have attempted to extend LLMs with image generation functionalities by leveraging VQ strategies [25]. These approaches discretize image features into tokenized codebook indices, allowing visual content to be processed alongside textual data. While VQ-based representations offer a structured way to align image and language tokens, they also introduce inherent drawbacks: quantization errors, reduced spatial fidelity, and constraints due to fixed-length codebooks. These issues lead to degradation in image quality and limit the model's flexibility in adapting to the diverse and nuanced anatomical details critical in medical imaging.

Furthermore, current VQ-based architectures often rely on static, non-adaptive codebooks. These rigid structures are difficult to scale and generalize [10, 11]. Incorporating new data distributions—such as CXRs acquired from different institutions, imaging protocols, or patient demographics—typically necessitates retraining the entire model with a new or expanded codebook. This retraining process is computationally intensive and impractical in dynamic clinical environments where continual learning and rapid adaptation are necessary. As a result, the scalability, robustness, and generalizability of current LLM-based frameworks remain severely limited in clinical deployment scenarios.

To overcome these challenges, there is an urgent need for a more flexible, extensible, and clinically grounded LLM framework capable of performing both medical image understanding and synthesis in a unified and end-to-end fashion. Such a model should incorporate adaptive encoding strategies that can dynamically represent image content without incurring significant quantization losses, support multi-granularity vision-language alignment, and be designed to generalize across heterogeneous data sources without extensive retraining. Additionally, it should preserve diagnostic interpretability and reliability, ensuring its outputs remain clinically meaningful and trustworthy.

Developing such a framework is a crucial step toward realizing general purpose, multimodal AI assistants for healthcare—systems that can not only read and interpret medical images but also generate them on demand, validate diagnostic hypotheses, and assist in interactive, clinician-centered decision-making.

**Challenge 3: Clinical Validation Gap: Ensuring Radiologist-Endorsed Evaluation for Unified LLMs in CXR Analysis and Synthesis.**

Nowadays, while a growing number of medical LLMs [54, 53, 55, 91] have been developed for CXR understanding and generation, their actual readiness for clinical application remains uncertain. Most models are validated only on benchmark datasets or automated metrics, which often fail to reflect real-world diagnostic needs. However, true clinical deployment requires rigorous, systematic evaluation by domain experts—namely, professional radiologists—who can assess the model's performance in terms of medical relevance, diagnostic accuracy, and interpretability.

To bridge this gap, we organized a comprehensive evaluation of our proposed unified multimodal large language model, MedXChat, to thoroughly assess its effectiveness and reliability across key CXR-related tasks. A panel of six certified radiologists with varying years of clinical experience was recruited to participate in this study. These experts were tasked with systematically reviewing and rating the model's outputs across three core applications: CXR report generation, medical VQA

and text-to-image CXR synthesis. For a comprehensive comparison, two additional state-of-the-art models were also included in the evaluation.

This expert-driven evaluation not only validates the clinical applicability of MedXChat but also provides critical insights into its strengths and limitations under real-world medical standards, setting a foundation for its integration into radiological workflows.

**Challenge 4: Interpretable Medical Multi-label Classification and Precise Visual Grounding.**

In clinical practice, CXRs often exhibit multiple coexisting pathologies, making multi-label classification [72, 112] and disease visual grounding [65, 12, 20] a fundamental requirement. However, beyond merely identifying the presence of diseases, interpretability—such as enabling dialog-based interaction and explaining findings directly on the CXR—is crucial for clinical adoption and trust. Traditional AI systems typically decouple classification from localization, lacking the integrated reasoning necessary to generate clinically meaningful explanations.

Recent advances in LLMs offer promising avenues for unified vision-language understanding [66, 93]. While these models have shown strong capabilities in text generation and general image comprehension, their application to interpretable multi-label classification and fine-grained visual grounding in medical imaging remains underdeveloped. Most existing LLMs [23, 94, 55] cannot explicitly associate multiple disease labels with their corresponding anatomical regions in CXRs, limiting their clinical relevance and usability.

This gap introduces several critical challenges. First, many thoracic diseases present with subtle, overlapping, or ambiguous radiographic features, requiring models to detect fine-grained visual cues [94]. Second, current LLMs generally lack visual grounding capabilities—the ability to localize specific image regions that support each predicted label. This spatial transparency is essential for validating model outputs and ensuring diagnostic credibility. Without such interpretability, predictions remain opaque and risk clinical rejection.

Addressing these limitations requires the development of unified frameworks that support both multi-label disease classification and region-to-label grounding. Potential solutions include prompt-based classification tuning, instruction-finetuned LLMs equipped with grounding heads, and attention-based cross-modal feature alignment. Bridging semantic prediction and spatial explainability is essential for building medical AI systems that are not only accurate, but also trustworthy and interpretable in real-world diagnostic workflows.

### 1.2.2 Motivations

In particular, the primary motivations for this thesis can be summarized in the following four aspects.

**Motivation 1: Develop A Two-stage Domain-level and Instance-level Information Bone CXR Suppression Network to Address Challenge 1.**

In the task of bone suppression for chest X-rays, there exist subtle differences in structure and texture between bone-included and bone-suppressed images, resulting in a relatively mild domain gap. Most existing methods focus primarily on instance-level features of individual images while overlooking the domain-level anatomical priors that are commonly shared across CXRs from different patients. This limitation leads to insufficient model robustness and consistency, significantly hindering the generalizability of current approaches in real-world clinical settings. To address this issue, we propose a two-stage bone suppression network that jointly leverages both domain-level and instance-level information. In the first stage, we introduce a Multi-head Codebook Attention (MCA) Learning Module to enhance the model's ability to capture and represent prior knowledge from the bone-suppressed CXR domain. In the second stage, we employ a Cross-Covariance Attention Block (CAB) Network to encode input-specific features from each CXR, guiding the generation process more precisely. To further strengthen instance-level feature extraction across both stages, we replace traditional Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs).

This enables the model to leverage long-range dependencies and attend to clinically relevant patterns across spatially distant regions in chest X-ray images. By integrating both domain-level and instance-level information, our approach not only enhances the generalization ability of the bone suppression network but also significantly improves the accuracy of downstream classification and segmentation tasks, making it more robust and reliable for real-world clinical applications.

**Motivation 2: Instruction-guided Fine-tuning of LLM with Stable Diffusion to Address Challenge 2.**

To address the challenges posed by imbalanced and noisy To develop a unified LLM capable of performing report generation, VQA, and image synthesis from CXRs, while addressing key limitations in current VQ-based image generation methods—such as quantization loss, fixed codebook constraints, and limited adaptability to new data distributions—we introduce MedXChat, a unified LLM framework that seamlessly integrates both interpretive and generative capabilities for CXR analysis. Our framework first utilizes GPT-4 to generate instruction-style data tailored for image synthesis, which includes special tokens to guide the generation process. We then fine-tune the LLM using a combination of MIMIC-CXR reports, VQA datasets, and the synthetic instruction data, enabling the model to simultaneously perform the three core tasks. Finally, we align the learned multimodal embeddings with image generation by fine-tuning a Stable Diffusion (SD) decoder using the CXR images paired with prompts containing the special tokens. Importantly, we keep most of the SD architecture frozen and train only the SD encoder and several lightweight zero-convolution layers to ensure efficiency and task-specific adaptability. This unified approach not only eliminates the need for fragmented task-specific models but also enhances performance, scalability, and interpretability in real-world clinical applications.

**Motivation 3: Radiologist-Led Evaluation of MedXChat Addresses Challenge 3.**

To enhance the clinical trustworthiness and practical utility of model

outputs, we propose a radiologist-led, multi-tier evaluation framework to systematically assess the clinical applicability of MedXChat—our unified multimodal LLM for CXR analysis and synthesis. We assembled an expert panel of six registered radiologists with varying levels of clinical experience. This panel evaluated MedXChat's performance across three core tasks—report generation, VQA, and text-to-image synthesis—through comparative assessments involving two additional medical LLMs. The evaluation process was structured into three stages. In the first stage, three junior radiologists conducted initial screenings to identify major issues and consistency errors in the model outputs. In the second stage, two senior radiologists with 7–10 years of clinical experience evaluated the outputs using eight clinical criteria, including factual accuracy, diagnostic completeness, adherence to radiological terminology, and interpretability. In the final stage, a chief radiologist with over 25 years of experience adjudicated any scoring discrepancies and ensured fairness and consistency in the overall evaluation. This expert-led clinical review not only provided a comprehensive audit of MedXChat's performance but also uncovered critical shortcomings not easily detected by automatic evaluation metrics—such as failure to detect rare diseases, insufficient clinical context in VQA responses, unnatural phrasing in generated reports and inaccurate generated medical images. These clinical insights offered valuable guidance for iterative model refinement, aligning MedXChat's technical capabilities more closely with real-world clinical expectations. By establishing this rigorous evaluation pipeline, we effectively bridge the gap between algorithmic performance and clinical implementation. The process ensures that MedXChat delivers outputs that are not only accurate but also clinically interpretable and trustworthy, laying a solid foundation for its deployment in high-stakes healthcare environments.

**Motivation 4: Prompt Engineering and Chain of Thought Framework Addresses Challenge 4.**

To tackle the challenge of enabling LLMs to perform interpretable multi-label classification and visual grounding for CXRs, we propose MedVisioChat, a unified framework that integrates visual grounding

capabilities with medical knowledge to support multi-round diagnostic dialogues. This framework transforms traditional medical image interpretation by allowing the model not only to classify diseases but also to localize them within the image, thereby enhancing diagnostic transparency and clinical trust. For multi-label classification, we design specific prompts that list fourteen common thoracic diseases and instruct the model to label each as "Positive" or "Negative." The input follows a structured format, such as: "Atelectasis: Negative." This format ensures that the model systematically assesses each disease category with binary labels, enabling comprehensive diagnostic coverage. We also design an accuracy reward and a format reward to generate chain of thought to enhance classification accuracy. To bridge semantic and visual modalities, we introduce a novel instruction-based prompt design incorporating special tokens such as <ref> and <box>. These tokens explicitly link each predicted disease label to its corresponding region in the CXR image, facilitating accurate and interpretable visual grounding. The model is thus trained not only to recognize disease categories but also to provide spatial explanations in natural language, enhancing interpretability. This integration of prompt engineering, instruction fine-tuning, and classification-aware visual alignment significantly improves both the diagnostic performance and explainability of the model. MedVisioChat provides a promising pathway toward building trustworthy and deployable AI diagnostic systems, with strong potential for real-world clinical adoption.

## 1.3 Thesis Contribution and Organization

This section summarizes the key contributions and outlines the structure of the thesis. In Chapter 2, we review existing literature on deep learning approaches for medical image analysis and synthesis, highlighting their relevance to the methods developed in this work. Chapters Chapters 3 to 6 present our proposed solutions to various challenges in medical image understanding and generation, each building on and extending the current state of the art. Lastly, Chapter 7 concludes the thesis and explores promising directions for future research.

The primary contribution of this thesis lies in advancing medical image analysis and synthesis through three key tasks: bone suppression in CXR, the development of a unified LLM for CXR understanding and generation, and an LLM framework for multi-disease classification with precise visual grounding.

For the CXR bone suppression task, we highlight the effectiveness of cross-domain image conversion techniques in mitigating the domain shift between bone-included and bone-suppressed CXRs. To this end, we propose an end-to-end two-stage framework that jointly performs cross-domain translation and bone suppression. Specifically, we introduce a Multi-head Codebook Attention (MCA) mechanism to capture domain-level priors from bone-suppressed CXRs, while enhancing instance-level feature representation through the integration of both vanilla and cross-covariance Transformer modules. Extensive experiments demonstrate that our method not only produces bone-suppressed images with superior reconstruction quality but also significantly enhances downstream tasks such as pneumonia classification and lesion segmentation.

For the unified LLM for CXR understanding and generation task, we emphasize the importance of developing a generalizable model capable of seamlessly handling diverse vision-language tasks within a single framework. To this end, we propose MedXChat, a unified LLM system that jointly supports CXR-to-report generation, CXR-VQA, and text-to-CXR image synthesis. We adopt a visual-language architecture that incorporates powerful pretrained components—such as ViT-based visual encoders and the mPLUG-Owl language model—and optimize them using lightweight strategies like delta-tuning and LoRA to maintain efficiency. Moreover, we address limitations in previous VQ-based models by introducing a novel instruction-driven diffusion synthesis module, which enables high-quality medical image generation without retraining large components. The system is further enhanced through the use of GPT-4-generated instruction data and stable diffusion, which guides the model's image generation ability with rich contextual prompts. Comprehensive evaluations demonstrate that MedXChat achieves superior performance across all tasks, delivering interpretable, clinically meaningful

results that set a new standard for unified medical vision-language understanding and generation.

For the radiologist-led evaluation of MedXChat, we designed a comprehensive multi-tier assessment protocol to systematically evaluate its clinical effectiveness across three key tasks: report generation, VQA, and text-to-image synthesis. A panel of six certified radiologists, representing a spectrum of clinical expertise from junior doctors to senior chief specialists, was convened to perform a blinded assessment and comparative analysis of MedXChat alongside two other medical LLMs. The evaluation process involved rigorous scoring across multiple dimensions including factual accuracy, diagnostic relevance, adherence to radiological terminology, and image realism. The results demonstrate that MedXChat consistently outperforms alternatives by generating clinically coherent diagnostic reports, delivering accurate and contextually grounded answers in VQA, and synthesizing CXR images that are both structurally plausible and diagnostically meaningful.

For the task of interpretable multi-label classification and precise visual grounding in CXRs, we propose a unified instruction-guided framework, MedVisioChat, that enables simultaneous disease classification and spatial localization with clinical interpretability. The model is designed to bridge the gap between textual and visual modalities by utilizing a prompt-engineered multi-label classification system, where each of the 14 disease categories is explicitly labeled as "Positive" or "Negative." To further strengthen interpretability, we integrate a chain-of-thought (CoT) mechanism into the classification pipeline. The CoT provides step-by-step reasoning behind the model's predictions. To enhance visual interpretability, we introduce special instruction tokens—such as <ref> for disease labels and <box> for bounding box annotations—that guide the model to align each diagnosis with its corresponding region in the image. This explicit alignment enables the model to accurately predict the visual grounding locations of diseases based on the classification results. Built upon a powerful LLM backbone and fine-tuned with LoRA on carefully curated instruction-following dialogues, our model achieves superior performance and outperforms existing baseline methods.

# Chapter 2

# Literature Review

In this chapter, we begin with an overview of Medical Deep Generative Modeling, introducing major model families including autoregressive models, variational inference–based generative models, generative adversarial networks, transformers, and diffusion models, with an emphasis on their relevance to medical imaging applications. We then discuss the recent advances in Large Language Models (LLMs), focusing on medical-specific LLMs, instruction tuning strategies, and prompt engineering techniques that enable effective multimodal interaction and domain adaptation. Building on this foundation, we explore a range of medical imaging tasks, covering bone suppression for improved diagnostic accuracy, automated report generation, visual question answering, disease classification, and visual grounding for explainable AI. For each task, we review representative methods, highlight key technical challenges, and analyze both the progress achieved and the gaps that remain. Together, these sections provide a comprehensive perspective on how deep generative modeling and large language models are shaping the future of medical image analysis and synthesis.

## 2.1 Medical Deep Generative Modeling

### 2.1.1 Autoregressive Models

Autoregressive (AR) models are a class of generative models that factorize the joint probability distribution of sequential or structured data into

a product of conditional probabilities:

$$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{<i}), \tag{2.1}$$

where each element $x_i$ is generated conditioned on all previously generated elements. This sequential dependency enables AR models to capture long-range dependencies and complex structural patterns. Initially popularized in natural language processing and speech synthesis, the autoregressive paradigm has been extended to two-dimensional image generation tasks via raster-scan or patch-based strategies, as exemplified by PixelRNN, PixelCNN, and their variants. By conditioning each generation step on the accumulated context, AR models can produce outputs that are both detailed and semantically coherent.

In the medical imaging domain, the autoregressive paradigm offers several advantages. First, its step-by-step generation allows for fine-grained control over the output, which is particularly valuable in applications where spatial consistency and anatomical plausibility are critical. For example, pixel-level AR models can synthesize medical images with subtle pathological features, while token-level AR decoders can produce diagnostic reports with logical flow and factual accuracy. Second, AR models are naturally suited for multimodal learning, supporting conditional generation tasks such as generating reports from images, synthesizing images from textual descriptions, or producing both modalities jointly. Third, their explicit conditional dependency mechanism facilitates the integration of structured prior knowledge—such as clinical ontologies or radiology lexicons—guiding the generation process toward domain-relevant outputs.

AR models have been explored in medical imaging synthesis for their ability to sequentially model spatial dependencies, either at the pixel or patch level. Wang et al. [102] demonstrated the potential of AR approaches for volumetric data. They treat 3D medical images as sequences of visual tokens— organized based on spatial proximity, contrast, and semantic similarity— and train the model to predict the next

token autoregressively. This method effectively captures long-range contextual information and achieves strong performance across nine downstream medical imaging tasks. Shin et al. [82] proposed the Recurrent Neural Cascade Model (RNCM), one of the earliest frameworks to integrate convolutional neural networks for chest X-ray feature extraction with an autoregressive long short-term memory (LSTM) decoder for sequential medical report generation. By first predicting high-level clinical tags and then conditioning sentence generation on these tags, the model improved factual accuracy and coherence, demonstrating the feasibility of deep learning–based automated chest X-ray interpretation and paving the way for subsequent vision–language approaches in radiology.

However, AR models also have certain limitations, particularly in high-resolution image synthesis where the sequential nature of generation incurs significant computational cost. The inherently serial decoding process results in slower inference compared to parallel generation methods, and early prediction errors may propagate through the sequence, degrading the overall quality. To mitigate these issues, hybrid approaches have emerged that combine AR components with other generative paradigms, such as variational inference or diffusion models, thereby retaining the strong sequential modeling capabilities of AR while improving efficiency and robustness. These advancements have expanded the applicability of AR models beyond their original domains, making them increasingly relevant for medical image synthesis, automated report generation, and multimodal medical AI systems.

### 2.1.2 Variational Autoencoder

The Variational Autoencoder (VAE), introduced by Kingma and Welling [48], is a generative model that formulates data generation as a probabilistic inference problem. Unlike conventional autoencoders that learn deterministic latent representations, VAEs introduce a set of latent variables whose posterior distribution given the data is generally intractable. To address this, VAEs employ variational inference to approximate the true posterior $p(\mathbf{z} \mid \mathbf{x})$ with a tractable variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$,

typically parameterized as a Gaussian whose mean and variance are predicted by an encoder network.

The generative process is modeled by a decoder network $p_\theta(\mathbf{x} \mid \mathbf{z})$, which reconstructs the input data from samples drawn from the latent distribution. Training is achieved by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})), \tag{2.2}$$

where the first term encourages faithful reconstruction of the input, and the second term regularizes the latent space by penalizing divergence from the prior distribution $p(\mathbf{z})$ (often a standard normal). VAEs have become a cornerstone in deep generative modeling due to their probabilistic formulation, which enables explicit likelihood estimation, smooth latent space interpolation, and predictive uncertainty quantification.

In medical imaging, VAEs and their variants (e.g., Conditional VAEs (CVAE) [86], $\beta$-VAEs [35]) have been widely adopted for synthesis, anomaly detection, and modality translation. For example, Chartsias et al. [8] proposed a multi-input CVAE for cross-modality synthesis between cardiac MR and CT, enabling anatomically consistent generation across modalities by conditioning the latent representation on multi-source image features. Similarly, Dalca et al. [18] introduced VoxelMorph, a VAE-based probabilistic image registration framework that learns a distribution over plausible deformation fields, allowing both accurate alignment and uncertainty estimation in brain MRI registration.

Extensions of the VAE framework have been developed to address limitations in capturing complex, multi-scale structures in medical images. Hierarchical VAEs [87, 95] introduce multiple layers of latent variables organized in a hierarchy, where higher-level latents encode global structural information and lower-level latents capture fine-grained details. Hierarchical VAEs have shown effectiveness in medical image synthesis by capturing anatomical priors and multi-modal dependencies. Kapoor et al. [44] introduced a Multiscale Metamorphic VAE for 3D brain MRI generation, modeling coarse-to-fine morphological transformations

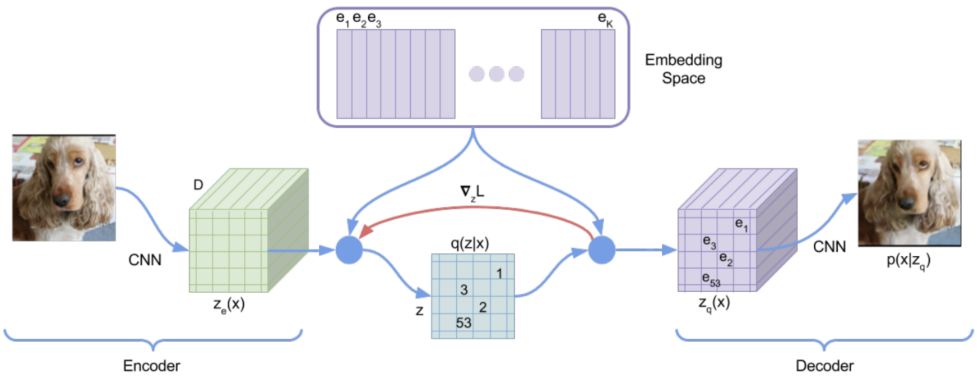



**Figure 2.1:** A typical VQ-VAE framework comprises three main components: an encoder, a learnable codebook containing discrete embedding vectors, and a decoder.From [96].

on a reference template to ensure anatomical consistency and superior fidelity compared to standard VAEs and GANs. More recently, Dorent et al. [21] proposed MHVAE, a Multi-modal Hierarchical VAE that fuses brain MRI and intra-operative ultrasound modalities into a shared hierarchical latent space and uses probabilistic fusion and adversarial training to generate missing modality images with high realism. These hierarchical structures enhance both the expressiveness and clinical plausibility of generated outputs.

As show in Figure. 2.1, Vector-Quantized VAE (VQ-VAE) [96] extends the VAE by replacing the continuous latent space with a discrete codebook learned jointly with the encoder and decoder. This discrete representation has been particularly beneficial for high-fidelity medical image synthesis and compression. For example, [30] proposed a VQ-VAE–based approach for mapping multi-sequence MRI scans into a unified discrete latent space, enabling non-adversarial cross-sequence synthesis via a Seq2Seq generator. Similarly, [115] introduced a VQ memory module into U-Net for weakly supervised segmentation, where discrete codebooks serve as a repository of anatomical features to enhance pseudo-label generation and overall robustness.

While VAEs have demonstrated utility in medical image analysis

and synthesis, several inherent limitations hinder their broader clinical applicability. A key drawback is their tendency to produce oversmoothed or blurry reconstructions in high-resolution modalities such as MRI or CT, primarily due to pixel-wise reconstruction losses (e.g., MSE) that average over plausible outputs and suppress diagnostically important fine details. The common assumption of an isotropic Gaussian prior is often too restrictive to capture the complex, multi-modal distributions of medical images, leading to suboptimal latent representations, particularly for rare pathologies. VAEs are also prone to posterior collapse, where latent variables fail to encode meaningful information when paired with highly expressive decoders, thereby compromising interpretability and limiting sensitivity to subtle abnormalities. Moreover, the fixed trade-off in the ELBO between reconstruction fidelity and latent regularization is rarely optimal across diverse clinical tasks—for example, anomaly detection may require prioritizing sensitivity over fidelity, whereas surgical planning demands both. Finally, VAEs trained on domain-specific datasets exhibit limited generalizability across imaging modalities, acquisition protocols, and institutions, with domain shifts often degrading real-world performance.

### 2.1.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [28] are a class of generative models that formulate the learning process as a two-player minimax game between a generator and a discriminator. The generator aims to produce samples that are indistinguishable from real data, while the discriminator learns to distinguish between real and generated samples. In other words, a discriminator $D$ and a generator $G$ play the following two-player minimax game with value function $V(G, D)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2.3}$$

This adversarial training paradigm encourages the generator to capture the underlying data distribution, enabling the synthesis of high-fidelity and perceptually realistic outputs without requiring explicit likelihood

estimation.

In medical imaging, GANs and their variants (e.g., conditional GANs, CycleGAN, StyleGAN) have been widely employed for a range of tasks, including image-to-image translation, modality conversion, super- resolution, and artifact reduction. For example, Wolterink et al. [109] utilized CycleGAN for unpaired MR-to-CT synthesis, facilitating attenuation correction in PET imaging without the need for paired training data. Similarly, Armanious et al. [3] proposed MedGAN, a cGAN-based framework that integrates perceptual, style, and adversarial losses for multi-modal medical image translation, achieving state-of-the-art performance in MR-to-CT and PET-to-CT conversion. Recent studies have further extended GANs to address domain adaptation and data scarcity challenges. Yang et al.[116] proposed a GAN-based framework for low-dose CT denoising, incorporating Wasserstein distance with a perceptual loss to achieve superior noise suppression while preserving fine anatomical details. More recently, Mahapatra et al.[63] introduced a self-supervised GAN for histopathology stain normalization, leveraging semantic guidance to maintain tissue structure and morphological fidelity during style transfer.

One of the most important variants of GANs is the Vector Quantized Generative Adversarial Network (VQGAN) [25], which combines the adversarial training mechanism of Generative Adversarial Networks (GANs) with the discrete latent representation learning capability of the Vector Quantized Variational Autoencoder (VQ-VAE). The core idea is to apply vector quantization to map the encoder outputs into a finite codebook $\mathcal{Z} = \{e_k \in \mathbb{R}^D\}_{k=1}^K$, resulting in sparse, discrete, and semantically rich latent representations.

Given an input image $x$, the encoder $E_\theta$ produces a continuous latent vector $z_e = E_\theta(x)$, which is then quantized via nearest-neighbor lookup:

$$z_q = e_k, \quad \text{where} \quad k = \arg\min_j \|z_e - e_j\|_2. \tag{2.4}$$

The decoder (generator) $G_\phi$ reconstructs the image from the quantized representation $\hat{x} = G_\phi(z_q)$. The model is trained by jointly optimizing the following objective:

$$\begin{aligned}\mathcal{L}_{\text{VQGAN}} = \mathcal{L}_{\text{rec}}(x, \hat{x}) + \beta \left\| \text{sg}[z_e] - e_k \right\|_2^2 + \left\| z_e - \text{sg}[e_k] \right\|_2^2 \\ + \lambda \mathcal{L}_{\text{GAN}}(x, \hat{x}) + \gamma \mathcal{L}_{\text{perc}}(x, \hat{x}),\end{aligned} \tag{2.5}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss (e.g., $L_1$ or $L_2$ distance), the second and third terms correspond to codebook update and commitment loss, $\text{sg}[\cdot]$ denotes the stop-gradient operation, $\mathcal{L}_{\text{GAN}}$ is the adversarial loss for enhancing high-frequency details, and $\mathcal{L}_{\text{perc}}$ is the perceptual loss (e.g., LPIPS) to maintain semantic consistency. VQGAN combines the efficiency of discrete latent representation learning from VQ-VAE with the high-fidelity synthesis capability of GANs, making it particularly effective in medical imaging tasks. It has been successfully applied in cross-modality medical image translation (e.g., MRI-to-CT), super-resolution reconstruction, histopathology image synthesis, and report-to-image generation, consistently preserving clinically relevant structures while generating realistic details. Furthermore, the discrete latent space makes VQGAN a strong backbone for Transformer-based generative models in radiology, enabling scalable autoregressive modeling over compact image tokens.

Despite their success in generating visually convincing images, GANs face challenges in clinical adoption due to issues such as mode collapse, training instability, and the lack of explicit uncertainty quantification. Second, while adversarial losses enhance perceptual realism, they may inadvertently introduce hallucinated structures, which is a critical safety concern in medical applications. Furthermore, the stop-gradient (SG) operation helps stabilize training by preventing interference between the encoder and the codebook, but it can slow convergence and lead to suboptimal discrete representations due to blocked gradient flow. This limitation may hinder fine-detail preservation and cross-modal consistency in complex medical imaging tasks, motivating strategies such as soft quantization to retain gradient information.

### 2.1.4 Transformer-based Models

Transformer-based models have emerged as a dominant architecture in deep learning, particularly for sequence modeling tasks. Their central innovation is the self-attention mechanism, which enables the model to capture long-range dependencies between elements in a sequence without relying on recurrence or convolution. This design allows for fully parallelizable training and efficient modeling of global context. The self-attention operation can be formally expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.6}$$

where $Q$, $K$, and $V$ are the query, key, and value matrices derived from the input embeddings, and $d_k$ is the dimensionality of the key vectors.

In generative modeling, Transformers are often employed in an autoregressive (AR) fashion, predicting the next token $x_t$ conditioned on all previously generated tokens $x_{<t}$. This probabilistic formulation can be written as:

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_{<t}), \tag{2.7}$$

where $T$ denotes the sequence length. This framework, originally developed for natural language processing (NLP), has been extended to computer vision and medical imaging, enabling tasks such as image synthesis, radiology report generation, and visual question answering (VQA). Vision Transformers (ViTs) treat images as sequences of fixed-size patches, while hybrid CNN–Transformer architectures employ convolutional layers for local feature extraction followed by Transformer layers for global context modeling.

Recent research has explored integrating Transformer architectures into generative frameworks to enhance medical image analysis and synthesis. Chen et al. [14] introduced a memory-driven Transformer for chest X-ray report generation, incorporating an external memory module to store and retrieve image-relevant textual patterns. These retrieved

patterns are integrated into the Transformer decoder's self-attention mechanism, enhancing the coherence of long reports and mitigating repetitive errors often observed in traditional CNN–RNN baselines. Evaluated on the MIMIC-CXR dataset, the approach achieved state-of-the-art performance across BLEU, METEOR, and clinical efficacy metrics. In parallel, Dalmaz et al. [19] developed ResViT, a GAN architecture augmented with Vision Transformer (ViT) modules. Its generator features Aggregated Residual Transformer (ART) blocks that blend convolutional precision with global context modeling, enabling accurate synthesis of missing MRI contrasts and MRI-to-CT transformations. This hybrid design demonstrated superior anatomical fidelity and structural coherence compared to CNN-only generative models.

In medical imaging, Transformer-based generative models offer several advantages. First, their ability to model long-range spatial dependencies is crucial for capturing clinically relevant structures that may be spatially distant yet semantically related, such as lesions and associated anatomical landmarks. Second, their scalability enables training on large-scale multimodal datasets that incorporate not only imaging data but also associated clinical reports and metadata. Third, their flexibility allows the unification of multiple medical tasks—such as cross-modality synthesis, segmentation, and diagnostic report generation—within a single framework.

### 2.1.5 Diffusion Models

Diffusion models are a class of generative models that learn to synthesize data by reversing a gradual noising process. Inspired by non-equilibrium thermodynamics, these models define a forward diffusion process that progressively adds Gaussian noise to the input data over $T$ steps, transforming the data distribution $q(\mathbf{x}_0)$ into an isotropic Gaussian distribution $q(\mathbf{x}_T)$. The model is then trained to learn the reverse denoising process, gradually reconstructing data from noise.

Formally, the forward process is defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \tag{2.8}$$

where $\beta_t \in (0,1)$ is a predefined variance schedule controlling the noise level at each timestep $t$. By iteratively applying this process, we obtain a closed-form distribution:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{2.9}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The generative process learns the reverse conditional distributions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ parameterized by a neural network, typically predicting either the mean or the added noise $\epsilon_\theta(\mathbf{x}_t, t)$:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}\right). \tag{2.10}$$

The training objective can be simplified to a denoising score matching loss [36]:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, t\right)\right\|_2^2\right], \tag{2.11}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Diffusion models have emerged as a powerful class of generative frameworks, achieving state-of-the-art performance across diverse synthesis tasks by modeling complex data distributions without adversarial training, thus avoiding common pitfalls of GANs such as mode collapse and training instability. In medical imaging, they have been successfully applied to high-fidelity MRI and CT synthesis, cross-modality translation, super-resolution, and anomaly detection. Their probabilistic formulation also enables explicit uncertainty quantification—an essential property for safe and interpretable clinical AI.

A notable example is [88], which leverages score-based diffusion to address ill-posed tasks such as accelerated MRI reconstruction and low-dose CT denoising. By integrating the learned score function $\nabla_x \log p_t(x)$ with data-consistency constraints, the method achieves both high perceptual quality and faithful anatomical preservation, outperforming compressed sensing and GAN-based baselines. Another representative work

is the 3D extension of Denoising Diffusion Probabilistic Models (DDPM) by Khader et al. [46], tailored for volumetric medical image generation. This approach adapts DDPM to operate directly on 3D voxel grids, synthesizing high-resolution MRI and CT volumes with preserved spatial coherence. A patch-based training strategy mitigates memory constraints, while conditional generation based on modality or anatomical labels enhances flexibility. Quantitative evaluations demonstrate superior structural fidelity over state-of-the-art GAN methods, highlighting the potential of diffusion models for realistic 3D data augmentation and rare case simulation.

Recent advances, such as latent diffusion [76], integrate variational autoencoders to perform generation in a compressed latent space, substantially reducing computational demands while maintaining synthesis quality. These developments further enhance the practicality of diffusion-based approaches in clinical workflows, combining scalability, realism, and uncertainty awareness.

## 2.2 Large Language Models

### 2.2.1 Medcial Large Language Models

Medical Large Language Models (Medical LLMs) are specialized adaptations of general-purpose large language models designed to process, understand, and generate medical information across text, image, and multimodal inputs. Built on transformer-based architectures, these models are pre-trained on vast corpora of biomedical literature, clinical notes, radiology reports, and sometimes multimodal datasets that pair medical images with text. The resulting language representations capture both domain-specific terminology and clinical reasoning patterns, enabling applications in automated report generation, medical question answering, clinical decision support, and patient education.

Unlike general LLMs, Medical LLMs incorporate domain adaptation strategies such as continual pre-training on curated medical text corpora, instruction tuning with clinician-authored prompts, and alignment

with clinical guidelines to improve factual accuracy and reduce hallucination. In multimodal settings, these models integrate medical image encoders (e.g., CNNs, Vision Transformers) with language decoders, enabling end-to-end reasoning over both visual and textual inputs—a critical capability for radiology and pathology applications.

Recent advances in medical large language models (LLMs) have been propelled by adapting general-purpose vision–language architectures to the clinical domain. LLaVA-Med [55] extends the Large Language and Vision Assistant (LLaVA) framework by integrating a medical image encoder with a domain-adapted LLaMA-based decoder. The model undergoes instruction tuning on curated medical image–text pairs and clinical dialogues, enabling it to perform a wide range of tasks, including medical visual question answering (Med-VQA), report summarization, and image–report cross-referencing. Domain-specific fine-tuning allows LLaVA-Med to achieve substantial improvements over the general-purpose LLaVA in both factual accuracy and clinical relevance. Similarly, XrayGPT [91] focuses on chest radiography by coupling a CLIP-based image encoder with a medical instruction-tuned GPT-4–style language model. Trained on large-scale chest X-ray–report datasets, XrayGPT is tailored for radiology-specific tasks such as structured report generation, differential diagnosis suggestion, and finding–impression alignment. The model incorporates disease-aware attention mechanisms to enhance sensitivity to subtle pathological patterns, leading to superior performance on benchmark datasets such as MIMIC-CXR and OpenI.

### 2.2.2 Instruction Tuning

Instruction tuning [107] has emerged as a powerful paradigm for enhancing the generalization and adaptability of large multimodal models. The core idea is to teach models how to follow diverse natural language instructions by training them on a broad collection of instruction-response pairs. Unlike traditional supervised learning that relies on rigid task-specific annotations, instruction tuning offers a flexible and scalable

framework for multi-task learning. Recent works such as Alpaca, Vicuna, and instruction-tuned GPT-4 have demonstrated impressive performance across a wide range of tasks, including question answering, reasoning, and content generation. In the medical domain, this paradigm is especially valuable, as it allows models to better align with the dynamic nature of clinical workflows, where task instructions vary based on diagnostic goals and user intent.

In medical imaging, where key tasks include report generation, VQA, and image synthesis, instruction tuning facilitates a unified language-driven interface to perform diverse functions. By training models to follow task-specific instructions—e.g., "generate a frontal chest X-ray showing mild cardiomegaly"—they can dynamically adapt to new tasks without requiring specialized heads or architectures. This approach is particularly effective in data-scarce scenarios, where high-quality domain-specific instructions can substitute for large-scale labeled datasets.

Building on this paradigm, MedXChat introduces instruction tuning with custom-designed special tokens (<Xray> and </Xray>) and leverages GPT-4 to generate rich instruction-following data tailored for chest X-ray analysis and generation. In particular, we construct a dedicated instruction dataset for the text-to-CXR image synthesis task, which includes domain-specific dialogues, clinical prompts, and corresponding model responses. The unified LLM is fine-tuned with Low-Rank Adaptation (LoRA), enabling efficient and scalable multi-task learning across modalities.

For text-to-image synthesis, we treat the content enclosed in <Xray> ... </Xray> as an embedded prompt for image generation. During training, the model learns to associate these structured prompts with corresponding radiological outputs. At inference time, the extracted prompt is passed to a Stable Diffusion model, which generates high-fidelity CXR images aligned with the clinical intent. This mechanism allows for controllable and explainable image synthesis guided by natural language. Unlike prior models such as LLMCXR and UniXGen, which lack explicit

prompt-grounded generation pathways, MedXChat enables a more semantically aligned and flexible multimodal pipeline for real-world medical scenarios.

### 2.2.3 Prompt Engineering

Prompt Engineering has emerged as a pivotal technique in aligning LLMs with downstream tasks. Brown et al. [6] demonstrated that LLMs, such as GPT-3, can perform surprisingly well on a wide range of tasks by conditioning on task-specific prompts without additional fine-tuning. Building on this foundation. Liu et al.[61] provided a systematic survey that categorized prompt-based learning into discrete prompting, soft prompting, and instruction-based prompting, revealing its potential to bridge the gap between pre-trained LLMs and domain-specific applications.

Inspired by these findings, we adopt prompt engineering to improve medical multi-label classification within our LLM framework. Specifically, we design structured prompt templates corresponding to 14 common thoracic disease categories in chest X-rays. To construct a high-quality instruction dataset, we leverage GPT-4 Turbo [66] to generate diverse and clinically relevant instruction–response pairs based on our domain-specific prompt templates. This instruction data is then used to fine-tune our LLM, enabling it to understand clinical classification tasks and accurately map visual input to discrete diagnostic categories. This approach allows for flexible integration of clinical intent into model behavior and enhances the interpretability and controllability of classification predictions.

### 2.2.4 Chain of Thought

Chain of Thought (CoT) reasoning has emerged as a crucial approach for enhancing the reasoning capability of LLMs. Wei et al. [108] first demonstrated that encouraging models to generate intermediate reasoning steps before producing the final answer significantly improves performance on arithmetic, commonsense, and symbolic reasoning tasks.

Subsequent work, such as Kojima et al. [50], revealed that even zero-shot CoT prompting can unlock hidden reasoning abilities in LLMs by appending simple trigger phrases like “Let’s think step by step".

Building on these insights, we adopt chain-of-thought prompting within our medical-domain framework. Specifically, we design a tailored reward function for medical LLM classification tasks. The reward comprises two components: (1) an accuracy reward that evaluates correctness across 14 diagnostic categories, and (2) a format reward that enforces the generation of both the chain-of-thought reasoning process and the final answer in the expected structure. Through this chain-of-thought–guided training, the model achieves improved performance in medical classification, demonstrating enhanced reasoning ability and prediction accuracy.

## 2.3 Medical Tasks

### 2.3.1 Bone Suppression

The bone suppression task involves removing bony structures from CXRs to reconstruct bone-free images [71]. This process enhances radiologists’ ability to detect lung diseases and improves the accuracy of deep learning–based lung disease classification by reducing rib artifacts and increasing the visibility of soft tissues. A conventional approach for bone suppression is dual-energy subtraction (DES) radiography, which acquires two X-ray images at different energy levels and subtracts them to separate bone from soft tissue. However, DES requires specialized hardware, imposes strict acquisition conditions, and exposes patients to increased radiation.

With the rapid development of artificial intelligence, deep learning–based methods—both supervised [89, 124, 29, 9] and unsupervised [24, 56]—are increasingly replacing DES for bone suppression. In supervised approaches, for example, the CNN-based Auto-Encoder [29] progressively downsamples and upsamples images to remove bones, while the IEDSR framework [58] enhances pneumothorax segmentation accuracy

by preprocessing CXRs with bone suppression to improve lung structure visibility. Similarly, Dilated cGAN [124] applies dilated conditional convolutions to capture multi-scale context and enforce semantic consistency, effectively suppressing bone structures.

Unsupervised methods also show promise. Pix2pix-MTdG [24] employs a pix2pix-based image-to-image translation network to jointly perform organ segmentation and bone suppression, while LoG-DRR [56] leverages domain knowledge from unpaired CT images via unsupervised domain adaptation. More recently, UDA [31] demonstrated cross-species bone suppression by transferring knowledge from labeled human CXRs to unlabeled animal CXRs through adversarial learning and feature alignment, preserving anatomical details without manual annotations.

Although these methods demonstrate the clinical value of bone suppression in disease classification, many existing models still face limitations in architectural design and output quality, leaving room for further improvements in both visual fidelity and diagnostic reliability.

### 2.3.2 Report Generation

Radiology report generation aims to automatically produce structured diagnostic reports from medical images, particularly CXRs, thereby reducing radiologists' workload and improving reporting consistency. The task requires accurately extracting visual features, modeling their relationship with medical findings, and generating coherent textual descriptions that capture both global context and localized abnormalities. Traditional CNN–RNN frameworks have struggled with long-range dependency modeling and clinical factual accuracy, motivating recent work that leverages Transformer-based architectures, cross-modal alignment, and external knowledge integration to enhance both linguistic fluency and clinical fidelity.

Representative methods include the conditioned Transformer [2], which injects structured clinical labels into the decoder to guide attention toward disease-relevant content, improving both BLEU scores and

clinical observation accuracy. Chen et al.[14] proposed a memory-driven Transformer that retrieves prototypical image–text patterns from an external memory, reducing repetitive sentences and improving long-report coherence on MIMIC-CXR. Wang et al.[101] introduced the Cross-Modal Prototype Driven Network (CMPD-Net), which learns semantic prototypes aligned across modalities to better capture disease-specific features, yielding superior performance in multi-condition scenarios compared to sequential or retrieval-based baselines.

Recent progress in medical LLMs has further advanced radiology report generation by leveraging powerful vision–language architectures adapted to the clinical domain. LLaVA-Med [55] extends the Large Language and Vision Assistant framework with a medical image encoder and a domain-adapted LLaMA-based decoder, instruction-tuned on curated medical image–text pairs and clinical dialogues. This enables it to generate reports with higher factual accuracy and richer clinical context compared to general-purpose models. Similarly, XrayGPT [91] focuses on chest radiography by combining a CLIP-based visual encoder with a medical instruction-tuned GPT-style language model. Optimized on large-scale chest X-ray–report datasets such as MIMIC-CXR, it incorporates disease-aware attention to improve recognition of subtle pathological patterns, leading to more accurate and clinically coherent reports.

### 2.3.3 Medical Visual Question Answering

Medical Visual Question Answering (Med-VQA) aims to automatically answer natural language questions about medical images, supporting applications such as clinical decision support, education, and patient communication. Compared with general-domain VQA, Med-VQA requires fine-grained visual understanding of domain-specific modalities (e.g., chest X-rays, CT, MRI, histopathology) and precise alignment with medical terminology and diagnostic reasoning. The task is challenging due to limited annotated datasets, high variability in imaging appearance, and the need for factual correctness in generated answers. Recent research has leveraged multimodal representation learning, external medical knowledge bases, and cross-modal attention mechanisms to

enhance both answer accuracy and clinical interpretability.

Among representative approaches, MedFuseNet [81] introduces a multimodal fusion network that integrates visual features from a convolutional backbone with textual embeddings from a language model via a co-attention mechanism, capturing complex interactions between image regions and question semantics to improve reasoning over clinically relevant areas. CGMVQA [74] (Coarse-to-Grained Multi-modal VQA) adopts a hierarchical reasoning strategy, first performing coarse localization of relevant anatomical structures and then refining attention to fine-grained pathological regions. By combining global and local context modeling, CGMVQA demonstrates superior performance on datasets like VQA-RAD and PathVQA, particularly in handling multi-step reasoning and rare disease cases.

More recently, large language models (LLMs) have advanced Med-VQA by enabling open-ended, instruction-following capabilities and improved contextual reasoning. RadFM [110] is a radiology-focused vision– language large model that integrates a high-capacity medical image encoder with an instruction-tuned LLM, enabling it to perform VQA, report summarization, and finding–impression alignment with disease-aware attention to subtle pathological features. LLaVA-Med [55] extends the general-purpose LLaVA framework to the medical domain by instruction-tuning on curated image–question–answer triples and clinical conversations. This domain adaptation significantly improves factual accuracy and clinical relevance over general-purpose models, enabling LLaVA-Med to handle diverse Med-VQA scenarios, including multi-image reasoning and report-grounded question answering.

### 2.3.4 Disease Classification

Disease classification in medical imaging, particularly CXR disease classification, aims to automatically identify the presence or absence of specific pathologies from radiographic images. This task is crucial for screening, triage, and aiding diagnostic workflows, requiring models to accurately capture both global anatomical patterns and localized pathological features. Early approaches were dominated by convolutional neural networks (CNNs), which extract hierarchical spatial and semantic representations to achieve high classification accuracy. For instance, mCNN [72] leverages multi-scale convolutional layers to model local details and global structures, demonstrating strong performance across multiple CXR benchmarks. More recently, transformer-based architectures such as CTransCNN [112] have been introduced to model long-range dependencies and enhance global context understanding, further boosting multi-label classification accuracy. These models are typically trained in a supervised manner on large-scale annotated datasets, producing only categorical labels without providing detailed explanations or fine-grained visual reasoning.

Recent advancements in LLMs have expanded disease classification beyond simple label prediction toward richer, more interpretable vision–language reasoning. Models such as PaLM-E [23], Med-PaLM M [94], and LLaVA-Med [55] integrate high-capacity language models with powerful visual encoders to process radiological images and output diagnostic findings in natural language. Med-PaLM M fine-tunes the PaLM-E framework with curated medical instruction datasets, achieving strong zero-shot and few-shot performance on both classification and question-answering benchmarks, while improving clinical factuality. Similarly, LLaVA-Med extends the LLaVA framework to the medical domain by instruction-tuning on paired medical images and reports, enabling dialogue-based interpretation with enhanced medical vocabulary and domain knowledge. These LLM-driven approaches not only match or surpass traditional supervised classifiers in accuracy but also provide transparent, context-rich diagnostic outputs that align more closely with clinical reporting practices.

### 2.3.5 Visual Grounding

Medical visual grounding has emerged as a key task for enhancing the explainability of AI-driven diagnostic systems by explicitly linking textual descriptions of clinical findings to their corresponding regions in medical images. Unlike conventional classification or captioning, which outputs a label or a sentence without spatial context, visual grounding provides localized visual evidence that clinicians can directly verify. This spatial linkage improves trust, supports second-opinion validation, and ensures that automated predictions are backed by concrete image-based reasoning. The task is especially critical in high-stakes domains such as radiology, where misinterpretation of subtle lesions can lead to severe diagnostic errors.

Early CNN-based methods, such as DACR [65], used class activation maps to highlight coarse regions relevant to the predicted findings, offering an initial step toward interpretability but with limited localization precision. More advanced Transformer-based architectures, including MedRPG [12] and SeqTR [7], integrate cross-modal attention mechanisms to align textual descriptions with specific image regions, enabling fine-grained lesion localization through bounding boxes or segmentation masks. These approaches not only improve alignment accuracy but also adapt to diverse modalities and pathologies, paving the way for clinically deployable systems where transparent, region-level justifications are essential for radiologists' decision-making and for effective patient–doctor communication.

### 2.3.6 Text-to-CXR Synthesis

Text-to-CXR synthesis aims to generate realistic CXR images directly from textual inputs such as radiology reports, diagnostic summaries, or disease labels. Unlike natural image generation, this task must ensure anatomical correctness while faithfully representing disease-specific features, avoiding the creation of hallucinated or erroneous structures that could

mislead clinical decisions. Recent approaches incorporate Vector Quantization (VQ) into the generative framework, such as VQGAN or VQ-VAE, which encode high-resolution medical images into a discrete latent space and then condition the generation process on text. This design improves structural consistency, enhances computational efficiency, and facilitates multimodal alignment, thereby achieving a balance between visual realism and clinical fidelity.

Representative works include UniXGen and LLM-CXR. UniXGen employs a unified multimodal generation framework that encodes CXRs into a discrete latent space and uses cross-attention to align textual descriptions with the VQ-encoded anatomical structures, enabling accurate synthesis from reports to images. LLM-CXR maps clinical semantic embeddings generated by a large language model into the discrete codebook space of a VQGAN, with the generator reconstructing high-fidelity CXR images during decoding. Both methods leverage the discrete representation benefits of VQ to enhance structural fidelity and disease specificity, achieving superior performance on benchmarks such as MIMIC-CXR compared to pixel-space or continuous-latent approaches.

## 2.4 Public Medical Dataset

This section provides an overview of experimental datasets commonly used to evaluate robust training methods in medical image analysis and synthesis. We begin by introducing widely used datasets in the medical imaging domain, highlighting their significance and practical applications. Specifically, we present public datasets used for bone suppression and various tasks related to unified multimodal large models in medical imaging.

### 2.4.1 Datasets for Bone Suppression

To evaluate the effectiveness of bone suppression techniques in medical imaging, we selected several CXR datasets that are highly relevant for assessing bone suppression image quality. These include the Bone Shadow Suppression X-ray dataset [83], the largest collection of paired

chest X-rays before and after bone suppression. For the downstream classification task, the Pneumonia Chest X-rays dataset [45] provides bone-suppressed CXRs containing both normal and pneumonia samples. The NIH Chest X-ray dataset [104] includes image labels for 14 distinct diseases extracted from associated radiology reports, supporting multi-label classification tasks. For downstream segmentation, the Chest X-ray Dataset for Tuberculosis Segmentation [41] consists of CXR images specifically curated for tuberculosis detection.

- **Bone Shadow Suppression X-ray Dataset** [83]: This dataset is the largest publicly available collection of paired chest X-rays before and after bone suppression, comprising 241 image pairs. We randomly split the dataset into 193 pairs (80%) for training and 48 pairs (20%) for testing. It serves as the primary benchmark for evaluating bone suppression performance.
- **Pneumonia Chest X-rays Dataset** [45]: This dataset includes 5,450 chest X-ray images, consisting of 1,575 normal and 3,875 pneumonia cases. Following the official data split, 5,216 images were used for training and 624 for testing. This dataset is employed to evaluate the impact of bone suppression on downstream pneumonia classification performance.
- **NIH ChestX-ray14 Dataset** [104]: Containing 112,120 frontal-view chest X-ray images from 30,805 patients, this dataset provides image-level annotations for 14 common thoracic diseases, extracted using natural language processing techniques from corresponding radiology reports. Each image may be associated with multiple disease labels. For our experiments, we used the standard data split with 86,524 images for training and 25,596 for testing.
- **Chest X-ray Dataset for Tuberculosis Segmentation** [41]: This dataset contains 704 chest X-ray images sourced from two collections: the Montgomery County Chest X-ray Set (USA) and the Shenzhen Chest X-ray Set (China). Each image is annotated with lung masks for segmentation. We randomly divided the dataset into 563 training

and 141 testing samples, which are used to assess the effectiveness of bone suppression in downstream tuberculosis segmentation tasks.

### 2.4.2 Datasets for Medical LLMs

For the evaluation of Medical LLMs, we conduct a comprehensive assessment using publicly available datasets that are widely recognized and commonly used in current research. Specifically, we focus on the MIMIC-CXR dataset, which supports multiple tasks in chest X-ray analysis, including report generation, VQA, classification, and visual grounding. This dataset provides a challenging benchmark for evaluating multimodal model performance across a range of clinically relevant tasks. Additionally, we incorporate the VinDr-CXR dataset, which focuses on visual grounding in over 100,000 chest X-ray scans and offers rich radiological annotations. By evaluating our methods on these diverse and challenging datasets, we demonstrate the robustness and effectiveness of our approach in handling various tasks in medical imaging, such as report generation, VQA, classification, and visual grounding.

- **MIMIC-CXR** [43] is the largest publicly available chest radiograph dataset, comprising over 370,000 images linked to free-text radiology reports. It covers 14 common thoracic disease categories and is widely used for a variety of multimodal learning tasks, including report generation, VQA, image classification, and visual grounding. In our study, we followed the official dataset split protocol to maintain consistency and ensure fair benchmarking. Specifically, we used 270,790 image-report pairs for training, 2,130 for validation, and 3,858 for testing. The dataset's scale and rich annotations make it a cornerstone for evaluating unified medical multimodal models.

- **VinDr-CXR** [64] is a large-scale chest X-ray dataset collected from multiple Vietnamese hospitals, comprising more than 100,000 CXR

scans. Among them, 18,000 images have been meticulously annotated by 17 board-certified radiologists with both local (bounding box-level) and global (image-level) labels across 22 critical findings. Structured into 15,000 training images and 3,000 test images, VinDr-CXR is particularly suitable for visual grounding and localization tasks. Additionally, it employs a customized DICOM image labeling platform to ensure annotation consistency and clinical reliability. This dataset serves as a valuable resource for training and evaluating models on fine-grained disease detection and spatial reasoning in chest radiology.

# Chapter 3

# Improving CXR Bone Suppression by Exploiting Domain-level and Instance-level Information

In this chapter, we explore the task of suppressing bone structures in CXRs using a limited paired dataset. Our approach focuses on eliminating bone-related interference while preserving critical anatomical details to enhance the performance of downstream clinical tasks. Specifically, we introduce a two-stage learning framework that jointly leverages domain-level and instance-level information. In the first stage, we learn a domain-specific prior through a Multi-head Codebook Attention (MCA) module embedded in a Vision Transformer (ViT)-based generative network. This enables the model to encode representative bone-suppressed features across the dataset. In the second stage, we integrate a Cross-Covariance Attention Blocks (CABs) network to extract fine-grained instance-level features from individual bone shadowed CXRs. These features are then used to guide the reconstruction of bone suppressed images. Extensive experiments on multiple datasets demonstrate that our method not only achieves superior image quality in terms of PSNR, SSIM, and MSE, but also significantly boosts performance in downstream clinical tasks such as pneumonia classification and tuberculosis segmentation. These results confirm the robustness and clinical value of our approach in real-world medical imaging applications.

## 3.1 Introduction

CXRs are a widely adopted imaging modality for pulmonary disease screening due to their speed, accessibility, and noninvasive nature. However, as a two-dimensional projection without depth information, CXRs suffer from anatomical overlaps, particularly bone structures like ribs and clavicles that obscure soft tissue details. This poses significant challenges for both radiologists and computer-aided diagnostic (CADx) systems, especially in detecting subtle lung abnormalities [40]. To address this, bone suppression (BS) techniques have been developed to reduce the visual interference of bones and improve diagnostic clarity.

Traditional BS methods, such as dual-energy subtraction (DES)[51], acquire X-rays at two energy levels to differentiate between soft tissue and bone. While effective, DES requires specialized and costly hardware, introduces additional radiation exposure[84], and is impractical for widespread clinical deployment. In recent years, deep learning has emerged as a promising alternative for bone suppression. Early works—both supervised [89, 29, 9, 124, 58] and unsupervised [24, 56, 31]—typically rely on CNN-based architectures such as U-Net [38] to learn mappings from bone-shadowed to boneless images. These models offer fast inference and cost-efficiency, outperforming traditional methods in both practicality and performance.

Following advancements in generative modeling, recent studies have explored Auto-Encoders [29, 52] and Generative Adversarial Networks (GANs) [125, 16, 124, 76] for bone suppression, inspired by the success of image-to-image translation tasks. However, these methods predominantly focus on instance-level information—capturing features specific to individual images—while overlooking the fact that CXRs, despite some variability, share domain-level commonalities. Neglecting this shared structural prior limits the generalizability and robustness of bone suppression models.

To address this gap, we propose a novel framework that jointly leverages domain-level and instance-level information for bone suppression. Central to our approach is the Multi-head Codebook Attention (MCA)

module, which replaces the traditional single-head vector quantization in VQ-based models [25]. MCA decomposes the latent representation into multiple heads, each attending to different aspects of the feature space, thereby enabling richer and more nuanced domain-level modeling. Unlike "hard" codebooks that rely on non-differentiable nearest neighbor searches, our "soft" codebook formulates feature encoding as a learnable attention-weighted combination over all code items. This not only preserves continuity in the representation space, but also enhances feature fidelity and reduces quantization artifacts.

Our framework adopts a two-stage optimization process. In Stage I, we learn a robust domain prior from bone-suppressed CXRs using a transformer-based auto-encoder and MCA. In Stage II, we integrate Cross-Covariance Attention Blocks (CABs) to extract instance-level information from bone-shadowed CXRs. The CAB network generates attention maps that highlight diagnostically relevant regions, which are then refined and encoded using a fine-tuned Vision Transformer (ViT) encoder. The resulting embeddings are decoded using the pretrained decoder from Stage I to synthesize high-quality boneless images.

We validate our method through extensive experiments on multiple public datasets. Using the Bone Shadow Suppression dataset [83] as the primary training resource and ground truth derived from DES imaging, we achieve state-of-the-art results in bone removal quality. Moreover, when applied to downstream tasks such as pneumonia classification [45] and NIH disease classification [104], our bone-suppressed CXRs lead to notable improvements in diagnostic performance, outperforming baselines including Stable Diffusion [76], RQ-VAE [52], and conventional BS methods like Auto-Encoder [29], IEDSR [58], and Dilated cGAN [124]. These results demonstrate that our joint modeling of domain and instance features not only produces more faithful boneless images but also enhances clinical utility, underscoring the potential of our method in real-world radiological workflows.

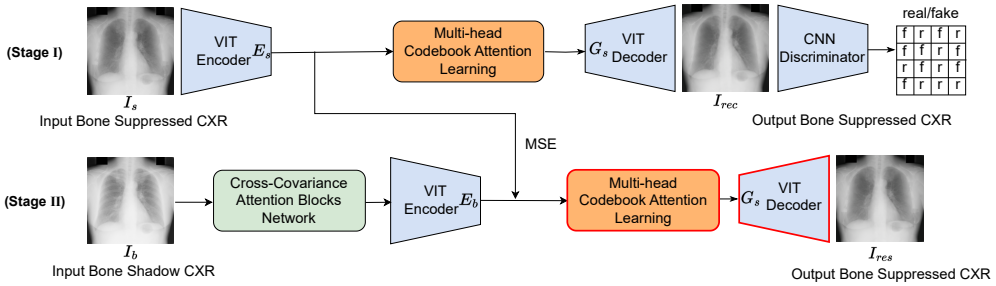



**Figure 3.1:** An overview of our two-stage learning-based bone-suppression framework: (a) In stage I, we first build a domain-level representative dictionary of boneless CXRs by adopting transformer encoder-decoder and learning a Multi-head Codebook Attention (MCA) in a self-reconstruction manner. (b) In the second stage, we leverage Cross-Covariance Attention Blocks (CABs) Network to enhance instance-level input quality for each CXR and generate its corresponding boneless CXR output based on the learned dictionary and decoder. Note, the MCA module, ViT encoder, and decoder are pre-trained in Stage I. In Stage II, the newly introduced CAB network is trained, and the ViT encoder is fine-tuned, while the MCA module and ViT decoder remain frozen (highlighted in red). For inference, only Stage II model is used.

## 3.2 Method

As illustrated in Fig. 6.1, our bone suppression framework consists of two sequential stages. In Stage I, we employ a transformer-based autoencoder architecture integrated with our proposed MCA module. This stage focuses on reconstructing boneless CXRs from input boneless images, enabling the model to effectively learn a domain-level codebook that captures shared structural priors through an attention-based mechanism [97].

In Stage II, we apply the pre-trained MCA module from Stage I in conjunction with a CABs network to process input bone-shadowed CXRs. The CABs network generates instance-aware attention maps, which are used to refine image representations and guide the suppression of bone structures more precisely. Both stages are jointly optimized during training; however, only Stage II is utilized during inference for efficient generation of bone-suppressed CXRs.

### 3.2.1 Multi-head Codebook Attention Learning Module (Stage I)

#### Multi-head Codebook

Given an input CXR image $I_s \in \mathbb{R}^{C\times H\times W}$, we first extract its latent feature map $\hat{\mathbf{Z}}s \in \mathbb{R}^{h\times w\times d}$ using the encoder $\boldsymbol{E}s$. Each $d$-dimensional vector $\hat{z}k$ in the feature map is then quantized by a learnable codebook to obtain the discrete representation $\mathbf{Z}s$. Unlike the conventional single-head codebook used in VQGAN [25], which defines a single codebook $\mathcal{C}SH = ci \in \mathbb{R}^d i = 0^{N-1}$ with $N$ $d$-dimensional code vectors, our approach introduces a novel multi-head codebook design $\mathcal{C}MH = \mathcal{C}^m m = 1^M$ (illustrated in Fig. 3.2 (b)) that consists of $M$ independent sub-codebooks (i.e., attention heads). Each sub-codebook $\mathcal{C}^m = c^m i \in \mathbb{R}^{\frac{d}{M}}{}_{i=1}^{N}$ contains $N$ vectors of dimension $\frac{d}{M}$, allowing each head to focus on a specific subspace of the feature representation.

To quantize a feature vector $\hat{z}_k$, we first split it evenly across channels into $M$ segments: $\hat{z}_k = \hat{z}_k^0 \oplus \hat{z}_k^1 \oplus \cdots \oplus \hat{z}_k^{M-1}$, where $\oplus$ denotes concatenation. Each segment $\hat{z}_k^m \in \mathbb{R}^{\frac{d}{M}}$ is independently quantized by its corresponding sub-codebook $\mathcal{C}^m$ by finding the nearest code vector:

$$z_k^m = \mathbf{q}(\hat{z}k^m) := \arg\min c_j^m \in \mathcal{C}^m \left|\hat{z}k^m - cj^m\right|, \quad z_k^m \in \mathbb{R}^{\frac{d}{M}}. \tag{3.1}$$

The final quantized representation $z_k$ is formed by concatenating all the selected sub-codebook vectors across the $M$ heads:

$$z_k = z_k^0 \oplus z_k^1 \oplus \cdots \oplus z_k^{M-1}. \tag{3.2}$$

Compared to the vanilla single-head approach, where each $\hat{z}_k$ can only be mapped to one of $N$ possible codes, our multi-head strategy increases the total number of encoding combinations to $N^M$. This exponentially enlarges the representational capacity of the quantizer without increasing the number of parameters, as each head operates in a lower-dimensional space. As a result, our design enables the model to capture

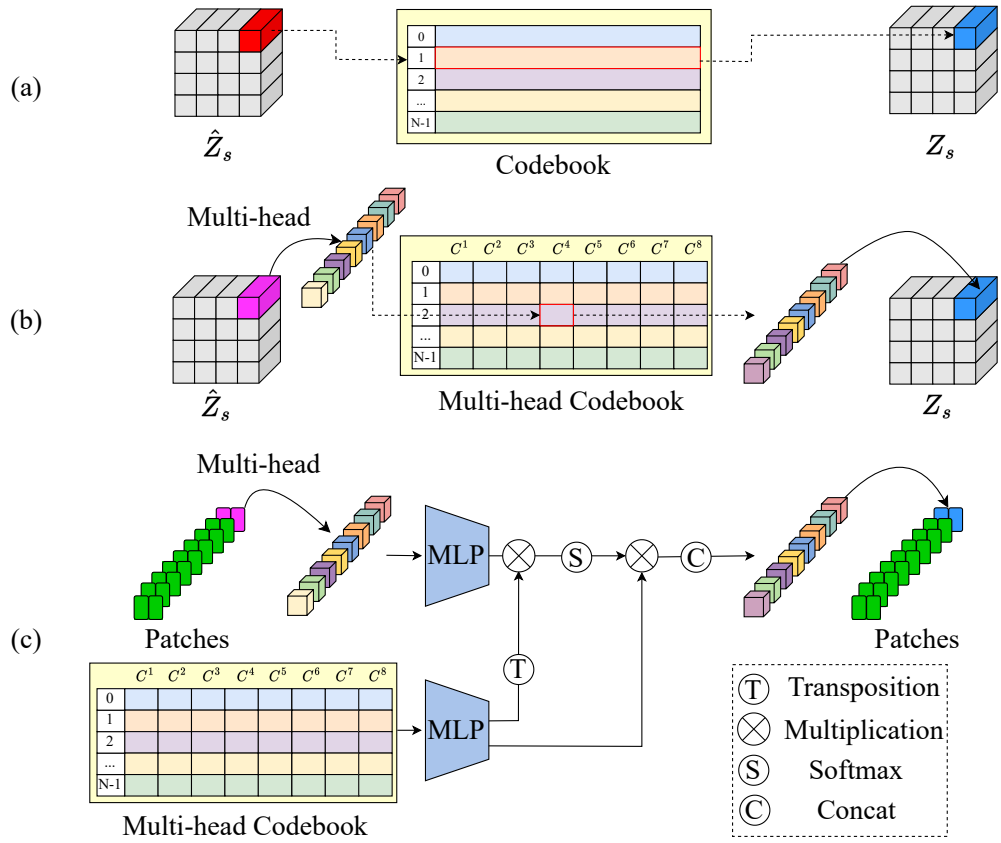



**Figure 3.2:** Comparison between the vanilla codebook learning scheme (a), multi-head codebook learning scheme (b) and our proposed multi-head codebook attention learning module (c).

more diverse and fine-grained patterns in the latent space, which is especially beneficial in the medical imaging domain where subtle variations (e.g., lesion texture, rib patterns) are clinically important.

### Codebook Attention

Traditional vector quantization, whether in the vanilla single-head setting or our multi-head codebook scheme, maps input features to the nearest codeword in a discrete codebook. While effective in compressing representations, this "hard" assignment strategy suffers from limited flexibility and may result in visually similar image patches being projected to the same code, leading to information loss and suboptimal reconstruction. To address these limitations, we propose a learnable *codebook attention* module (see Fig. 3.2 (c)), which replaces hard quantization

with a more expressive and differentiable soft encoding mechanism.

Given a multi-head codebook $\mathcal{C}MH \in \mathbb{R}^{H\times E\times D}$, where $H$ is the number of attention heads, $E$ is the number of codewords per sub- codebook, and $D$ is the dimensionality of each codeword, we interpret the codebook as both key and value embeddings. In parallel, the encoded feature map is split into $H$ heads to form the query embeddings $\hat{\mathbf{Z}}MH \in \mathbb{R}^{H\times h\times w\times \frac{d}{H}}$, where $d$ is the channel dimension of the original latent feature. For each spatial location, we compute a cross-attention score between the query and each codeword, enabling a soft weighted combination over all entries in the codebook:

$$\text{Att}(\mathbf{Q}\text{codebook}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\text{codebook}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3.3}$$

where $\mathbf{Q}$codebook denotes the query vectors derived from the input features, while $\mathbf{K}$ and $\mathbf{V}$ are the key and value vectors taken from the multi-head codebook itself. As noted, $\mathbf{Q}$codebook is computed from image features, while the codebook remains fixed during the attention computation[1].

This soft-attention-based code assignment mechanism offers several advantages over hard vector quantization. First, it transforms the inherently discrete representation space into a continuous one, which improves gradient flow and reduces quantization error. Second, by aggregating information from all codewords—rather than selecting a single nearest code—our method significantly enriches the representational capacity of the model without increasing memory consumption. Third, the entire encoding process becomes fully differentiable, eliminating the need for the non-differentiable nearest neighbor search and the stop-gradient trick used in prior works such as VQGAN [25].

Overall, the proposed codebook attention module provides a flexible and powerful alternative to conventional quantization, enabling the

[1] $\mathbf{Q}_{\text{codebook}}$ is extracted from the image feature embedding, while $\mathbf{K}$ and $\mathbf{V}$ are the shared codewords from the multi-head codebook.

model to learn more nuanced and semantically meaningful representations. In the context of bone suppression, this translates to improved fidelity in soft tissue reconstruction and more robust generalization across diverse anatomical patterns, as further evidenced in our experimental results.

**Vision Transformer Encoder and Decoder**

To capture more expressive and instance-specific features from CXRs, we employ a Vision Transformer (ViT) [22] as both the encoder and decoder in our architecture. Unlike convolutional neural networks that are inherently local in receptive field, ViTs leverage global self-attention mechanisms, making them well-suited for modeling long-range dependencies and subtle pathological patterns in medical images.

Given an input image $I_s \in \mathbb{R}^{C\times H\times W}$, we first divide it into non-overlapping square patches of size $P \times P$. Each patch is then flattened into a 1D vector, resulting in a sequence of $N_p = \frac{HW}{P^2}$ tokens. This yields a matrix $I_p \in \mathbb{R}^{N_p \times P^2 C}$, where each row corresponds to a flattened patch vector. This process effectively transforms the spatial structure of the image into a sequential form, enabling the application of transformer-based architectures.

To project each patch token into a latent feature space, a trainable linear projection is applied: $\mathbf{E}_p = \mathbf{I}_p \mathbf{W}_p{}^\top$, where $\mathbf{W}_p \in \mathbb{R}^{D\times P^2 C}$ maps the input to an embedding dimension $D$, producing $\mathbf{E}p \in \mathbb{R}^{N_p \times D}$. Since transformers lack inherent spatial awareness, a learnable 1D positional embedding $\mathbf{E}pos \in \mathbb{R}^{N_p \times D}$ is added to encode spatial information:

$$\mathbf{E} = \mathbf{E}p + \mathbf{E}pos. \tag{3.4}$$

The resulting sequence $\mathbf{E}$ is then fed into the transformer encoder.

The encoder comprises 12 stacked transformer blocks, each consisting of a Multi-Head Self-Attention (MSA) layer, a Feed-Forward Network (FFN), and Layer Normalization (LN) operations. Within each MSA layer, the input embeddings are linearly projected into queries ($\mathbf{Q}$),

keys (**K**), and values (**V**), and the attention scores are computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right), \tag{3.5}$$

where $d_k$ is the dimension of the query/key vectors. These attention weights are then used to produce the output:

$$\mathbf{F} = \mathbf{A}\mathbf{V}. \tag{3.6}$$

The attention output is followed by residual connections and layer normalization. The FFN then maps the representations through a two-layer MLP with GELU activation, increasing and reducing dimensionality to enable abstract feature learning. This structure allows the encoder to build rich contextual representations across the entire image.

The decoder mirrors the encoder in structure, also comprising 12 transformer blocks. The encoded sequence is first added with a new set of positional embeddings and then processed by the decoder layers. Finally, the output sequence is reshaped back into the 2D spatial format and passed through a convolutional projection head to reconstruct the output image (e.g., the bone-suppressed version of the original input). This symmetric design facilitates effective end-to-end learning of both encoding and generation processes in our model.

Overall, the use of ViT as encoder and decoder enables our framework to efficiently model global contextual information and extract high-quality instance-level representations, which are crucial for accurate bone suppression and soft-tissue enhancement in medical imaging.

#### Training Objectives

With the encoded feature map $\mathbf{Z}_s$ output from MCA module, the reconstructed boneless CXR can be reconstructed through the Decoder $G_s$: $I_{rec} = G_s(\mathbf{Z}_s)$. We basically follow the training objectives of VQGAN [25] to adopt the $L_2$ reconstruction loss $\mathcal{L}_{L2}$, adversarial loss $\mathcal{L}_{adv}$ [39], and

perceptual loss [103] $\mathcal{L}_{per}$ which balancing the GAN loss and the reconstruction loss as the supervision for image quality. In addition, the differentiable loss function $\mathcal{L}_{\text{VQ}}$ which is adopted in VQGAN to achieve end-to-end training:

$$\mathcal{L}_{\text{VQGAN}} = \|\text{sg}[\boldsymbol{E}_s(I_s)] - \mathbf{Z}_s\|_2^2 + \|\text{sg}\left[\mathbf{Z}_s\right] - \boldsymbol{E}_s(I_s)\|_2^2, \tag{3.7}$$

where sg denotes the stop-gradient operation. In our methodology, this operation is not applied, as our MCA mechanism is fully differentiable. The objective function is formulated as follows:

$$\mathcal{L}_{\text{VQ}} = \|\boldsymbol{E}_s(I_s) - \mathbf{Z}_s\|^2. \tag{3.8}$$

The total loss functions for our multi-head codebook attention learning procedure are described as follows:

$$\mathcal{L}_{MCA} = \mathcal{L}_{L2} + \mathcal{L}_{per} + \mathcal{L}_{VQ} + \lambda * \mathcal{L}_{adv}, \tag{3.9}$$

where $\lambda$ is adaptive weight [25].

### 3.2.2 Instance Feature Extraction and Bone Suppression (Stage II)

In the second stage of our framework, we introduce a Cross-Covariance Attention Block (CAB) network, which generates an informative attention map to emphasize clinically relevant regions within the CXR image. This spatially guided attention helps the subsequent Vision Transformer (ViT) encoder focus on key anatomical structures, thereby enabling the extraction of more discriminative instance-level features.

These refined features are then passed through the previously trained Multi-head Codebook Attention (MCA) module, where they are adaptively encoded by attending to the learned discrete representations in the codebook. This encoding process not only enriches the feature representation but also ensures consistency with the distribution learned in Stage I.

Finally, the encoded features are fed into the decoder $G_s$ trained in Stage I, which reconstructs the final bone-suppressed chest X-ray. It is important to note that during inference, only Stage II is activated, ensuring an efficient and lightweight deployment without sacrificing performance.

**Cross-Covariance Attention Blocks (CABs) Network**

To capture fine-grained structural details and enhance instance-level representation, we adopt a U-Net-style architecture built upon transformer blocks enhanced with CABs, following the design philosophy of Restormer [121]. This network serves as an effective backbone for extracting essential structural information from bone-shadowed CXRs.

As illustrated in Fig. 3.3, the input image of size $H \times W$ is first processed through a symmetric encoder-decoder framework. The encoder progressively downsamples the image to a compact feature representation of size $\frac{H}{8} \times \frac{W}{8} \times 384$, which is then restored back to its original spatial resolution using progressive upsampling in the decoder. At each scale within the encoder and decoder, multiple CABs are applied. These CABs employ a gated attention mechanism to operate across the channel dimension, effectively modeling long-range dependencies without relying on patch-based tokens as in conventional vision transformers.

Unlike traditional ViTs that rely on token embeddings and patch-wise attention, CABs compute cross-covariance attention over full- channel feature maps. This allows the model to capture global contextual information while preserving fine anatomical structures such as ribs, soft tissue contours, and background variations [79]. By integrating low-level texture details with high-level semantic features, CABs generate an informative attention map that emphasizes bone structures and relevant contextual cues, enabling better focus on task-critical regions while suppressing irrelevant background noise.

Supervised by our bone suppression objective, the CAB-enhanced feature maps act as a high-resolution "information map," which precisely highlights anatomical features related to bone shadows and soft

tissue. Unlike Restormer, which is tailored for generic image restoration tasks using complex multi-objective loss functions such as $\ell_1$ loss, our model is specifically optimized for bone suppression using a mean squared error (MSE) loss between the original and bone-suppressed images. This adaptation ensures that the CAB modules are finely tuned to capture rib-related visual patterns, aiding the subsequent suppression process.

Finally, the generated information map, rich in instance-level detail, is added back to the original image to emphasize key features. This combined representation is then fed into a transformer-based encoder, which extracts the final instance-level features that are subsequently processed by the MCA module for bone removal and image reconstruction.

**Bone Suppression**

Once the enhanced instance features are obtained from the CAB, we further extract high-level semantic representations using an additional Vision Transformer encoder $\boldsymbol{E}_b$. This encoder processes the CAB-enhanced input CXR with bone shadows and generates a feature map denoted as $\hat{\mathbf{Z}}_b$. However, since these features are extracted from bone-shadowed images, they inherently contain domain-specific noise that may hinder direct reconstruction via the decoder trained on boneless images.

To address this domain shift, we enforce a feature-level alignment between $\hat{\mathbf{Z}}_b$ and the boneless feature representation $\hat{\mathbf{Z}}s$ that was previously learned in Stage I. Specifically, we introduce a feature consistency loss using mean squared error (MSE) to minimize the distance between these two representations:

$$\mathcal{L}\mathrm{f} = \left|\hat{\mathbf{Z}}_s - \hat{\mathbf{Z}}_b\right|^2. \tag{3.10}$$

This loss encourages the feature embeddings from bone-shadowed inputs to align closely with their boneless counterparts, enabling the model to effectively perform bone suppression in a latent feature space.

Following this alignment step, we apply the Multi-head Codebook Attention (MCA) mechanism to discretize and encode the aligned feature map $\hat{\mathbf{Z}}_b$. Each feature vector in $\hat{\mathbf{Z}}_b$ is softly quantized via attention into a compact latent representation $\mathbf{Z}_b$, using the multi-head codebook that was learned during Stage I. This process preserves high-level semantics while removing bone-related visual patterns.

Finally, the encoded representation $\mathbf{Z}_b$ is passed through the decoder $G_s$—also pre-trained in Stage I—to synthesize the final bone-suppressed chest X-ray image. Notably, by reusing the decoder trained on clean boneless data, we ensure that the reconstructed image inherits structural fidelity and realism consistent with clean domain characteristics.

#### Training Objectives

We optimize only the CAB modules and the feature encoder $\boldsymbol{E}_b$ during Stage II, while the multi-head codebook and decoder are directly inherited from Stage I without further updates. As a result, the optimization objective in this stage is simplified and focuses solely on enforcing feature-level consistency between bone-shadowed and bone-suppressed representations. Specifically, we employ a single loss term—the feature similarity loss—defined as:$\mathcal{L}_{\text{f}} = \|\hat{\mathbf{Z}}_s - \hat{\mathbf{Z}}_b\|^2$, which encourages the features extracted from bone-shadowed CXRs to closely align with the corresponding boneless representations learned in the previous stage, thereby facilitating effective bone suppression through latent space alignment.

## 3.3 Experiment Settings

### 3.3.1 Implementation Details

We adopt ViT models with encoder and decoder components pre-trained on ImageNet [120] to extract and reconstruct features from CXR images. In Stage I, we fine-tune both the ViT encoder and decoder in conjunction with the proposed MCA module. In Stage II, we fine-tune a separate ViT

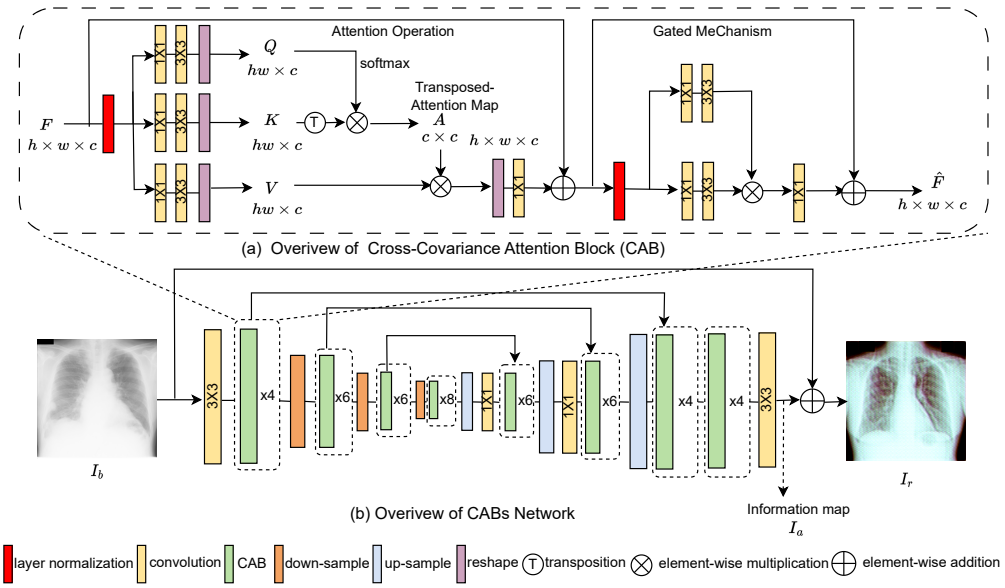



**Figure 3.3:** We employ several (a) cross-covariance attention blocks (CABs) modules in the (b) CAB network in Stage II, which is a UNet-style transformer network to extract the essential instance information map $I_a$ from the input bone shadow CXR $I_b$. We add $I_a$ with the highlighted information back to input CXR $I_b$ to produce the re-processed CXR $I_r$. We then feed this $I_r$ into the CNN encoders as in Fig. 1 of the main manuscript. For each CAB module, we adopt cross-covariance channel-wise attention operation as in [121] instead of vanilla token-wise attention operation to capture the important instance information.

encoder along with the Cross-Covariance Attention Block (CAB) module to perform instance-aware bone suppression.

For the codebook configuration, we set the total number of code vectors (i.e., codebook size) to 1024 and the embedding dimension to 256. The multi-head structure is composed of 8 attention heads, resulting in each head operating on a 32-dimensional subspace. This configuration preserves the overall embedding dimensionality and is found to offer a good trade-off between performance and complexity. We empirically evaluated configurations with 4, 8, and 16 heads, and determined that 8 heads yield the best balance in terms of reconstruction quality and computational efficiency.

Our models are trained using the Adam optimizer [47] on two NVIDIA RTX A5000 GPUs with a mini-batch size of 4 per GPU. The learning rate is fixed at $2 \times 10^{-4}$. We train the model for approximately 300 epochs during the self-generation phase (Stage I) and an additional 200 epochs during the bone suppression phase (Stage II).

### 3.3.2 Datasets Details

Our training set is constructed from the Bone Shadow Suppression X-ray dataset [83], which provides the largest publicly available collection of paired chest X-ray images before and after bone suppression. This dataset enables supervised learning for accurate rib structure removal. To further evaluate the clinical utility of our proposed bone suppression method, we conduct downstream classification experiments on two additional datasets: the Pneumonia Chest X-rays dataset [45] and the NIH ChestX-ray14 dataset [104]. Moreover, to assess its impact on segmentation tasks, we perform a comparative study for tuberculosis detection on the Chest X-ray Dataset for Tuberculosis Segmentation [41]. These evaluations demonstrate the generalizability and effectiveness of our approach in real-world diagnostic scenarios.

## 3.4 Evaluation and Discussion

In this section, we conduct a comprehensive evaluation of enhancement performance across multiple methods, considering not only overall image quality but also their effectiveness in key downstream medical tasks. Specifically, we assess performance in disease classification, tuberculosis segmentation and performance on higher-resolution CXRs. By jointly evaluating visual improvements and task-specific outcomes, we aim to offer a holistic understanding of how each method contributes to real-world medical image synthesis applications.

**Table 3.1:** Bone-suppression performance comparison on the X-ray bone suppression dataset [83].

| Method | PSNR↑ | SSIM↑ | MSE↓ |
|---|---|---|---|
| Stable Diffusion | 20.5 | 0.850 | $7.3 \times 10^{-3}$ |
| Star GAN2 | 24.6 | 0.871 | $5.2 \times 10^{-3}$ |
| Cycle GAN | 25.1 | 0.936 | $4.0 \times 10^{-3}$ |
| Pix2pix | 31.5 | 0.952 | $8.3 \times 10^{-4}$ |
| RQ VAE | 31.6 | 0.943 | $8.2 \times 10^{-4}$ |
| Auto-Encoder | 13.8 | 0.826 | $4.2 \times 10^{-2}$ |
| IEDSR | 23.9 | 0.959 | $5.1 \times 10^{-3}$ |
| Dilated cGAN | 32.3 | 0.942 | $7.5 \times 10^{-4}$ |
| **Ours** | **35.0** | **0.975** | $\mathbf{4.6 \times 10^{-4}}$ |

### 3.4.1 Image Quality Enhancement

For quantitative evaluation, we employ three standard metrics for image generation tasks: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [106], and Mean Squared Error (MSE) [78], to assess the quality of the generated bone-suppressed CXRs. For the pneumonia classification task, we report four binary classification metrics: sensitivity (SEN), specificity (SPE), area under the ROC curve (AUC), and classification accuracy (ACC). Additionally, for the NIH dataset, we calculate classification accuracy for each disease category and compute an overall weighted accuracy to reflect performance across all categories.

Table 3.1 compares the bone suppression performance of our proposed method with a range of competitive approaches, including both general-purpose image-to-image translation models and domain-specific bone suppression methods. The compared models encompass Stable Diffusion [76], StarGAN2 [16], CycleGAN [125], Pix2pix [38], RQ-VAE [52], and VQGAN [25] (used as a baseline in Table 3.2), as well as dedicated bone suppression techniques such as Auto-Encoder [29], IEDSR [58], and Dilated cGAN [124].

To ensure a fair comparison, we fine-tuned or re-trained all models on the publicly available Bone Shadow Suppression dataset [83]. Specifically, for general-purpose models, we adopted their officially released pre-trained weights: Stable Diffusion (trained on LAION-5B), StarGAN2

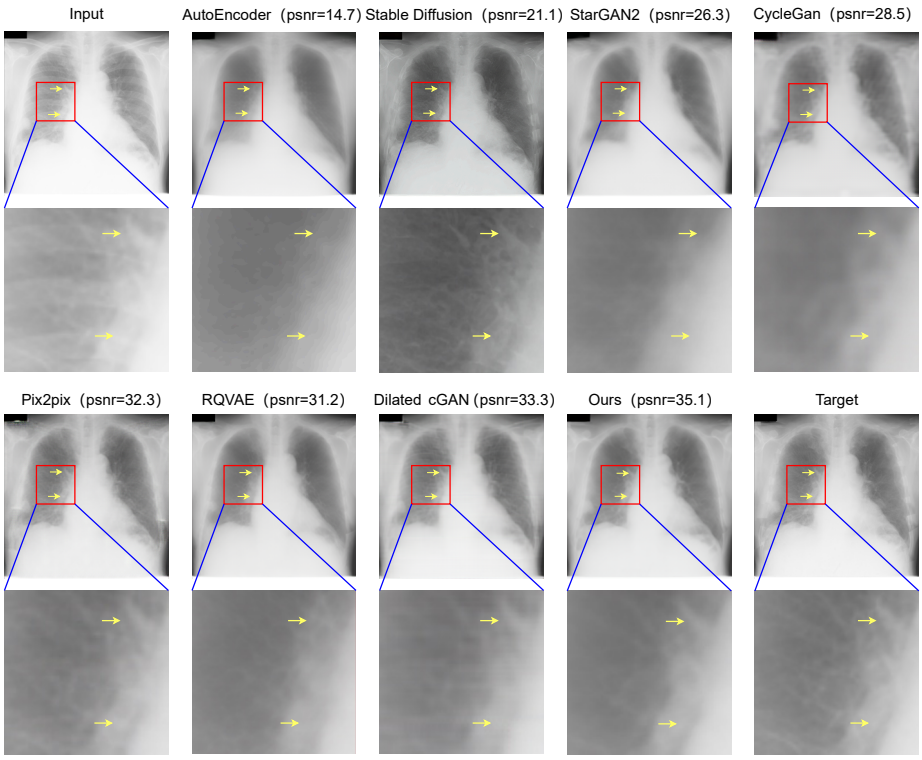



**Figure 3.4:** Visualization of bone suppression using different methods. The first image is the input bone-shadowed CXR. Then, we respectively show the boneless CXRs generated by using AutoEncoder [29], Cycle GAN [125], Pix2pix [38], RQ VAE [52], Dilated cGAN [124] and our method. The last image is the ground-truth boneless CXR.

(CelebA-HQ), CycleGAN and Pix2Pix (Facade dataset), and fine-tuned them on the bone suppression dataset. In contrast, the specialized methods Auto-Encoder, IEDSR, and Dilated cGAN were trained from scratch on the same dataset.

As shown in Table 3.1, our approach achieves the best performance across all quantitative metrics (PSNR, SSIM, and MSE), clearly outperforming both traditional and state-of-the-art methods. While recent techniques such as RQ-VAE and Dilated cGAN demonstrate improved results compared to older baselines, they still fall short of our method. To further validate the robustness of these performance gains, we conducted paired statistical significance tests across PSNR, SSIM, and MSE.

**Table 3.2:** Ablation study of bone suppression performance on the X-ray bone suppression dataset [83] and downstream classification on the X-ray pneumonia dataset [45].

| Component | | | | Performance Metrics | | | | |
|---|---|---|---|---|---|---|---|---|
| ViT | Multi-head | Soft codebook | CAB | PSNR ↑ | SSIM ↑ | MSE ↓ | #Params ↓ | Classification Accuracy on dataset [45]↑ |
| | | | | 30.6 | 0.949 | $1.1 \times 10^{-3}$ | 89.2M | 91.5% |
| ✓ | | | | 30.9 | 0.951 | $1.0 \times 10^{-3}$ | 189.5M | 91.9% |
| | ✓ | | | 32.0 | 0.957 | $7.5 \times 10^{-4}$ | 89.2M | 92.5% |
| | | ✓ | | 32.1 | 0.958 | $7.5 \times 10^{-4}$ | 89.2M | 92.5% |
| | | | ✓ | 30.8 | 0.950 | $1.0 \times 10^{-3}$ | 115.3M | 91.8% |
| ✓ | ✓ | | | 32.4 | 0.960 | $7.3 \times 10^{-4}$ | 189.5M | 93.0% |
| ✓ | | ✓ | | 32.6 | 0.961 | $7.2 \times 10^{-4}$ | 189.5M | 93.1% |
| ✓ | | | ✓ | 31.5 | 0.954 | $8.8 \times 10^{-4}$ | 215.6M | 92.2% |
| | ✓ | ✓ | | 33.6 | 0.965 | $6.0 \times 10^{-4}$ | 89.2M | 93.5% |
| | ✓ | | ✓ | 32.8 | 0.960 | $7.1 \times 10^{-4}$ | 115.3M | 93.2% |
| | | ✓ | ✓ | 33.0 | 0.961 | $6.6 \times 10^{-4}$ | 115.3M | 93.3% |
| ✓ | ✓ | ✓ | | 34.0 | 0.968 | $5.8 \times 10^{-4}$ | 189.5M | 93.6% |
| ✓ | ✓ | | ✓ | 33.0 | 0.962 | $6.6 \times 10^{-4}$ | 215.6M | 93.3% |
| ✓ | | ✓ | ✓ | 33.2 | 0.964 | $6.3 \times 10^{-4}$ | 215.6M | 93.4% |
| | ✓ | ✓ | ✓ | 34.1 | 0.970 | $5.7 \times 10^{-4}$ | 115.3M | 93.6% |
| ✓ | ✓ | ✓ | ✓ | **35.0** | **0.975** | $\mathbf{4.6 \times 10^{-4}}$ | 215.6M | **94.1%** |

The results show that the improvements achieved by our method are statistically significant—PSNR ($p = 0.024$), SSIM ($p = 0.018$), and MSE ($p = 0.032$)—confirming that the observed advantages are not due to random variation but represent consistent and meaningful enhancements. Visual comparisons in Figure 3.4 further confirm these findings: our method yields sharper structural details in the lung region, with greater fidelity to the ground truth bone-suppressed images.

Moreover, ablation studies summarized in Table 3.2 validate the effectiveness of our proposed multi-head codebook attention (MCA). In particular, replacing VQGAN's vanilla hard codebook (Row 3) with our single-head soft codebook (Row 6), and ultimately with the full MCA design (Row 18), significantly improves performance. Unlike conventional hard quantization that discretizes feature representations—potentially losing fine-grained details—our soft codebook leverages attention- weighted combinations of multiple codewords, allowing for flexible and expressive encoding. This strategy not only preserves important anatomical structures but also improves pathological region detection through adaptive feature representation.

In addition, our method reduces visual artifacts, enhances image consistency before and after suppression, and promotes more stable and

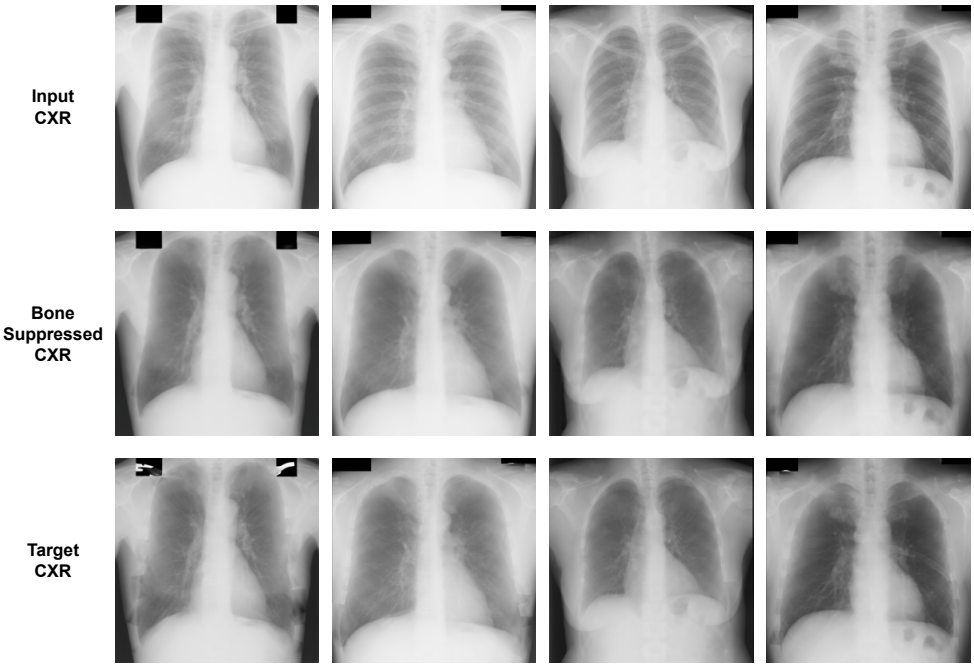



**Figure 3.5:** Visualization of the bone suppression results. The first row shows the input bone CXRs, the second row shows generated boneless CXRs using our method, and the last row shows the target boneless CXRs (*i.e.*, ground-truth) obtained by using the DES chest radiography.

efficient training. Even under limited data conditions, the training converges faster and yields superior results. Figure 3.5 presents example outputs, including the input CXR, the bone-suppressed result from our model, and the corresponding ground truth image generated by DES radiography, further demonstrating the visual quality and clinical relevance of our framework.

### 3.4.2 Disease Classification

To further demonstrate the practical utility of our bone suppression framework, we conducted downstream evaluations on two classification tasks: pneumonia binary classification and NIH 14-category classification. In particular, we applied our pre-trained bone suppression model—trained on the Bone Shadow Suppression X-ray dataset [83]—to generate boneless chest X-rays (CXRs) from the pneumonia dataset [45] and the NIH

**Table 3.3:** Classification performance comparison on the X-ray pneumonia dataset [45]. The bone-suppression model is trained on the X-ray bone suppression dataset. We use a ResNet-50 pretrained on ImageNet and then fine-tuned on the pneumonia dataset for classification.

| Method | SEN / SPE / AUC / ACC (%) |
|---|---|
| No Suppression | 91.4 / 91.8 / 89.5 / 91.7 |
| Stable Diffusion | 90.6 / 91.3 / 88.6 / 91.0 |
| Star GAN2 | 90.2 / 92.1 / 89.1 / 91.4 |
| Cycle GAN | 91.0 / 92.1 / 89.5 / 91.7 |
| Pix2pix | 91.9 / 91.8 / 89.7 / 91.8 |
| RQ VAE | 93.2 / 91.5 / 90.2 / 92.1 |
| Auto-Encoder | 88.0 / 89.8 / 86.4 / 89.1 |
| IEDSR | 90.9 / 91.6 / 89.2 / 91.5 |
| Dilated cGAN | 93.2 / 92.6 / 91.5 / 92.8 |
| **Ours** | **93.6 / 94.4 / 93.3 / 94.1** |

dataset [104].

We then fine-tuned a ResNet-50 classifier [33], initialized with ImageNet pre-trained weights, using either the original CXRs or the corresponding bone-suppressed versions produced by different methods. The goal was to assess how bone suppression impacts diagnostic performance.

Table 3.3 reports the sensitivity, specificity, AUC, and overall classification accuracy for the pneumonia dataset. Results show that our bone suppression method significantly boosts classification performance. Specifically, the overall accuracy increased from 91.7% (original CXRs) to 94.1% with our method. In contrast, only a few methods (Pix2Pix, RQ-VAE, and Dilated cGAN) yielded marginal improvements, while others failed to enhance—and in some cases even degraded—the classifier's performance. Notably, all evaluation metrics (SEN, SPE, AUC) improved with our method, underscoring its effectiveness in enhancing the diagnostic signal of pneumonia.

Interestingly, although Stable Diffusion underperformed in traditional image fidelity metrics such as PSNR and SSIM (see Table 3.1), it achieved competitive results in the clinical classification task. This

**Table 3.4:** disease classification performance comparison on the nih chest x-ray dataset [104]. The bone-suppression model is trained based on x-ray bone suppression dataset. We use a resnet-50 pretrained and subsequently fine-tuned based on imagenet and nih chest x-ray dataset as our classification network (Classification Accuracy [%])

| Disease / Method | No Suppression | Stable Diffusion | Star GAN2 | Cycle GAN | Pix2pix | RQ VAE | Auto-Encoder | IEDSR | Dilated cGAN | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| Atelectasis | **70.9** | 64.5 | 67.3 | 68.1 | 69.1 | 69.3 | 58.8 | 68.0 | 68.9 | 70.0 |
| Cardiomegaly | **61.9** | 58.0 | 57.4 | 56.9 | 57.9 | 57.7 | 57.4 | 56.6 | 59.9 | 58.2 |
| Consolidation | 69.8 | 69.2 | 70.4 | 72.6 | 71.3 | 72.6 | 61.5 | 72.4 | 70.3 | **74.4** |
| Edema | 80.8 | 80.6 | 81.2 | 82.4 | 81.5 | 83.1 | 73.2 | 81.9 | 82.1 | **85.6** |
| Effusion | 78.0 | 75.4 | 79.3 | 78.5 | 79.9 | **81.2** | 64.7 | 78.2 | 77.8 | 80.3 |
| Emphysema | 69.7 | 66.8 | 67.1 | 68.3 | 69.1 | 68.3 | 64.1 | 68.5 | 69.4 | **70.8** |
| Fibrosis | 69.6 | 68.2 | 67.9 | 68.2 | 68.4 | 70.1 | 64.5 | 68.1 | 70.5 | **71.7** |
| Hernia | 81.8 | 77.6 | 80.1 | 81.3 | 80.7 | 75.3 | 74.3 | 80.7 | **83.2** | 81.8 |
| Infiltration | 63.0 | 61.2 | 62.3 | 61.8 | 62.7 | 63.7 | 61.0 | 61.2 | 64.2 | **64.9** |
| Mass | 64.9 | 67.3 | 65.2 | 65.7 | 66.6 | 66.6 | 61.2 | 65.2 | 66.4 | **67.9** |
| Nodule | 79.8 | 77.5 | 78.2 | 77.8 | 76.9 | 80.5 | 67.3 | 77.5 | 81.4 | **83.5** |
| Pleural thickening | 64.9 | 58.0 | 63.2 | 64.4 | 65.6 | 67.4 | 62.9 | 64.0 | 66.1 | **68.9** |
| Pneumonia | 57.8 | 61.2 | 62.7 | 68.1 | 67.3 | 72.4 | 56.1 | 67.7 | 64.2 | **73.4** |
| Pneumothorax | 73.0 | 66.7 | 69.3 | 71.2 | 70.9 | 73.3 | 64.5 | 70.9 | 72.5 | **74.0** |
| Weighted Avg | 68.4 | 65.8 | 67.1 | 67.3 | 67.9 | 68.9 | 61.8 | 67.2 | 68.7 | **71.3** |

indicates that pixel-wise similarity metrics may not always reflect improvements in disease-relevant features—especially in regions occluded by ribs. Our method, by contrast, achieves superior scores in both image quality and clinical relevance, suggesting a better balance.

For the NIH dataset, Table 3.4 presents classification accuracy across 14 disease categories. Our approach improved the weighted classification accuracy from 68.4% to 71.3%. In contrast, most other methods led to a drop in performance, with only RQ-VAE and Dilated cGAN yielding modest gains. While some diseases (e.g., Atelectasis and Cardiomegaly) did not benefit from bone suppression—likely due to their reliance on coarse anatomical structures—most categories showed improved classification with our boneless images. Notable gains were observed in diseases affecting finer lung structures, such as Consolidation, Edema, Nodule, Pleural Thickening, and Pneumonia, supporting the idea that rib shadows can interfere with disease localization in subtle cases.

To further validate these observations, we visualized Grad-CAM attention maps [80] in Figure 3.6 for both normal and pneumonia cases. A radiologist annotated the lesion regions in pneumonia samples (highlighted with red contours in Figure 3.6(b)). As shown, classifiers trained on original CXRs tend to focus on irrelevant regions (e.g., lung centers in normal cases or peripheral areas in pneumonia). In contrast, attention

**Table 3.5:** Tuberculosis segmentation performance comparison on Chest X-ray Dataset for Tuberculosis Segmentation [41] (%).

| Method | Accuracy ↑ | Dice Coefficient ↑ | Jaccard Index ↑ |
|---|---|---|---|
| No Suppression | 96.2 | 92.8 | 87.4 |
| Stable Diffusion | 94.8 | 91.6 | 86.2 |
| Star GAN2 | 95.5 | 92.2 | 86.7 |
| Cycle GAN | 95.7 | 92.4 | 86.9 |
| Pix2pix | 96.5 | 93.0 | 87.5 |
| RQ VAE | 96.6 | 93.2 | 87.7 |
| Auto-Encoder | 91.0 | 88.5 | 83.6 |
| IEDSR | 95.6 | 92.3 | 86.7 |
| Dilated cGAN | 96.8 | 93.3 | 87.9 |
| **Ours** | **97.6** | **94.0** | **88.7** |

maps from our bone-suppressed inputs better align with radiologist-identified lesions, suggesting that bone suppression helps models focus on clinically meaningful features and improves interpretability.

### 3.4.3 Tuberculosis Segmentation

We further validate the effectiveness and generalizability of our bone suppression model on the Chest X-ray Dataset for Tuberculosis Segmentation [41]. Specifically, we train a U-Net segmentation network using paired lung masks with both original CXRs and bone-suppressed images generated by different methods. As summarized in Table 3.5, our approach leads to consistent performance improvements across all metrics: segmentation accuracy increases from 96.2% to 97.6%, Dice coefficient from 92.8% to 94.0%, and Jaccard index from 87.4% to 88.7%. Among competing methods, only Pix2Pix, RQ-VAE, and Dilated cGAN produce bone-suppressed images that outperform the baseline without suppression, underscoring the superior segmentation benefits of our model.

### 3.4.4 Performance on Higher-resolution CXRs

We further performed ablation studies using higher-resolution images, which can potentially improve model performance but require significantly more GPU memory. Given the memory constraints of the A5000

**Table 3.6:** Bone-suppression performance comparison at 512×512 resolution on the X-ray bone suppression dataset [83].

| Method | PSNR↑ | SSIM↑ | MSE↓ |
|---|---|---|---|
| Stable Diffusion | 24.9 | 0.854 | $4.5 \times 10^{-3}$ |
| StarGAN2 | 28.3 | 0.876 | $2.1 \times 10^{-3}$ |
| CycleGAN | 28.7 | 0.941 | $1.8 \times 10^{-3}$ |
| Pix2pix | 36.1 | 0.956 | $1.1 \times 10^{-4}$ |
| RQ-VAE | 36.3 | 0.947 | $1.0 \times 10^{-4}$ |
| Auto-Encoder | 17.3 | 0.831 | $1.7 \times 10^{-2}$ |
| IEDSR | 27.5 | 0.965 | $2.0 \times 10^{-3}$ |
| Dilated cGAN | 37.1 | 0.946 | $9.5 \times 10^{-5}$ |
| **Ours** | **40.6** | **0.981** | $\mathbf{8.8 \times 10^{-5}}$ |

GPU, we adopted a patch-based training strategy with 512 x 512 resolution chest X-rays. Each 512 × 512 image was divided into four non-overlapping patches of 256 x 256 pixels for both training and inference. During testing, the patches were individually processed and then reassembled to reconstruct the full-size output.

As shown in Table 3.6, our model achieves improved performance at the higher resolution of 512 x 512. The comparison with other methods demonstrates the robustness and adaptability of our learning strategy to high-resolution inputs. Furthermore, as presented in Table 3.7, the bone-suppressed images generated by our approach yield superior disease classification accuracy compared to all other baselines under the same resolution setting.

### 3.4.5 Ablation Study

Table 3.2 presents the results of our ablation study, highlighting the individual contributions of key components in our model toward enhancing generation quality. The symbol "+" denotes the inclusion of a specific component. As shown, incorporating the ViT-based encoder-decoder, multi-head codebook attention (MCA), and context-aware blocks (CAB) into the VQGAN [25] baseline leads to consistent improvements in both

**Table 3.7:** Classification performance (Accuracy [%]) comparison using 512×512 images from the X-ray pneumonia dataset [45].

| **Method** | **Accuracy↑** |
|---|---|
| No Suppression | 92.3 |
| Stable Diffusion | 91.6 |
| StarGAN2 | 92.1 |
| CycleGAN | 92.6 |
| Pix2pix | 93.0 |
| RQ VAE | 93.3 |
| Auto-Encoder | 90.1 |
| IEDSR | 92.1 |
| Dilated cGAN | 94.0 |
| **Ours** | **95.2** |

image quality and downstream classification performance on the X-ray bone suppression dataset [83].

Among these, our proposed MCA stands out as the core innovation. To validate its effectiveness, we conducted a comparative visualization study of bone suppression results using three variants: the standard codebook, the multi-head codebook, and our MCA. This analysis clearly demonstrates the superiority of MCA in generating sharper, more anatomically consistent outputs.

**Ablation on Codebook Heads.** To assess the impact of the number of codebook heads on model performance, we conducted experiments using 4, 8, and 16 heads. As reported in Table 3.8, the model exhibits stable performance across these configurations. Notably, the use of 8 heads achieves the best balance between performance and computational cost, indicating it is a favorable setting for our architecture.

**Ablation on Sparse Attention.** We further explored the use of sparse attention mechanisms to promote better semantic disentanglement. Specifically, we replaced the standard softmax function with sparsemax, which enforces sparsity in the attention weights. While this theoretically encourages a more selective and interpretable code combination, we observed a slight performance decline—from a PSNR/SSIM of 35.0/0.975

**Table 3.8:** Ablation study on codebook heads.

| Metric | 4 | 16 | 8 |
|---|---|---|---|
| PSNR ↑ | 34.5 | 34.8 | **35.0** |
| SSIM ↑ | 0.972 | 0.973 | **0.975** |
| MSE ↓ | $5.1 \times 10^{-4}$ | $4.9 \times 10^{-4}$ | $\mathbf{4.6 \times 10^{-4}}$ |
| Params ↓ | 215.6M | 215.6M | 215.6M |
| Accuracy ↑ | 93.8% | 93.9% | **94.1%** |

to 34.9/0.973. This suggests that although sparsity may enhance disentanglement, it can also introduce trade-offs that do not necessarily improve image reconstruction quality. Exploring alternative sparsity-inducing strategies remains an avenue for future research.

### 3.4.6 Limitation

Although our method demonstrates strong performance, it faces certain limitations—primarily the high computational cost associated with the ViT and CAB modules. To address this, future work will explore more efficient and lightweight backbone designs, such as integrating sparse attention mechanisms within the ViT architecture or applying model pruning techniques to reduce complexity. Furthermore, the paired bone-suppressed CXR dataset employed in this study is relatively limited in scale and diversity. Expanding the dataset with more varied and representative samples could further boost the robustness and generalizability of our framework.

## 3.5 Summary

In this work, we present a two-stage learning-based framework for CXR bone suppression. Our method first leverages a MCA module to capture domain-level contextual information from the target domain. Subsequently, a ViT encoder-decoder architecture, enhanced with CABs, is employed to extract fine-grained instance-level features. By effectively integrating both global and local representations, our approach surpasses existing state-of-the-art generative methods in bone suppression quality. Moreover, the boneless CXRs produced by our model significantly

improve performance in downstream classification and segmentation tasks.

The contributions of this work are summarized as follows:

- We introduce a MCA mechanism to capture domain-level representations from the target boneless CXRs. By leveraging attention and a multi-head structure, MCA effectively encodes domain priors, enhancing the model's generalization capability when processing original CXRs.
- To obtain more informative instance-level features, we employ a hybrid approach combining vanilla and cross-covariance Transformer mechanisms. A ViT-based encoder-decoder framework is used to encode input features and reconstruct boneless CXRs, while a CAB network highlights clinically relevant regions to guide feature extraction.
- Extensive experiments validate that our bone-suppressed CXR images not only yield superior reconstruction quality but also significantly enhance performance in downstream tasks, such as pneumonia classification.

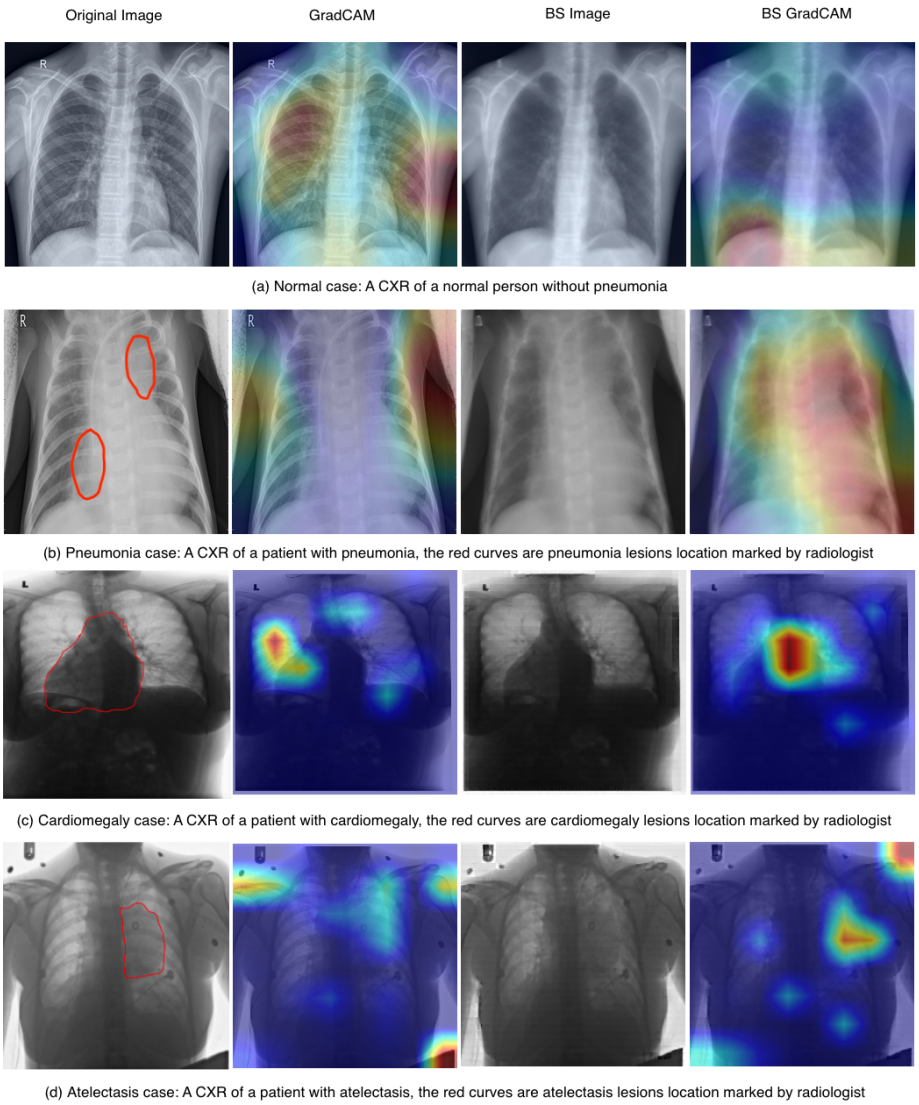



**Figure 3.6:** Two examples of Grad-CAM generated attention maps for pneumonia classification using original and bone-suppressed (BS) CXRs. Figure a is a normal chest X-ray. The Gradcam classification method has false positives in the screening of lesions. After using the new algorithm to remove the rib shadow, the false positive rate of Gradcam is significantly reduced. Figures b–d present representative abnormal cases, including pneumonia, cardiomegaly, and atelectasis. In all cases, radiologist-annotated pathological regions are indicated by red curves. When Grad-CAM is applied to the original CXRs, the attention maps are often distracted by rib structures, resulting in false negatives, diffuse activations, or suboptimal localization of clinically relevant regions. After applying the proposed rib-removal (bone suppression) algorithm, Grad-CAM activations become more concentrated on the true pathological areas—such as pneumonic infiltrates, the enlarged cardiac silhouette, and atelectatic lung regions—thereby improving disease localization, interpretability, and classification reliability.

# Chapter 4

# MedXChat: A Unified Multimodal Large Language Model Framework towards CXRs Understanding and Generation

In this work, we address the challenge of unified modeling for multimodal medical image tasks, which remains largely underexplored in the context of CXRs. Existing approaches typically focus on isolated tasks such as report generation or VQA, lacking a cohesive framework capable of handling multiple vision-language modalities simultaneously. To bridge this gap, we propose MedXChat, a unified multimodal framework built upon a large vision-language model that supports various tasks, including image synthesis, VQA, and report generation for CXR images. Our framework integrates an instruction data construction, an instruction-following Stable Diffusion module for image generation, and efficient task-specific adaptation via LoRA-based fine-tuning. Extensive experiments demonstrate that MedXChat not only produces high-quality and anatomically realistic CXR images but also achieves strong performance across diverse downstream tasks. Notably, the generated images preserve clinically significant attributes such as disease location, severity, and anatomical perspective, contributing to improved accuracy in diagnostic tasks. Our results highlight the potential of unified LLMs in advancing intelligent medical imaging and facilitating more effective clinical decision support.

## 4.1 Introduction

Multi-modal tasks that integrate image and text modalities play a pivotal role in advancing the field of medical image analysis by enabling machines to both understand and reason about complex clinical information. These tasks have seen growing importance in recent years, particularly in areas such as medical report generation [53, 91, 54, 97, 17, 105], VQA [114, 91, 54], and visual grounding[54, 53], which together form the foundation for intelligent and interpretable clinical decision support systems. These multi-modal tasks enable AI models to go beyond basic classification, facilitating more nuanced capabilities such as generating descriptive diagnostic reports, interacting with users through natural language, and identifying specific pathological regions on radiographs. LLMs, particularly those based on transformer architectures [97], have shown remarkable promise in supporting such multi-modal tasks. Their ability to process and synthesize large-scale text data allows them to capture subtle semantic nuances, making them well-suited for understanding complex medical language. When extended to vision-language scenarios, these models can bridge the semantic gap between visual content and textual descriptions, thereby enhancing the interpretability and utility of medical imaging outputs. For instance, LLMs can be prompted with natural language questions regarding a patient's chest X-ray, and return answers based on contextual visual cues—a functionality that is increasingly being explored in medical VQA models [114, 91, 54]. The development of high-capacity models such as LLaMA2 [93] and ChatGPT-4 [66] has significantly accelerated progress in vision-language research, with many studies demonstrating their capabilities in general-purpose multi-modal tasks [49, 27, 111]. These models excel at instruction- following, reasoning, and image-grounded dialogue generation, making them attractive candidates for adaptation to medical domains. As a result, researchers have begun to incorporate these LLMs into various clinical tasks, including automated report generation from radiographs, retrieval of relevant case examples, and interpretation of patient-specific queries [91].

Despite these advances, significant gaps remain in the current research landscape. Most existing works treat multi-modal medical tasks in isolation, developing separate models for report generation, VQA, and visual grounding. This leads to fragmented workflows, redundant computations, and increased development overhead. There is a clear need for unified frameworks that can seamlessly handle multiple vision-language tasks within a single model, allowing shared representations and joint optimization across tasks. Such unified solutions would not only streamline deployment in clinical environments but also lead to performance gains by leveraging multi-task learning and shared knowledge. While text-to-image generation has gained traction in the general domain, its application in the medical field—especially for synthesizing clinically realistic images from text—is still in its infancy. Compared to tasks like report generation and VQA, text-to-medical-image synthesis presents unique challenges, including the need for anatomical precision, pathophysiological accuracy, and high fidelity to disease-specific manifestations. The limited availability of paired textual and imaging data further complicates training in this setting. Consequently, few studies have tackled this task, and even fewer have attempted to integrate it into a broader multi-modal system. Furthermore, most current approaches rely heavily on visual tokenization or vector quantization to bridge image and text domains [54, 53], which introduces challenges such as quantization error and fixed vocabulary bottlenecks. These limitations hinder scalability and limit the model's ability to generalize to new tasks, unseen pathologies, or alternative imaging modalities. While transformer-based LLMs and recent advancements in vision-language modeling have substantially improved the capabilities of medical AI, there remains a pressing need for next-generation frameworks that are unified, generalizable, and capable of high-fidelity generation and reasoning across all core multi-modal tasks. Addressing these gaps will be essential for building intelligent, interactive, and clinically reliable assistants that can function in real-world healthcare settings.

While a unified framework for large-scale multimodal medical data processing holds significant promise, such solutions are still notably lacking, particularly in conventional medical imaging domains such as CXRs.

Despite the increasing integration of vision-language models into healthcare, only two notable frameworks currently exist that support both image interpretation and image generation within the same system: UniXGen [53], a non-LLM-based method, and LLM-CXR [54], a more recent approach that incorporates LLMs. UniXGen employs a VQ strategy [25] to unify visual and textual representations, allowing it to tackle report generation and visual understanding by mapping CXR images into discrete tokens. This mapping facilitates a common representation space between image and text modalities. However, the use of VQ introduces quantization errors that can degrade semantic fidelity—an issue particularly problematic in medical imaging, where fine-grained anatomical detail is essential. Moreover, because the token vocabulary is fixed, any expansion to new datasets, imaging modalities, or medical terminologies often requires re-training the entire model, which adds substantial computational burden and limits its practical scalability. LLM-CXR [54] builds on this VQ-based foundation but incorporates instruction-tuned LLMs to enhance reasoning and cross-modal alignment. By aligning discrete image tokens with language prompts via instruction tuning, it enables interactive tasks like medical report generation and VQA. Nevertheless, it inherits the core limitations of VQ: a rigid codebook design, loss of fine semantic detail, and limited adaptability across diverse datasets or unseen token combinations. These issues are further exacerbated when handling rare pathologies or subtle radiographic patterns, where generalization becomes critical. In practice, these VQ-based frameworks fall short in handling the real-world diversity found in clinical environments. Medical imaging data often vary widely across institutions due to differences in imaging protocols, devices, demographics, and disease prevalence. The reliance on static codebooks and fixed token spaces limits these frameworks' ability to generalize across such heterogeneous domains. This also restricts the inclusion of more expressive or novel medical concepts that lie outside the pre-defined token vocabulary. For instance, when facing rare diseases or emerging diagnostic categories, the models lack flexibility unless retrained extensively—a major barrier for real-time clinical adaptation. Furthermore, VQ-based frameworks tend to suffer from high memory and compute overhead,

especially during the encoding-decoding stages, which becomes a bottleneck for deployment in time-sensitive or resource-limited settings. The lack of modularity also means updates to the image processing pipeline require simultaneous adjustment of the entire system, further reducing efficiency and maintainability. Altogether, these limitations underscore the urgent need for a more efficient and generalizable multimodal framework that overcomes the bottlenecks of VQ-based methods. A promising direction is to adopt continuous visual representations (e.g., CLIP-style features) instead of discrete token spaces, enabling richer semantic alignment while preserving computational efficiency. Such frameworks should also support modular training, allowing different components (e.g., language model, vision encoder, image generator) to be updated independently. Moreover, unified systems must be able to scale across tasks—report generation, VQA, visual grounding, and text-to-image synthesis—without compromising accuracy, interpretability, or clinical realism. Ultimately, the goal is to move toward intelligent medical assistants that not only understand but also interact meaningfully with diverse forms of medical data across real-world clinical workflows.

To overcome these issues and advance multimodal medical data processing, we propose MedXChat—a unified LLM framework designed to function as an intelligent medical assistant. MedXChat supports a broad range of clinical tasks, including image-to-report generation, visual question answering, and report-to-image synthesis. It addresses the image-text gap through three core strategies: First, instead of relying on discrete representations produced by VQ-based encoding [54, 53], MedXChat adopts visual features extracted using a pre-trained CLIP encoder [69]. This encoder leverages contrastive learning to explicitly align image and text features, enabling a tighter semantic mapping that is especially effective for fine-grained medical image data. Second, we utilize the closed-source ChatGPT-4 API [66] as an instructor to generate high-quality instruction data. These instructions are created in the form of dialogues that incorporate both medical reports and visual cues, which are then used to fine-tune our open-source LLM, improving its capability to handle multimodal clinical queries. Third, for text-to-image synthesis,

MedXChat bypasses the traditional intermediate token conversion process and instead directly generates CXR images from the textual prompts produced by the fine-tuned LLM. This is achieved using the instruction-following capabilities of the Stable Diffusion model [76]. By decoupling the image generation module from the language model, we can fine-tune only the diffusion component without retraining the entire framework, thus improving efficiency and modularity.

We validate the effectiveness of MedXChat through extensive experiments on three benchmark tasks: CXR-to-report generation, CXR-VQA, and text-to-CXR synthesis. Our model consistently outperforms both LLM-based methods [54, 53, 91, 110, 55] and non-LLM-based approaches [13, 14, 17]. To further assess the clinical utility of our system, we engaged a board-certified radiologist to evaluate 20 generated reports and 20 CXRs. Comparisons against UniXGen [53] and LLM-CXR [54] show that MedXChat yields superior results, reinforcing its practical potential in real-world medical applications.

## 4.2 Method

We introduce MedXChat, a unified LLM framework designed to process multimodal medical inputs—such as images and text—and generate task-specific outputs accordingly. An overview of the proposed framework is illustrated in the bottom right of Fig. 6.1. To build our system, we adopt the off-the-shelf LLM architecture mPLUG-Owl [118], chosen for its demonstrated success in open-source instruction-tuned language models [90, 15, 68]. During training, the majority of the model parameters—including the backbone LLM and the SD component—are frozen to maintain stability and reduce computational costs. For adapting the LLM, we employ a delta tuning strategy via Low-Rank Adaptation (LoRA) [37]. The input to MedXChat can include both a CXR image and a user-provided text instruction. The image is first processed by a pre-trained CLIP visual encoder (ViT-L/14) [69] to extract fine-grained, text-aligned regional visual features—crucial for capturing detailed semantics in medical images. These features are then combined with the

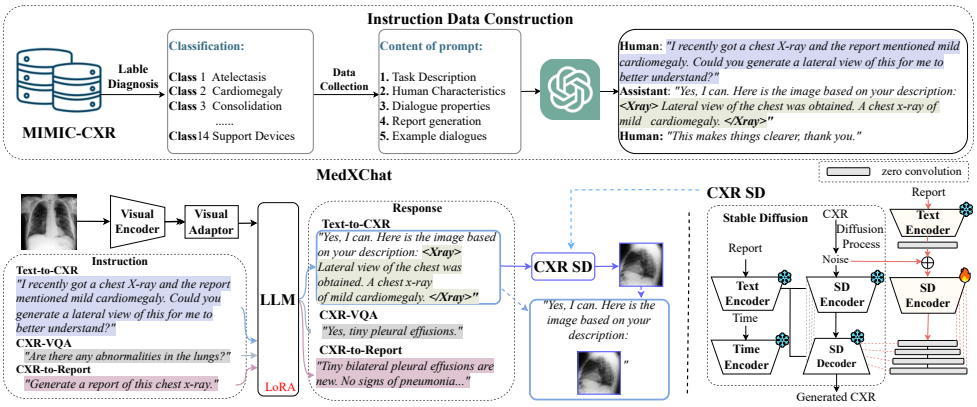



**Figure 4.1: Overview of our MedXChat framework**, including a preparation stage (dashed boxes) for constructing instruction data (top row) and fine-tuning the Stable Diffusion model using CXR images (referred to as CXR-SD), followed by an instruction tuning stage (solid box) where our multimodal MedXChat is actively trained.

user instruction to form a unified input sequence, which is passed to the LLM. Based on the input, the model generates an appropriate response. Notably, when a CXR image generation is requested, the LLM outputs a text prompt that serves as input to the CXR-SD module, a fine-tuned version of the original Stable Diffusion model tailored for CXR synthesis. The training process is divided into two stages: (1) a preparation stage, where we construct instruction-following datasets (Sec. 4.2.1) and fine-tune the SD component on CXR data (Sec. 4.2.2), and (2) the instruction tuning stage, where the full multimodal framework is optimized using task-specific objectives (Sec. 4.2.3 & 4.2.4).

### 4.2.1 Instruction Data Construction

Most current open-source LLMs [91, 55] lack the capability to handle visual inputs, limiting their applicability in tasks such as medical image generation. In contrast, our unified MedXChat framework is specifically designed to support multimodal processing, including image generation. To enable our end-to-end LLM to engage in text-to-CXR interactions in a clinician-like conversational style, we require high-quality instruction data to guide the fine-tuning process.

To construct these instruction datasets, we leverage the ChatGPT-4

API [66] alongside prompt engineering techniques. Our prompts are designed to produce detailed textual passages containing five key components, as shown in the top panel of Fig. 6.1: **1. Text Description** provides a general guideline for the response, *e.g.*, *'Please help construct three dialogues between a person and a helpful assistant. Each chest X-ray image is represented by* $< Xray >$ *DESCRIPTION* $< /Xray >$*, where DESCRIPTION is a textual description or report of the X-ray image"*. **2. Human Characteristics** allows users to instruct the assistant on their individual need in generating chest X-ray images, *e.g.*, whether the posteroanterior (PA) view or the lateral view is preferred. **3. Dialogue Properties** defines the properties of a good dialogue to ensure logical flow and restrict the content to only the pertinent parts of the report. **4. Report Generation** involves selecting a medical report from the MIMIC-CXR dataset and completing it with the view information. **5. Example Dialogues** provides two typical conversation examples within our framework.

To support structured text-to-CXR generation, we introduce special tokens $< Xray >$ and $< /Xray >$ to demarcate text segments that correspond to visual data. For instance: "The constructed dialogue must contain questions according to the following input chest X-ray image, which contains only the relevant part of the report between $< Xray >$ and $< /Xray >$: $< Xray >$ content $< /Xray >$", where the placeholder content is filled with the view-augmented report of a given CXR.

To enrich the reports with view-specific metadata, we classify the view type (PA vs. lateral) using a transformer-based encoder trained on all CXRs from the MIMIC-CXR training set. The identified view (e.g., "PA view of the chest was obtained") is prepended to each report. These augmented reports are then used for fine-tuning both the LLM and CXR-SD, replacing the original raw reports to provide enhanced context.

From the MIMIC-CXR dataset—which contains paired CXRs and reports for 14 diagnostic categories—we randomly select 200 reports per category, resulting in 2800 reports. Each report is passed to ChatGPT-4 to generate three distinct dialogue samples, yielding a total of 8400 dialogue segments. These are then used to fine-tune our LLM using LoRA-based delta tuning, enabling it to perform medical dialogue analysis and

generation.

Once trained, the LLM can identify the special tokens $< Xray >$ and $< /Xray >$ during conversation. When a report segment is enclosed by these markers, the model recognizes it as a prompt for image synthesis. The enclosed text is then forwarded to our fine-tuned CXR-SD model, which generates a corresponding chest X-ray image. To further support user-specified views, we fine-tune the SD model using the view-augmented report-image pairs, ensuring accurate synthesis of CXRs in the requested orientations.

### 4.2.2 CXR Stable Diffusion(SD)

To generate CXR images from clinical text, we leverage a pre-trained SD model [76]. Following recent fine-tuning strategies [122], we adapt SD using the MIMIC-CXR dataset, integrating a lightweight zero-convolution module for efficient training, as illustrated in Fig. 6.1 (bottom left).

Stable Diffusion models the data distribution $p(z)$ by progressively denoising a normally distributed variable through an encoder-decoder architecture. This process approximates the reverse of a predefined Markov chain over $T$ time steps. At each time step $T$, text features are first encoded using a time encoder $E_t$ via positional encoding techniques.

Beginning with a sample CXR $z_0$, the model applies a forward diffusion process $q$, incrementally adding noise according to:

$$q\left(z_t \mid z_0\right) = \mathcal{N}\left(z_t; \alpha_t z_0, \sigma_t^2 \mathbb{I}\right). \tag{4.1}$$

After $T$ steps, the clean image is fully transformed into noise $z_t$, which is then input into a U-Net-based encoder-decoder network [77]. This architecture comprises four downsampling and four upsampling blocks, each connected via skip connections. The core blocks feature a combination of ResNet layers and ViTs, each equipped with self-attention and cross-attention mechanisms [97] to enhance spatial and semantic alignment.

For conditional generation, the model incorporates text prompts encoded by the CLIP text encoder $E_{clip}$ [69]. These embeddings are fused with the noisy input to drive the reverse diffusion process:

$$\begin{aligned} p_\theta\left(z_{t-1} \mid z_t\right) &:= q\left(z_{t-1} \mid z_t, \epsilon_\theta\left(z_t, t\right)\right) \\ &= \mathcal{N}\left(z_{t-1}; \mu_\theta\left(z_t, t\right), \sigma_{t|t-1}^2 \frac{\sigma_{t-1}^2}{\sigma_t^2} \mathbb{I}\right), \end{aligned} \tag{4.2}$$

where $\epsilon_\theta$ is the noise prediction function implemented via U-Net. Through iterative denoising, a high-quality CXR image is generated.

Although the base SD model already demonstrates strong text-to-image capabilities, domain-specific fine-tuning is essential to ensure clinical relevance and fidelity in medical applications. To this end, we preserve the core capabilities of the pre-trained model by introducing a lightweight fine-tuning module. Specifically, we only train the SD encoder $E_{sd}$ and a set of $1 \times 1$ zero-convolution layers, whose weights are initially set to zero to avoid disrupting the original feature representations. These parameters are gradually learned during training, enabling the model to adapt to the medical domain while maintaining efficiency.

During fine-tuning, the input clinical report is first passed through the text encoder and a zero-convolution layer. The resulting embedding is concatenated with noise $z_t$ and fed into the SD encoder. This encoder comprises four downsampling blocks, each linked to the corresponding upsampling blocks through trainable zero-convolution layers. The resulting features are then fused with the outputs of the frozen pre-trained SD encoder, producing the final CXR image.

By restricting training to a small subset of parameters, our approach offers rapid fine-tuning without compromising the pre-trained model's performance. This strategy enables the generation of clinically relevant and fine-grained chest X-ray images, supporting downstream medical applications efficiently.

### 4.2.3 Multi Model Tasks

The proposed MedXChat framework, illustrated in Fig. 6.1 (bottom right), offers a unified solution that supports three key tasks: CXR-to-Report, CXR-VQA, and Text-to-CXR, detailed as follows:

**CXR-to-Report.** In this task, a chest X-ray image is first processed by a pre-trained CLIP visual encoder (ViT-L/14) [69] to extract visual features that are aligned with textual semantics. These features are then combined with a user instruction, such as *"Generate a report of this chest X-ray"*, and fed into the LLM. The LLM subsequently generates a concise radiology report based on the visual information embedded in the input image.

**CXR-VQA.** This task involves three inputs: a CXR image, its associated medical report, and a user-provided question, such as *"Is there cardiomegaly?"*. The LLM processes these multimodal inputs and outputs a response based on its analysis. For example, it may respond with *"No, cardiomegaly is not present"*, leveraging both the visual content and the textual context to provide accurate answers.

**Text-to-CXR.** In this generative task, the LLM is fine-tuned on the dialogue data described in Sec. 4.2.1. An example input might be:

*"Human: Hi there, I recently had a chest X-ray and the report mentioned 'right lower lobe consolidation'. I'm not sure what that means. Could you generate a lateral view of a chest X-ray with that condition for me?"*

Given this prompt, the LLM generates a medically relevant response and a textual description for image synthesis, for instance:

*Assistant: Absolutely. Consolidation in the right lower lobe of your lungs typically indicates an infection, such as pneumonia. Here is an image based on the description you provided:* $< Xray >$ *Lateral view of the chest was obtained. New right lower lobe consolidation is consistent with pneumonia.* $< /Xray >$

The LLM then extracts the content enclosed within the special tokens $< Xray >$ and $< /Xray >$ as the generation prompt. This prompt is passed to the CXR-SD module, which synthesizes a chest X-ray image matching the described condition. This enables MedXChat to provide

users with tailored visual representations in response to their clinical inquiries, thereby enhancing interpretability and engagement.

### 4.2.4 Training Objectives

Our training framework comprises two primary objectives: fine-tuning the LLM during the instruction tuning stage and optimizing the CXR-SD model during the preparation stage. The goal of the LLM is to generate structured target outputs that include an explanation, input prompt, and an autoregressive response—closely following the instruction-following paradigm established in ChatGPT training [90, 70, 6]. The loss is computed solely on the response tokens generated by the model, ensuring alignment with its autoregressive decoding mechanism.

In parallel, the CXR-SD model is optimized following the training strategy of the original Stable Diffusion (SD) framework [76], which centers on denoising. Specifically, it aims to predict and remove the noise added to input images across a series of temporal steps, guided by text-based prompts. Through this iterative denoising process, the model progressively reconstructs high-fidelity chest X-ray images, restoring structural details step by step.

**Large Language Model.** With the dialogues $\left(\mathbf{X}_{\mathrm{q}}^{1}, \mathbf{X}_{\mathrm{a}}^{1}, \cdots, \mathbf{X}_{\mathrm{q}}^{T}, \mathbf{X}_{\mathrm{a}}^{T}\right)$ obtained utilizing ChatGPT-4 in Sec. 4.2.1, where $\mathbf{X}_{\mathrm{q}}$ is the question, $\mathbf{X}_{\mathrm{a}}$ is the answer, and T is the total number of dialogues, we fine-tune LLM with LoRA [37] through minimizing the loss $L_a$:

$$L_a = \sum_{i=1}^{L} -\log p\left(\mathbf{X}_{\mathrm{a}}^{i} \mid \mathbf{X}_{\mathrm{q}}^{t}, < i, \mathbf{X}_{\mathrm{a}}^{t}, < i\right) \tag{4.3}$$

for a sequence of length L. As a result, the enhanced LLM is now adept at crafting intelligent and contextually relevant medical responses.

**Stable Diffusion.** Given an initial image $z_0$, the image diffusion process incrementally introduces noise, resulting in a sequence of increasingly noisy images $z_t$, where $t$ denotes the timestep of noise addition. Under a set of conditions, comprising the timestep $t$ and textual prompts $c_t$, a network $\epsilon_\theta$ is trained to predict the noise pattern that has been added

to the noisy image $z_t$, enabling the reverse process of noise reduction to reconstruct the original image. The overall learning objective for the entire diffusion model is to minimize the loss $\mathcal{L}_{SD}$:

$$\mathcal{L}_{SD} = \mathbb{E}_{z_0, \boldsymbol{t}, \boldsymbol{c}_t, \epsilon \sim \mathcal{N}(0,1)} \left[ \| \epsilon - \epsilon_\theta \left( \boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t \right) ) \|_2^2 \right]. \tag{4.4}$$

## 4.3 Experiment Settings

We benchmark our model against key state-of-the-art (SOTA) models for CXR-related multimodal generation using LLMs. For CXR-to-Report, our model's performance is mainly benchmarked against UniXGen [53], XrayGPT [91], and LLM-CXR [54]. For CXR-VQA, we compare with XrayGPT [91], ELIXR [114], and LLM-CXR [54]. For Text-to-CXR, we evaluate our model against UniXGen [53] and LLM-CXR [54]. It is highlighted that LLM-CXR and our model are uniquely equipped to handle all of these diverse tasks using a single integrated model. Whereas, XrayGPT [91] cannot handle Text-to-CXR generation and UniXGen cannot handle CXR-VQA task.

### 4.3.1 Implementation Details

We set the learning rate as $2 \times 10^{-5}$, the number of epochs as 5, and the batch size as 2. We train our model with 8.4M and 7.1B trainable and frozen parameters on two NVIDIA RTX 3090 GPUs. The deployed MedXChat model contains 15.5B parameters. On a single RTX 3090 GPU, the average inference time is approximately 2 s per study for CXR-to-Report generation, 2 s per query for CXR-VQA, and 10 s per study for Text-to-CXR synthesis.

**Downstream Classification.** To evaluate the clinical utility of our Text-to-CXR generation, we apply our generated CXRs to classifying the 14 diagnostic categories in MIMIC-CXR. We adopt a classification model proposed in [32], which gains high classification accuracy from its increased model depth via a residual learning framework, train it by all CXR images in MIMIC-CXR training set, and test it on the generated CXR images produced by different models.

### 4.3.2 Dataset

We validate the methods on the MIMIC-CXR dataset [43], which stands as the largest publicly available dataset encompassing chest radiographs along with free-text reports. It covers 14 categories related to lung disease diagnosis. Part of the CXR images are obtained in both PA and lateral views. In our research, we followed the official dataset division guidelines of MIMIC-CXR to maintain consistency and enable fair comparisons. Consequently, our training dataset comprises 270,790 samples of image-report pairs, accompanied by 2,130 and 3,858 samples for validation and testing, respectively.

## 4.4 Evaluation and Discussion

### 4.4.1 Evaluation Metrics

**CXR-to-Report.** We include both NLP standard evaluation metrics such as BLEU-4 [67], METEOR [5], ROUGE-L [59], and CIDEr [98] to evaluate the quality of the generated diagnostic reports, and the AUROC and F1 score metrics to assess the accuracy of disease classification based on generated reports. As per [54], we report AUROC and F1 in three forms: Micro, Macro, and Weighted, derived from the frequency deviation of categories, with each category being treated as equally important and weighted average, respectively. They are calculated based on six diagnostic categories: No Findings, Pneumothorax, Edema, Pleural Effusion, Consolidation or Pneumonia, and Lung lesion.

**CXR-VQA.** We follow the VQA performance evaluation framework from ELIX-R [114] to enable the comparison with other LLM-based methods[1]. Specifically, from the MIMIC-CXR dataset, we randomly choose eight samples from each of the six distinct diagnostic categories. We either inquiry about specific lesions or query the presence, location, and severity of the findings in each CXR image.

[1] This choice is necessitated by the unavailability of CXR-VQA results for other LLM-based methods

**Table 4.1:** CXR-to-Report: AUROC and F1. † marks quoted results from [54].

| AUROC↑ | Micro | Macro | Weighted | NoF. | Pmtx. | Edem. | PEff. | Csdn./Pna. | LLsn. |
|---|---|---|---|---|---|---|---|---|---|
| UniXGen-256† [53] | 0.577 | 0.533 | 0.541 | 0.564 | 0.530 | 0.542 | 0.533 | 0.516 | 0.513 |
| XrayGPT† [91] | 0.595 | 0.552 | 0.576 | 0.592 | 0.511 | 0.590 | 0.595 | 0.515 | 0.511 |
| LLM-CXR† [54] | 0.654 | 0.586 | 0.628 | **0.698** | 0.532 | 0.612 | 0.635 | **0.540** | 0.501 |
| MedXChat (Ours) | **0.672** | **0.599** | **0.666** | 0.583 | **0.549** | **0.633** | **0.783** | 0.530 | **0.519** |
| F1↑ | Micro | Macro | Weighted | NoF. | Pmtx. | Edem. | PEff. | Csdn./Pna. | LLsn. |
| UniXGen-256† [53] | 0.281 | 0.187 | 0.256 | 0.411 | 0.083 | 0.226 | 0.215 | 0.132 | 0.055 |
| XrayGPT† [91] | 0.314 | 0.227 | 0.320 | 0.371 | 0.049 | 0.333 | 0.404 | 0.143 | **0.058** |
| LLM-CXR† [54] | 0.414 | 0.283 | 0.408 | **0.562** | 0.083 | 0.370 | 0.455 | **0.198** | 0.030 |
| MedXChat (Ours) | **0.420** | **0.292** | **0.436** | 0.318 | **0.092** | **0.398** | **0.718** | 0.177 | 0.049 |

**Table 4.2:** CXR-to-Report: NLP Metrics. † marks quoted results from respective papers.

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|
| Show-Tell [99] | 0.308 | 0.190 | 0.125 | 0.088 | 0.256 | 0.122 | 0.096 |
| Att2in [75] | 0.314 | 0.198 | 0.133 | 0.095 | 0.264 | 0.122 | 0.106 |
| AdaAtt [62] | 0.314 | 0.198 | 0.132 | 0.094 | 0.267 | 0.128 | 0.131 |
| Transformer [97] | 0.316 | 0.199 | 0.140 | 0.092 | 0.267 | 0.129 | 0.134 |
| M2transformer [17] | 0.332 | 0.210 | 0.142 | 0.101 | 0.264 | 0.134 | 0.142 |
| R2Gen† [14] | 0.353 | 0.218 | 0.145 | 0.103 | 0.277 | 0.142 | 0.141 |
| PPKED† [60] | 0.36 | 0.224 | 0.149 | 0.106 | 0.284 | 0.149 | 0.237 |
| R2GenCMN† [13] | 0.353 | 0.218 | 0.148 | 0.106 | 0.278 | 0.142 | 0.143 |
| GSK† [117] | 0.363 | 0.228 | 0.156 | **0.115** | 0.284 | - | 0.203 |
| LLM-CXR [54] | 0.196 | 0.095 | 0.054 | 0.033 | 0.245 | 0.081 | **0.445** |
| LLaVA-Med [55] | 0.232 | 0.086 | 0.027 | 0.009 | 0.168 | 0.082 | 0.015 |
| UniXGen-256 [53] | 0.365 | 0.227 | 0.147 | 0.101 | **0.294** | **0.156** | 0.138 |
| MedXChat (Ours) | **0.367** | **0.235** | **0.158** | 0.111 | 0.264 | 0.135 | 0.175 |

**Text-to-CXR.** We employ the Frechet Inception Distance (FID), a widely recognized metric for image generation quality assessment, in conjunction with a downstream classification task to evaluate the quality of the images produced by our model.

### 4.4.2 Performance Comparison

**CXR-to-Report.** We evaluate our model using AUROC and F1 scores, as presented in Table 4.1, which are widely recognized metrics for assessing classification performance and the clinical validity of generated radiology reports. AUROC (Area Under the Receiver Operating Characteristic curve) reflects the model's capability to distinguish between different disease conditions, while F1 score balances precision and recall, providing a holistic view of prediction quality in imbalanced clinical

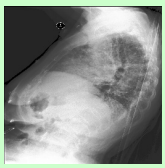
| Input CXR | Ground-Truth | LLM-CXR | Ours |
|---|---|---|---|
|  | impression : no acute cardiopulmonary process . findings : lungs are clear . there is no consolidation effusion or edema . the cardiomediastinal silhouette is within normal limits . no acute osseous abnormalities . | No acute cardiopulmonary process. | impression : no acute cardiopulmonary process . findings : the lungs are clear without focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is within normal limits . no acute osseous abnormalities . no free air below the right hemidiaphragm is seen . no pleural effusion or pneumothorax is seen . |
|  | impression : findings compatible with pulmonary fibrosis with likely superimposed edema . findings : ap and lateral views of the chest . low lung volumes are seen compatible with patients history of fibrosis . diffusely increased interstitial markings are seen throughout the lungs but these appear overall slightly worse when compared to prior . cardiomediastinal silhouette is grossly. | - | impression : mild pulmonary edema. findings : ap and lateral views of the chest were obtained. there is increased interstitial markings consistent with mild pulmonary edema. there is no pleural effusion or pneumothorax. the cardiac silhouette is enlarged but stable. mediastinal contours are unremarkable. |

**Figure 4.2:** Visual examples for the CXR-to-Report Task

datasets. Compared with existing models, our proposed MedXChat consistently outperforms all baselines in both AUROC and F1 across nearly all diagnostic categories. In terms of AUROC, MedXChat achieves the highest micro (0.672), macro (0.599), and weighted (0.666) averages, outperforming LLM-CXR [54], which was previously the strongest baseline. MedXChat also achieves the best AUROC scores in five out of seven disease categories, notably outperforming others in Pneumothorax (0.549), Edema (0.633), Pleural Effusion (0.783), and Lung lesion (0.519). In terms of F1 scores, which are particularly informative for evaluating performance on imbalanced datasets, MedXChat again ranks highest in overall micro (0.420), macro (0.292), and weighted (0.436) scores. The model outperforms LLM-CXR by a significant margin, especially in conditions like Pneumothorax (0.092 vs. 0.083), Edema (0.398 vs. 0.370) and Pleural Effusion (0.718 vs. 0.455), highlighting its robustness in handling both frequent and rare conditions. The non-LLM-based UniXGen [53] shows noticeably weaker results in both AUROC and F1, reinforcing the critical role of large language models in generating accurate and semantically coherent medical reports. These results confirm that MedXChat not only produces radiology reports with high discriminative ability (AUROC) but also ensures consistency and correctness in predictions (F1), making it a clinically valuable tool for automated CXR interpretation.

Table 5.4 presents a comprehensive comparison of natural language generation (NLG) metrics, including BLEU-1 through BLEU-4, ROUGE,

METEOR, and CIDEr. These metrics collectively assess lexical overlap, fluency, and semantic relevance between generated reports and ground-truth references. We evaluate not only large-scale multimodal models such as LLM-CXR [54] and UniXGen [53], but also traditional captioning-based methods like Show-Tell [99], Att2in [113], AdaAtt [62], Transformer [97], and M2Transformer [17]. MedXChat outperforms all previous approaches on most NLG metrics. Specifically, it achieves the highest scores in BLEU-1 (0.367), BLEU-2 (0.235), and BLEU-3 (0.158), reflecting strong n-gram consistency and high overlap with reference reports. Although GSK [55] narrowly surpasses MedXChat on BLEU-4 (0.115 vs. 0.111), our model shows superior performance in earlier BLEU levels, which are more indicative of high-fidelity surface-level phrasing. In terms of CIDEr, our model achieves a competitive score of 0.175, outperforming all baselines except for LLM-CXR (0.445). However, it is important to note that the CIDEr value reported for LLM-CXR is based solely on the impression section of the reports, which are shorter and more focused. When evaluating the full report, LLM-CXR's CIDEr score significantly decreases[2]. This suggests that MedXChat provides better overall consistency in generating complete and clinically relevant reports. Additionally, our model outperforms UniXGen [53] in all evaluated BLEU and CIDEr metrics, and performs comparably in ROUGE (0.264 vs. 0.294) and METEOR (0.135 vs. 0.156). Compared to text-only baselines, MedXChat exhibits much stronger linguistic coherence, indicating the advantages of integrating both visual and textual cues through a unified vision-language framework. These results highlight MedXChat's ability to generate radiology reports that are not only lexically accurate but also semantically meaningful and clinically reliable—an essential requirement for real-world deployment in medical AI systems.

Figure 4.2 presents qualitative comparisons between ground-truth radiology reports, the baseline model LLM-CXR [54], and our proposed MedXChat system. These visual examples offer valuable insights into

[2]As confirmed in our re-implementation, CIDEr on full-text LLM-CXR reports is substantially lower than the original value.

**Table 4.3:** CXR-VQA: Accuracy by topic. † marks quoted results from [54]. "SST" stands for Size, Severity, Type.

| Accuracy↑ | All | Presence | Location | SST |
|---|---|---|---|---|
| ELIXR† [114] | 54.8% | 64.5% | 41.0% | 25.0% |
| XrayGPT† [91] | 25.2% | 27.4% | 21.9% | 20.3% |
| RadFM† [110] | 32.7% | 34.5% | 31.3% | 20.8% |
| LLM-CXR† [54] | 44.8% | 41.3% | 50.0% | 62.5% |
| LLaVA-Med [55] | 53.1% | 53.8% | 51.0% | 56.3% |
| MedXChat (Ours) | **61.2%** | **61.5%** | **56.3%** | **68.8%** |

the generative capabilities of each model in producing clinically meaningful and comprehensive CXR reports. In the first case, LLM-CXR generates a minimal report with a single sentence, failing to include specific findings or an impression section. By contrast, MedXChat produces a well-structured report that closely follows clinical standards. It includes both an "Impression" summarizing the diagnostic outcome and detailed "Findings" describing the absence of consolidation, effusion, or pneumothorax. Additionally, it notes the clarity of the lungs and the normal appearance of the cardiomediastinal silhouette—details that are consistent with the ground-truth report written by a radiologist. The second example further demonstrates MedXChat's superiority in handling complex, multi-view cases. LLM-CXR fails to generate any output, while MedXChat produces a complete report that identifies "mild pulmonary edema," "increased interstitial markings," and "no pleural effusion or pneumothorax." Notably, it incorporates findings from both frontal and lateral views, referencing lateral chest views and describing relevant anatomical and pathological observations. This highlights MedXChat's ability to process and interpret multi-view inputs—a critical capability for real-world clinical scenarios. These qualitative comparisons demonstrate that MedXChat generates richer, more informative, and clinically aligned reports compared to baseline models. It not only captures key radiological features but also mimics professional reporting structure and language. These findings underscore the model's potential to assist radiologists in generating high-quality reports with reduced workload and increased diagnostic consistency.

**Table 4.4:** CXR-VQA: Accuracy by Diagnosis. † marks quoted results from [54].

| **Diagnosis** | XrayGPT† | RadFM† | LLM-CXR† | LLaVA-Med | MedXChat (Ours) |
|---|---|---|---|---|---|
| All | 25.2% | 32.7% | 44.8% | 53.1% | **61.2%** |
| No Findings | 42.5% | 61.3% | **71.3%** | 37.5% | 50.0% |
| Pneumothorax | 18.8% | 23.8% | 22.5% | 54.2% | **58.3%** |
| Edema | 26.3% | 31.3% | 53.8% | 58.3% | **72.9%** |
| Pleural Effusion | 20.0% | 26.3% | 53.8% | 60.4% | **79.2%** |
| Consolidation / Pneumonia | 25.0% | 34.4% | 39.1% | **56.3%** | 50.0% |
| Lung Lesion | 17.2% | **40.6%** | **50.0%** | 37.5% | 37.5% |

**CXR-VQA.** CXR-VQA performance is shown in Table 4.3 and Table 4.4. Based on the results presented in Table 4.3, we provide a comprehensive analysis of MedXChat's performance on the CXR-VQA task, focusing on three critical dimensions: Presence, Location, and SST (Size, Severity, and Type of abnormalities). MedXChat achieves the highest overall accuracy of 61.2%, surpassing all baseline models. This includes LLaVA-Med (53.1%), LLM-CXR (44.8%), and ELIXR (54.8%), demonstrating MedXChat's superior generalization ability and robust vision-language reasoning in clinical settings. In terms of Presence accuracy, which measures whether the model correctly identifies the existence of abnormalities, MedXChat achieves 61.5%, closely rivaling ELIXR (64.5%) while significantly outperforming XrayGPT (27.4%) and RadFM (34.5%). This suggests MedXChat is highly reliable in initial diagnostic screening scenarios where sensitivity to abnormal findings is crucial. For Location accuracy, MedXChat again leads with 56.3%, exceeding LLaVA-Med (51.0%) and LLM-CXR (50.0%). This indicates MedXChat's enhanced capability to interpret spatial cues in radiographic images and pinpoint the anatomical regions of concern—critical for clinical decisions involving targeted treatment. Most notably, in the SST category—arguably the most challenging due to the need for nuanced understanding of lesion

size, severity, and type—MedXChat reaches an impressive 68.8% accuracy. This clearly outperforms LLM-CXR (62.5%) and far surpasses models like ELIXR (25.0%), reflecting MedXChat's fine-grained comprehension of complex radiological descriptors. MedXChat consistently outperforms all baselines across all dimensions in the CXR-VQA task, establishing itself as a highly effective model for radiology-focused visual question answering. Its strengths in multi-faceted diagnostic understanding suggest strong clinical applicability, especially for generating interpretable, high-precision answers in real-world medical settings. Future improvements may focus on integrating multi-view imaging data and enhancing interpretability through multimodal attention mechanisms.

Based on the results in Table 4.4, MedXChat demonstrates superior diagnostic understanding across multiple chest pathologies within the CXR-VQA task. With an overall accuracy of 61.2%, MedXChat significantly outperforms all comparison baselines, including LLaVA-Med (53.1%), LLM-CXR (44.8%), and RadFM (32.7%), reinforcing its effectiveness in answering clinically relevant visual questions. In Edema diagnosis, MedXChat achieves an accuracy of 72.9%, outperforming LLaVA-Med (58.3%) and LLM-CXR (53.8%) by a wide margin. This result highlights the model's ability to recognize diffuse patterns and subtle fluid accumulations that are often challenging for automated systems. For Pleural Effusion, MedXChat attains the highest accuracy of 79.2%, surpassing all other baselines including LLaVA-Med (60.4%) and RadFM (26.3%). This demonstrates the model's precision in detecting pleural fluid abnormalities that require high spatial and contextual understanding. In Pneumothorax, a condition that often presents with subtle yet critical visual cues, MedXChat reaches 58.3% accuracy—more than doubling the performance of XrayGPT (18.8%) and significantly outperforming LLM-CXR (22.5%). This suggests the model is capable of identifying life-threatening findings with high sensitivity. Even for more common diagnoses like Consolidation/Pneumonia, MedXChat delivers competitive performance (50.0%), only slightly behind LLaVA-Med (56.3%). In No Findings, the model balances specificity and sensitivity with a solid 50.0% score. While Lung Lesion remains a challenging category for most

models—including MedXChat (37.5%)—the overall pattern shows MedXChat's consistent superiority in handling both common and complex diagnostic labels. These results underscore MedXChat's potential for real-world clinical use, particularly in tasks demanding accurate and interpretable answers to diagnostic queries across a wide spectrum of thoracic conditions. Future enhancements may focus on underrepresented categories like Lung Lesion through targeted data augmentation or multi-task learning strategies.

Figure 4.3 presents a representative visual question answering (VQA) example, highlighting MedXChat's ability to handle complex clinical queries grounded in radiographic content. In this scenario, the model accurately identifies findings such as opacities in the right lower lobe and left infrahilar area, and provides clinically coherent answers like pneumonia when asked about their implications. Furthermore, when prompted about the recommended course of action, the model appropriately responds with follow-up to resolution, reflecting an understanding of standard clinical management pathways. Importantly, it also shows diagnostic reasoning capability—for example, in response to the question "Is there a possibility of a pulmonary nodule being present?", the model states: "Yes, a pulmonary nodule cannot be excluded", demonstrating its ability to express uncertainty in a medically appropriate manner. In follow-up queries such as "What should be done if the finding of the nodule persists?", the model suggests further evaluation with chest radiographs, aligning with common clinical protocols. Finally, for anatomical or functional insights like pulmonary vascular engagement, the model correctly identifies mild pulmonary vascular engagement is present, indicating its nuanced understanding of subtle radiographic features. This example illustrates not only the model's capacity to extract and contextualize visual features but also its ability to respond to a variety of question types—diagnostic, procedural, and descriptive—mirroring the nature of real-world clinical dialogues. Such capabilities are crucial for enabling explainable AI and interactive medical assistants in clinical workflows.

**Text-to-CXR.** Table 5.6 presents quantitative comparisons for the Text-to-CXR synthesis task, highlighting two key evaluation metrics: Fréchet

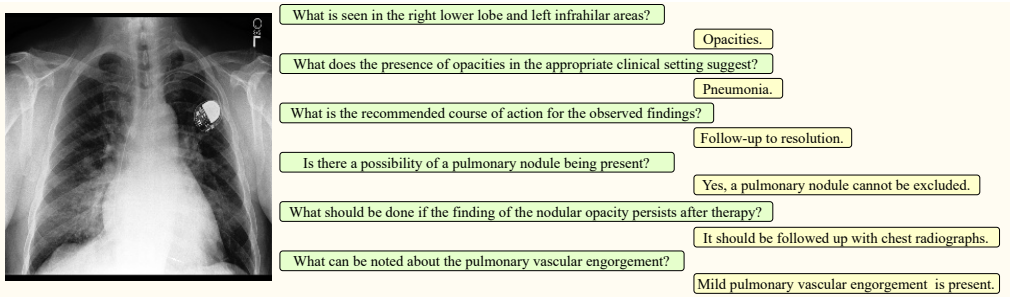



**Figure 4.3:** Visual examples for the CXR-VQA Task

**Table 4.5:** Text-to-CXR: FID and classification accuracy.

| Method | FID↓ | Classification Accuracy↑ |
|---|---|---|
| UniXGen [53] | 106.17 | 67.2% |
| LLM-CXR [54] | 73.29 | 68.6% |
| MedXChat (Ours) | **43.46** | **71.5%** |

Inception Distance (FID) and classification accuracy using a downstream diagnostic classifier. Our model, MedXChat, achieves the lowest FID score of 43.46, substantially outperforming UniXGen (106.17) and LLM-CXR (73.29). This significant reduction in FID indicates that the synthetic chest X-rays generated by MedXChat are much closer in distribution to real CXR images, reflecting superior visual fidelity and text-image alignment. To further assess clinical validity, a classifier trained on real MIMIC-CXR images was used to evaluate the generated CXRs. MedXChat again outperforms all baselines with a classification accuracy of 71.5%, surpassing both LLM-CXR (68.6%) and UniXGen (67.2%). This result implies that not only are the generated images visually similar to real data, but they also preserve critical diagnostic information necessary for downstream tasks. It demonstrates MedXChat's strong capability in generating medically plausible images conditioned on clinical prompts. Moreover, the consistently poor performance of the non-LLM-based UniXGen across both metrics reinforces the conclusion that LLM-based frameworks are better equipped to bridge the semantic gap between complex clinical text and image content. These results, combined with earlier evaluations in report generation and VQA, underscore MedXChat's versatility and effectiveness as a unified multimodal medical assistant.

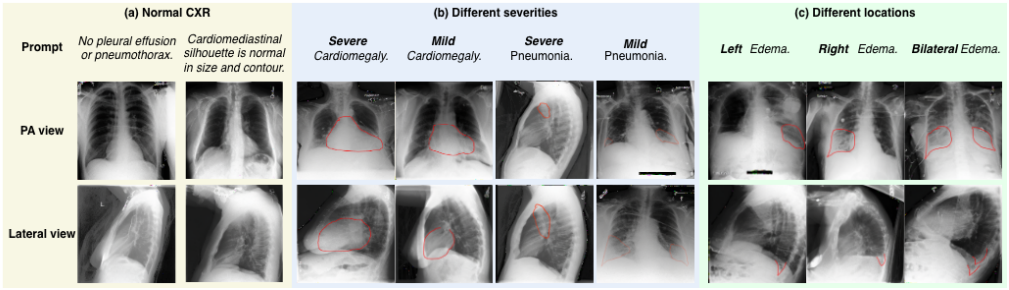



**Figure 4.4:** Visual examples showcasing our Text-to-CXR generation. The delineated image regions correspond to the users' prompts, expertly identified by a trained Radiologist.

**Table 4.6:** Radiologist evaluation on 40 randomly selected generation cases.

| #Prefer / Total | Reports | CXRs |
|---|---|---|
| UniXGen [53] | 1/20 | 7/20 |
| LLM-CXR [54] | 0/20 | 0/20 |
| MedXChat (Ours) | **19/20** | **13/20** |

Figure 4.4 presents compelling qualitative results that highlight MedXChat's superior capability in Text-to-CXR generation across multiple axes of clinical relevance. Each column illustrates a different textual prompt condition—ranging from normal findings (left), varying disease severity (middle), to spatial distribution of pathology (right). In all cases, the model is able to synthesize highly realistic CXRs that align closely with the specified clinical descriptors. One notable strength of our model is its capacity to generate both posteroanterior (PA) and lateral-view CXRs—a feature absent in existing methods such as LLM-CXR and UniXGen, which are limited to PA views. This dual-view generation greatly enhances the model's potential in supporting real-world radiology workflows, where lateral views are critical for diagnosing certain thoracic conditions. Furthermore, the model's responsiveness to subtle semantic differences in prompts is particularly impressive. For instance, in the "Severe Cardiomegaly" vs. "Mild Cardiomegaly" examples (Fig. 4.4, middle column), the generated heart silhouettes appropriately reflect variations in heart size and contour. Similarly, the rightmost set of examples demonstrates how MedXChat accurately differentiates between left, right, and

bilateral edema, producing anatomically coherent features that reflect these distinctions. Importantly, the red contour overlays in the figure represent expert radiologist annotations, validating the alignment between generated pathologies and their intended textual descriptions. This further reinforces the clinical plausibility and controllability of our synthesis process. Collectively, these qualitative results underscore MedXChat's strength in translating nuanced medical text into diagnostically meaningful images.

**Radiologist Evaluation.** To further validate the clinical effectiveness of MedXChat, we conducted a blinded expert evaluation by a board-certified radiologist, assessing 40 randomly selected samples that included 20 radiology reports and 20 posteroanterior (PA) CXR images generated by MedXChat, UniXGen, and LLM-CXR, as summarized in Table 4.6. Due to the inability of UniXGen and LLM-CXR to synthesize lateral-view CXRs, only frontal-view generations were considered to ensure a fair comparison. The evaluation results reveal a striking preference for MedXChat: 95% of generated reports (19/20) and 65% of generated CXRs (13/20) were favored over those from competing methods. In contrast, UniXGen received minimal preference (1 report and 7 CXRs), while LLM-CXR was not preferred in any instance. These findings provide strong evidence that our model not only produces clinically relevant and coherent textual content but also generates realistic and diagnostically meaningful CXR images. The radiologist's overwhelming endorsement underscores the potential of MedXChat as a reliable and interpretable AI assistant in real-world medical settings.

### 4.4.3 Limitation

While MedXChat demonstrates superior performance across CXR-related tasks, several limitations remain. First, the model's reliance on the MIMIC-CXR dataset may limit its generalizability to images from different institutions or imaging devices. Second, although our framework supports fine-grained CXR generation, ensuring clinical realism under diverse disease combinations or rare conditions remains a challenge. Lastly, the instruction-tuning process, while effective, is dependent on synthetic

dialogues, which may not fully capture real-world clinical interactions. Future work will explore domain adaptation and the incorporation of real physician-patient conversations to enhance model robustness and usability.

## 4.5 Summary

This chapter introduces MedXChat, a unified multimodal large language model framework tailored for comprehensive chest X-ray (CXR) analysis. MedXChat seamlessly integrates three key tasks—CXR-to-Report, CXR-VQA, and Text-to-CXR—within a single architecture. By leveraging instruction tuning and fine-tuning strategies on both language and image generation components, MedXChat effectively understands and generates clinically relevant content from multimodal inputs. Extensive experiments and expert evaluations validate its superiority over existing approaches in diagnostic accuracy, language quality, and image fidelity, demonstrating its strong potential as an AI assistant for medical imaging applications.

Our contributions can be summarized as follows.

- We introduce MedXChat, a unified LLM framework that supports both interpretive and generative functions for CXR images by facilitating the bidirectional exchange of information between medical texts and CXR images. Our framework achieves superior performance in CXR-to-Report, CXR-VQA, and Text-to-CXR tasks compared to existing multimodal benchmarks.
- We have pioneered an innovative Text-to-CXR synthesis method that leverages the instruction-following capabilities of an off-the-shelf Stable Diffusion model. This approach preserves the fine-grained feature generation essential for high-quality CXR imaging while exploiting the robust generative power of a large pre-trained model.
- MedXChat also extends its innovation to practical applications by using natural language as input, thereby breaking traditional task

boundaries. This unified approach simplifies the training process for medical professionals by integrating diverse tasks into a single environment, and it enhances clinical utility by providing rich visual context that can support radiologist education and consultation.

- We have curated a substantial set of instructional data from the MIMIC-CXR dataset, specifically tailored for Text-to-CXR generation. Moreover, we fine-tuned our Stable Diffusion model using a comprehensive collection of medical data. All these resources will be open-sourced to advance further research in medical image generation.

# Chapter 5

# Constructing A Unified Vision-Language Model for Chest X-Ray Diagnosis, Medical Education, and Data Augmentation

In this chapter, we present a systematic validation of MedXChat, a unified multimodal large language model framework meticulously designed for CXR applications. MedXChat integrates three core functionalities—disease diagnosis, medical education, and data augmentation—within a coherent and complete architecture. To comprehensively assess its performance and clinical applicability, we evaluate MedXChat using both computational metrics and a structured clinical assessment conducted by a panel of radiologists, comprising three junior radiologists with 3 years of experience, two senior radiologists with 7–10 years of experience, and one supervising radiologist with 25 years of experience. The evaluation results demonstrate that MedXChat achieves high efficiency and reliability across diverse diagnostic and educational tasks in real-world healthcare scenarios.

## 5.1 Introduction

In the previous chapter, we introduced MedXChat, a unified LLM framework meticulously designed for CXR applications. Unlike existing systems that focus on isolated tasks, MedXChat integrates three complementary functionalities—disease diagnosis, medical education, and data augmentation—within a single, coherent architecture. This design enables the model to address a wide range of clinical and educational scenarios, from automated report generation and interactive VQA to high-fidelity CXR image synthesis for training and algorithm development. Such integration not only enhances the versatility of the system but also facilitates cross-modal interactions, allowing information from one task to inform and improve another. However, the validation of MedXChat thus far has relied primarily on conventional computational metrics, which, despite being widely adopted, exhibit several well-documented limitations that may compromise the reliability and clinical relevance of the evaluation.

Automated evaluation of text generation systems—such as those developed for radiology report generation—typically involves comparing generated reports with reference reports to assess semantic accuracy. Conventional natural language generation (NLG) metrics, such as BLEU [67], remain the most frequently used tools for this purpose. These metrics primarily quantify n-gram overlap between the generated and reference texts, providing a surface-level measure of similarity. However, they often overlook deeper aspects of lexical variety, syntactic diversity, and semantic equivalence, all of which are crucial for accurately conveying clinical meaning. Prior studies [57] have repeatedly highlighted inherent limitations of such n-gram-based approaches. First, because these metrics rely on rigid string-matching, they often fail to handle paraphrasing effectively. For example, BLEU and METEOR [73] may disproportionately reward a particular phrasing while penalizing semantically equivalent variations that deviate from the reference structure. To mitigate this, embedding-based methods such as BERTScore [123] have been introduced, leveraging contextualized token

representations to better capture semantic similarity beyond mere surface form.

Moreover, capturing clinically relevant content poses another major challenge. Standard NLG metrics often struggle to assess whether critical diagnostic entities—such as lesion types, anatomical locations, or severity indicators—are correctly identified and reported. Consequently, researchers have increasingly paired these metrics with clinical information extraction frameworks such as CheXbert [85] and RadGraph [42], which produce F1 scores for clinically relevant entities or relationships. This combination offers a more domain-aware perspective on evaluation. Nevertheless, even when such metrics are combined, their alignment with expert human judgment remains suboptimal [57]. To address this, RadCliQ [119] was proposed, linearly combining BLEU, BERTScore, CheXbert, and RadGraph F1 scores, with weights learned from regression against human-marked error scores. While RadCliQ has shown improved correlation with radiologists' assessments, its reliance on a limited and expensive set of human-annotated samples restricts scalability. Furthermore, its scoring scheme inherently places more weight on errors in clinical findings than on linguistic fluency, potentially overlooking issues in readability, coherence, and overall narrative quality—factors that directly influence clinical usability.

A similar challenge arises in the evaluation of medical image generation. One of the most widely adopted metrics in this domain is the Fréchet Inception Distance (FID) [34], which measures the statistical distance between the feature distributions of generated and real images in the Inception network's embedding space. While FID has been highly influential in assessing generative adversarial networks (GANs) and other natural image synthesis models, it is not inherently suited for medical imaging tasks. First, the Inception model is pretrained on natural images (ImageNet), which means its learned features are biased toward everyday objects and textures rather than subtle pathological patterns—such as microcalcifications or faint pulmonary opacities—that are critical for accurate diagnosis. As a result, FID may rate two images as similar despite clinically significant differences. Second, FID focuses on global

distribution alignment and does not assess whether specific anatomical structures or lesions are faithfully represented, potentially overestimating the quality of images that are visually appealing but diagnostically misleading. Third, FID scores are highly sensitive to dataset size and distribution, producing unstable results when the number of generated samples is small or when class imbalance exists—common scenarios in rare disease modeling and specialized radiology datasets. Furthermore, FID does not account for clinical task performance: a generated image could achieve a low FID score yet still fail to support downstream diagnostic tasks.

These limitations underscore a broader challenge: evaluation metrics developed for open-domain text or natural images often fail to capture the nuanced, domain-specific requirements of medical AI systems. For a unified framework like MedXChat—which spans both text-based and image-based modalities—exclusive reliance on such generic metrics risks underestimating or misrepresenting its true clinical value. Rigorous validation, therefore, demands domain-adapted, task-aware, and clinically interpretable evaluation frameworks that jointly assess semantic correctness and diagnostic relevance.

To this end, we adopted a dual evaluation strategy for MedXChat. Conventional computational metrics were retained to provide quantitative benchmarks, while a structured clinical evaluation was conducted to ensure that the results reflected real-world applicability and reliability. A panel of six radiologists—comprising three junior radiologists with 3 years of experience, two senior radiologists with 7–10 years of experience, and one supervising radiologist with 25 years of expertise— systematically reviewed and scored MedXChat's outputs across automated report generation, VQA, and high-fidelity CXR synthesis, following predefined clinical criteria. The evaluation revealed that MedXChat not only achieved strong performance on computational benchmarks but also demonstrated high efficiency and robust reliability in expert assessments, confirming its feasibility and clinical potential across diverse diagnostic scenarios. Beyond these quantitative and expert-driven assessments, feedback from the six participating radiologists also highlights

MedXChat's strong potential for medical education. Through hands-on use across the three core tasks—report generation, VQA, and CXR synthesis—the clinicians noted that the model provides clear reasoning, clinically aligned explanations, and diverse examples that can support trainees in understanding diagnostic patterns and refining interpretive skills. Their evaluations indicate that MedXChat not only performs reliably in clinical contexts but also offers meaningful pedagogical value by facilitating interactive learning and enabling medical students to engage with complex radiological concepts in a structured and accessible manner.

## 5.2 Method

### 5.2.1 Model Recap and Evaluation

As illustrated in Figure 5.1, MedXChat is developed as a unified multimodal large language model tailored for CXR–related applications. It supports three complementary core functionalities: CXR-to-Report automated report generation, CXR-VQA interactive visual question answering, and Text-to-Image high-fidelity CXR image synthesis. By integrating these capabilities within a single coherent architecture, MedXChat is capable of addressing diverse clinical and educational scenarios. The system is designed for both professional use—including radiologists, clinicians, and medical students—and non-professional use, such as by patients, caregivers, and healthcare agencies, thus accommodating multiple user roles and bridging the gap between expert-driven diagnostics and broader public accessibility.

The model construction process involves comprehensive dataset integration, targeted model fine-tuning, and task-specific adaptation. For report generation, MedXChat is trained on the MIMIC-IV dataset; for CXR-VQA, it leverages the MIMIC-CXR-VQA dataset; and for image synthesis, it utilizes the MIMIC-IV dataset augmented with a synthetic "MIMIC-T2II" dataset generated via a large language model to enrich

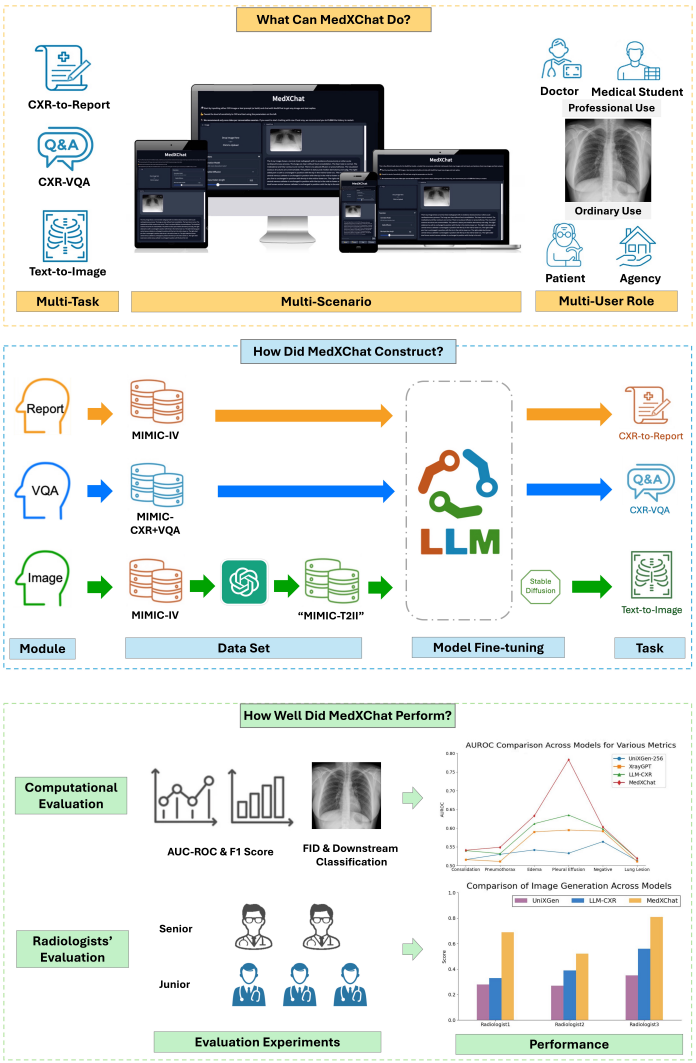



**Figure 5.1:** The study workflow for MedXChat development and evaluation.

training diversity. A Stable Diffusion backbone is incorporated for producing high-quality, realistic CXR images. These task-specific components are seamlessly unified under the multimodal LLM framework, enabling the model to be jointly fine-tuned for robust performance across all three task domains—thereby enhancing both cross-task generalization and modality-specific optimization.

Performance evaluation follows a dual-assessment strategy, combining quantitative computational metrics with qualitative expert clinical assessment. Computational evaluation employs AUC-ROC and F1

scores for text-based tasks, and FID together with downstream classification accuracy for assessing image synthesis quality and diagnostic utility. Complementing these automated measures, a structured clinical evaluation is conducted by a panel of six radiologists: three junior radiologists (3 years of experience), two senior radiologists (7–10 years of experience), and one supervising radiologist (25 years of experience). The experts systematically score MedXChat's outputs for accuracy, clarity, and clinical relevance across all tasks. Evaluation results consistently show that MedXChat outperforms strong baselines, demonstrating high efficiency, diagnostic reliability, and adaptability across real-world diagnostic and educational scenarios.

### 5.2.2 Dataset Construction

Figure 5.2 presents an overview of the dataset selection and preparation pipeline for the three tasks evaluated in this study: CXR-to-Report generation, CXR-VQA, and Text-to-Image synthesis. Each task utilizes data derived from the MIMIC-CXR dataset, with task-specific processing, sampling, and filtering to ensure consistency, clinical relevance, and training efficiency.

For the CXR-to-Report task, we employed the MIMIC-CXR dataset—the largest publicly available chest radiography dataset—comprising 377,110 images from 227,835 radiographic studies conducted at Beth Israel Deaconess Medical Center, Boston, MA. Each study includes DICOM-format images and corresponding free-text radiology reports. The dataset covers 14 diagnostic categories, including pneumonia, cardiomegaly, atelectasis, pleural effusion, and others. Some studies contain both PA and lateral views. To ensure consistency across experiments, we adhered to the official train/validation/test splits released with the dataset.

During preprocessing, we excluded samples with missing reports, low-resolution images, or ambiguous labels. Furthermore, we applied stratification to maintain balanced disease distribution across subsets, mitigating potential bias toward more common pathologies. After filtering, the final dataset for this task includes 270,790 image-report pairs for

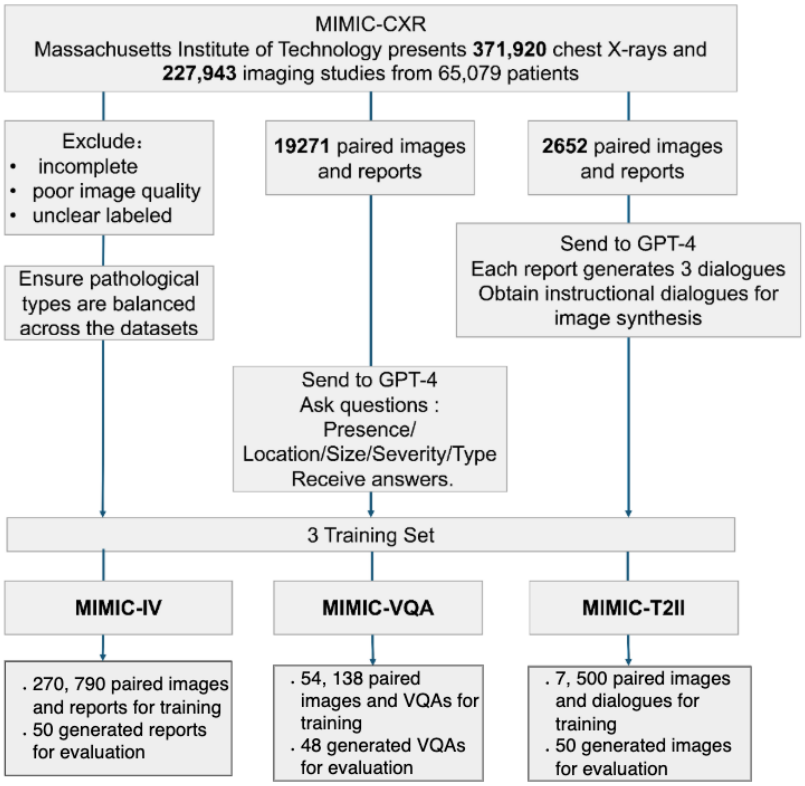



**Figure 5.2:** Flowchart illustrates the dataset preparation and training workflow for MedXChat. After excluding incomplete, poor-quality, and unclearly labelled data selected from the MIMIC-CXR dataset, 19,271 paired images and reports were processed for visual question answering (VQA) using GPT-4. Additionally, 2,652 paired images and reports were used to generate instructional dialogues for image synthesis. The data was split into three training sets: MIMIC-IV for report generation, MIMIC-VQA for question answering, and MIMIC-T2II for text-to-image synthesis, with distinct evaluation sets.

training.

For the CXR-VQA task, we adopted the data construction method from LLM-CXR [54], selecting 19,217 paired images and reports from MIMIC-CXR. These pairs were further processed using GPT-4, guided by instruction prompts adapted from the ELIXR benchmark [114]. Each image-report pair was used to generate multiple VQA samples, addressing key diagnostic questions across four dimensions: presence, location, size, and severity/type of radiological findings. The final dataset consists of 54,138 image-question-answer triplets for training, offering a rich

**Table 5.1:** The computational metrics and two-level radiologists' evaluation criteria to compare the performance of MedXChat with other LLM models in multitasks.

| Task | Compared Models | Computational Metrics | Junior Radiologists | Senior Radiologists |
|---|---|---|---|---|
| CXR-to-Report | LLM-CXR<br>UniXGen | AUC<br>F1 | 0 = non-compliant<br>0.5 = partially compliant<br>1 = fully compliant | 1. Accuracy (impression and lesion localization)<br>2. Detail (complete structures and description)<br>3. Location (left/right, lower/upper)<br>4. Severity (mild, moderate, severe)<br>5. Consistency (descriptions and impressions)<br>6. Tone (objective)<br>7. Comparison (clinical history)<br>8. Terminology (medical terms and grammar)<br>Scoring each item from 0 to 10 points |
| CXR-VQA | XrayGPT<br>RadFM<br>LLM-CXR<br>LLaVA-Med | Accuracy | 0 = non-compliant<br>0.5 = partially compliant<br>1 = fully compliant | None |
| Text-to-CXR | LLM-CXR<br>UniXGen | FID<br>Classification | 0 = non-compliant<br>0.5 = partially compliant<br>1 = fully compliant | 1. Consistency (alignment with the report)<br>2. Clarity (sharpness/contrast)<br>3. Integrity (completeness of chest structures)<br>4. Centering (symmetry and central positioning)<br>5. Recognizability (identifiability of lesions)<br>Scoring each item from 0 to 10 points |

resource for evaluating multimodal understanding.

For the Text-to-Image generation task, we created a specialized instruction dataset tailored for image synthesis in the clinical domain. From the 14 disease categories in MIMIC-CXR, we sampled 200 reports per category, resulting in 2,652 unique reports. Each report was processed using GPT-4 with carefully crafted instruction prompts to simulate three interactive dialogues per report, generating medically grounded instructions for chest X-ray image synthesis. This yielded a total of approximately 7,500 instruction-dialogue pairs for training.

For radiologist evaluation, we randomly selected 50 CXRs for report generation and 48 CXRs evenly sampled from six disease categories—No Findings, Pneumothorax, Edema, Pleural Effusion, Consolidation/ Pneumonia, and Lung Lesion (eight CXRs per category) —for question answering on Presence, Location and Size, Severity, and Type. In addition, 50 generated CXRs derived from dialogues were included.

## 5.3 Radiologists' Evaluation

### 5.3.1 Evaluation Implementation

As shown in Table 5.1, we evaluated the performance of MedXChat through both computational metrics and an expert panel consisting of 6 radiologists. An expert panel of six board-certified radiologists was invited to assess the clinical quality of outputs, while standard computational metrics were used to benchmark performance across three core tasks: CXR-to-Report, CXR-VQA, and Text-to-Image generation.

We conducted a comprehensive expert evaluation involving six board-certified radiologists to assess the clinical quality of outputs generated by MedXChat and baseline models. The panel included three junior radiologists (XX1, XX2, XX3, each with 3 years of experience), two senior radiologists (XX4, XX5, with 7–10 years of experience), and one supervisor radiologist (XX6, with 25 years of experience). Junior radiologists independently ranked for results generated by UniXGen, LLM-CXR, and our model. Senior radiologists evaluated the detailed criteria of generated reports and images. The supervisor radiologist resolved any uncertain scores and biases. All evaluations were performed in a blinded setting—radiologists were unaware of which model generated which output. Scores were assigned using a standardized 3-point scale: 0 for non-compliant, 0.5 for partially compliant, and 1 for fully compliant.

For the CXR-to-Report task, we selected 50 generated reports from UniXGen, LLM-CXR, and our MedXChat model, consistent with the setup used in the computational evaluation. In the first phase, three junior radiologists independently scored each report based on its overall clinical coherence and informativeness. In the second phase, two senior radiologists evaluated each report using eight criteria, including accuracy, detail, lesion localization, severity assessment, internal consistency, tone, clinical comparison, and terminology usage. Scores that diverged by more than 0.3 between raters were reviewed and resolved by the supervisor radiologist.

In the CXR-VQA task, 48 image-question pairs were evaluated, covering six common clinical categories: Pneumothorax, Edema, Pleural Effusion, Consolidation/Pneumonia, Lung Lesion, and No Findings. The two senior radiologists and the supervisor each led a team of three junior radiologists, and all groups independently rated the answers generated by MedXChat and four baseline models: LLM-CXR, XrayGPT, RadFM, and LLaVA-Med. The scoring followed the ELIXR framework, considering the presence, location and size, severity, and type of each radiographic finding.

For the Text-to-Image task, 50 freestyle CXR reports were randomly selected from the MIMIC-CXR dataset and used to generate images with UniXGen, LLM-CXR, and MedXChat. The evaluation followed a two-phase process. First, three junior radiologists scored each image holistically. Then, two senior radiologists rated the images across five detailed criteria: alignment with the report, visual clarity, anatomical completeness, centering, and recognizability of key abnormalities. Discrepancies above 0.3 points were re-assessed by the supervisor. To further verify clinical fidelity, we also trained a diagnostic classifier on real CXRs and tested its performance on the generated images to evaluate whether they preserved diagnostic cues.

### 5.3.2 Evaluation Results

#### CXR-to-Report

As shown in Figure 5.3 (top), MedXChat achieved consistently superior performance in the radiologist-based evaluation compared to baseline models. The three junior radiologists assigned MedXChat an average score of 0.51, substantially higher than 0.26 for UniXGen and 0.18 for LLM-CXR. This performance gap was evident across all three evaluators: Radiologist1 scored MedXChat at 0.38 versus 0.31 (UniXGen) and 0.22 (LLM-CXR); Radiologist2 rated MedXChat at 0.43 compared to 0.21 and 0.18; and Radiologist 3 gave the highest score to MedXChat at 0.67, more than double the ratings for both baselines.

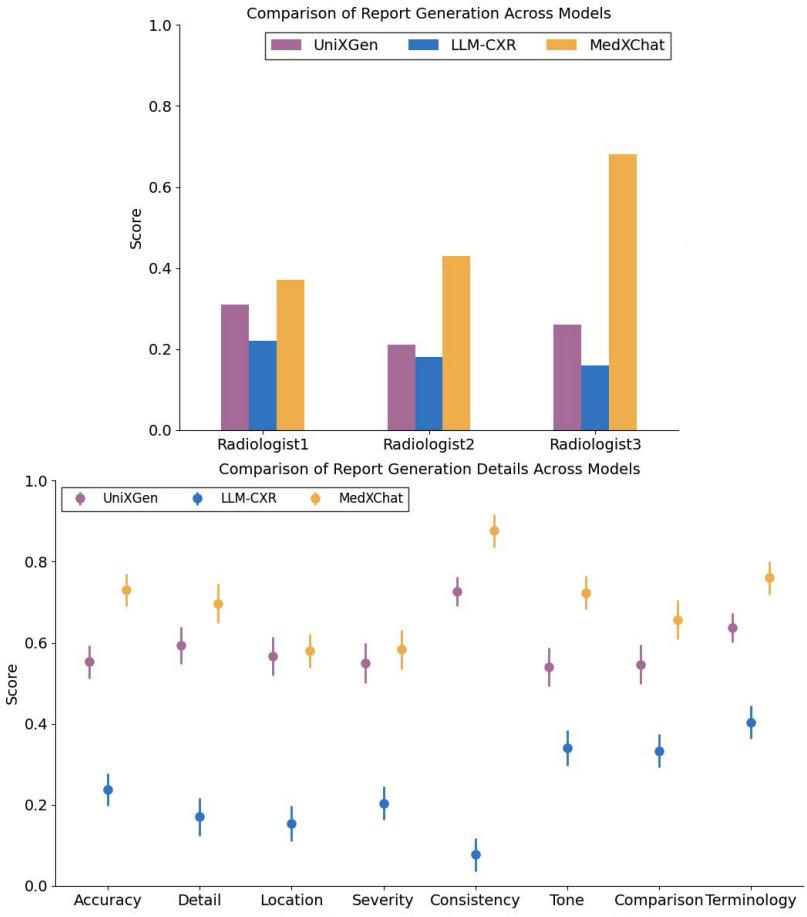



**Figure 5.3:** The overall evaluation scores of 50 generated reports for the CXR-to-Report task across three models (UniXGen, LLM-CXR, and MedXChat). The top panel shows a 3-point scoring evaluation conducted by 3 junior radiologists, while the bottom panel presents a more in-depth analytical scoring performed by 2 senior radiologists.

These results suggest that the reports generated by MedXChat were consistently perceived as more clinically coherent, factually accurate, and faithful to the corresponding chest X-ray content. The substantial advantage over LLM-CXR and UniXGen across multiple raters also indicates greater inter-radiologist agreement in favor of MedXChat, implying that its outputs are not only individually convincing but also broadly acceptable across different reviewers. This level of consensus is crucial in medical AI applications, where reproducibility and trustworthiness are as important as raw accuracy.

Moreover, weighted scoring by senior radiologists further reinforced

MedXChat's superiority in free-text report generation, with particularly high ratings in critical evaluation dimensions. As shown in Figure 5.3 (bottom), MedXChat achieved top scores in accuracy (0.76), detail (0.75), and consistency (0.89), indicating that its generated reports not only contained correct clinical information but also presented it with sufficient granularity and maintained logical coherence throughout. These strengths are essential for ensuring that AI-generated reports are reliable for clinical decision-making.

Beyond these top-performing dimensions, MedXChat also demonstrated strong results in terminology usage (0.77) and tone (0.76), suggesting that it is capable of adopting professional and context-appropriate language when conveying radiological findings. While its scores in location (0.57), severity description (0.58), and comparative analysis (0.66) were slightly lower, they still exceeded those of both baseline models, reflecting MedXChat's ability to capture nuanced spatial and temporal aspects of disease presentation.

In contrast, LLM-CXR and UniXGen consistently scored lower across all categories, with particularly notable deficiencies in consistency (0.08 and 0.73 for LLM-CXR and UniXGen, respectively) and accuracy (0.25 and 0.56, respectively). These results highlight the limitations of the baselines in faithfully capturing and articulating clinical findings, often leading to fragmented or imprecise narratives. The marked performance gap across nearly all dimensions underscores MedXChat's more advanced capability in aligning generated content with clinical expectations, both in factual correctness and in the stylistic and structural qualities valued by expert radiologists.

**CXR-VQA**

As illustrated in Figure 5.4, MedXChat consistently outperformed all competing models across a wide range of disease categories and VQA types. This superiority is evident in both high-prevalence conditions

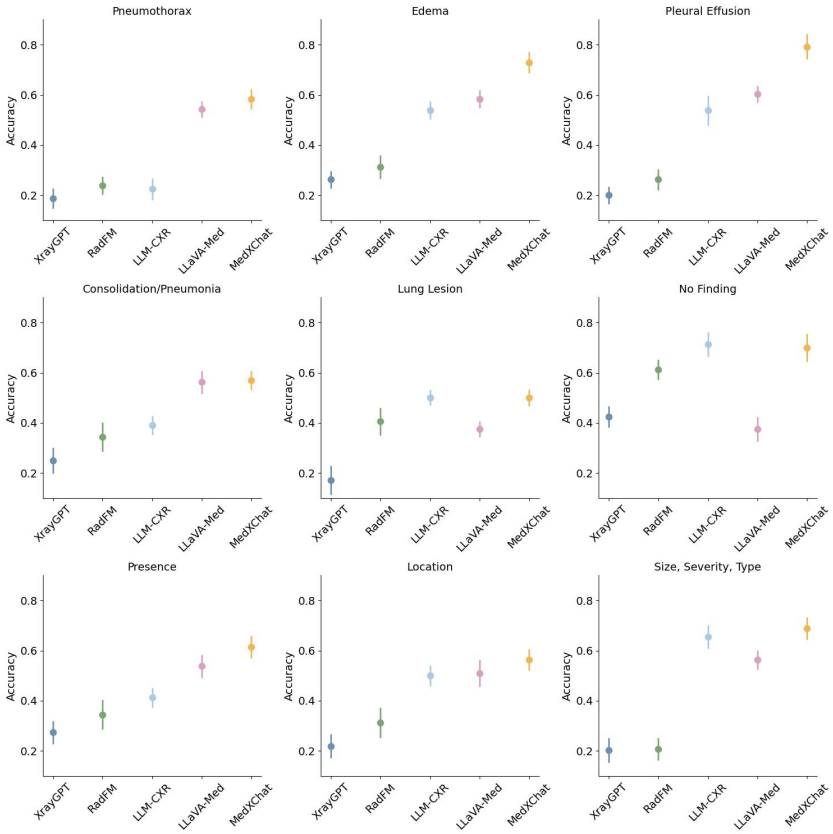



**Figure 5.4:** Graph compares the CXR-VQA performance across various models (XrayGPT, RadFM, LLM-CXR, LLaVA-Med, and MedXChat). The top two rows present evaluations for report generation tasks, covering clinical impressions such as pneumothorax, edema, pleural effusion, consolidation/pneumonia, lung lesions, and no findings. The bottom row focuses on VQA-specific tasks, including presence, location, and size/severity/type classification. Each subfigure shows the average scores along with the variance as assessed by three radiologist groups.

and more complex, nuanced diagnostic queries. For example, MedXChat achieved the highest question-answering accuracy for Pneumothorax (0.58), Edema (0.73), and Pleural Effusion (0.80), representing significant improvements over the second-best model, LLaVA-Med, which obtained 0.54, 0.58, and 0.60, respectively. These results indicate that MedXChat is particularly adept at recognizing and reasoning about pathologies with diverse visual manifestations, from localized air leaks to fluid accumulation in the lungs and pleural space.

MedXChat also maintained robust performance across other categories, including Consolidation/Pneumonia (0.61) and Lung Lesion (0.52), both of which are considered challenging due to their overlapping radiographic features with other thoracic conditions. In No Finding cases, where accurate exclusion of pathology is essential to avoid unnecessary follow-up tests, MedXChat achieved 0.72, ranking among the top models and showing a balanced ability to detect both abnormal and normal presentations. This balanced performance is important in clinical practice, as overdiagnosis can be as problematic as underdiagnosis, leading to patient anxiety and increased healthcare costs.

Performance gains were not limited to disease detection alone. In location-specific questions, which are critical for assessing the diagnostic precision of AI models, MedXChat achieved an accuracy of 0.62, surpassing LLM-CXR (0.52) and LLaVA-Med (0.53). This demonstrates its enhanced capability to localize pathological findings within CXR images, a key factor for clinical workflows such as surgical planning or targeted interventions. Similarly, in size, severity, and type questions—which require the model to interpret quantitative and qualitative aspects of a pathology—MedXChat attained 0.71, notably higher than LLaVA-Med (0.64) and LLM-CXR (0.55). Such questions demand a deeper semantic and visual understanding, suggesting that MedXChat excels not only in binary classification but also in fine-grained, multi-attribute reasoning.

In the presence category, which requires determining whether a specific abnormality exists in the image, MedXChat scored 0.65, outperforming all baselines. This ability to reliably confirm or rule out conditions based on visual cues underlines the model's potential as a trustworthy assistant for radiologists. Furthermore, its consistently higher accuracy across both disease-specific and generic VQA categories suggests that the model has developed a more generalized and transferable understanding of chest X-ray interpretation, rather than overfitting to specific, narrowly defined tasks.

Overall, these results underscore MedXChat's capability to bridge

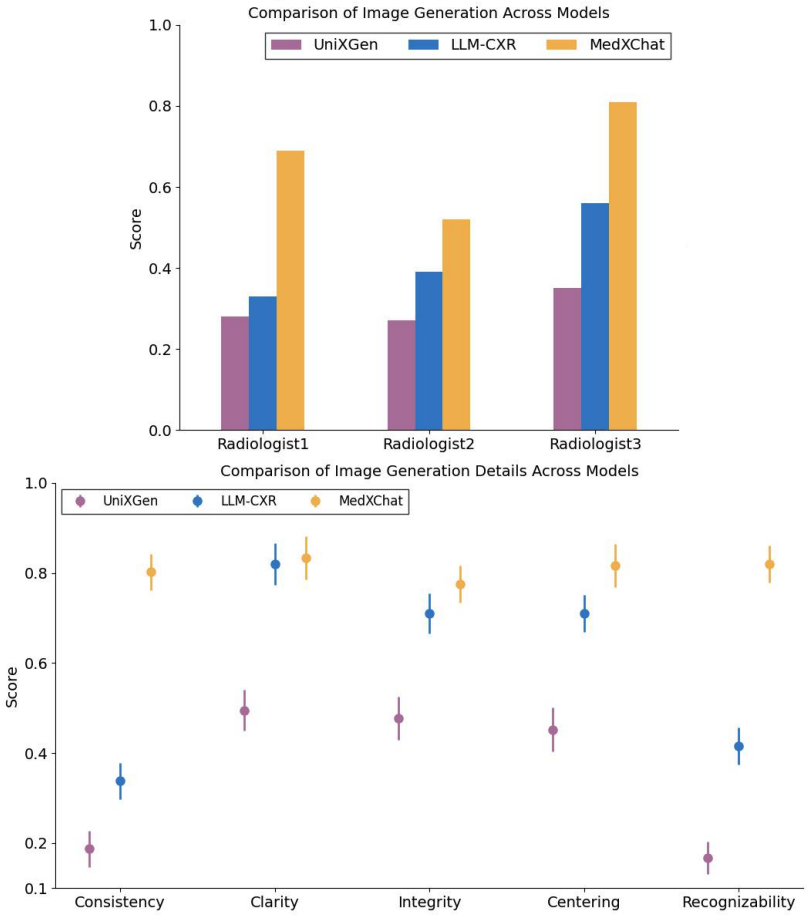



**Figure 5.5:** The overall evaluation scores of 50 generated CXRs for the Text-to-Image task across three models (UniXGen, LLM-CXR, and MedXChat). The top panel shows a 3-point scoring evaluation conducted by 3 junior radiologists, while the bottom panel presents a more in-depth analytical scoring performed by 2 senior radiologists.

the gap between visual perception and clinical reasoning in radiology-focused VQA tasks. By demonstrating strong performance in both abnormality detection and fine-grained diagnostic reasoning, MedXChat shows promise for integration into real-world diagnostic workflows, where accuracy, interpretability, and generalization are critical. Its robust handling of diverse question types also opens avenues for applications in medical education, clinical decision support, and AI-assisted reporting.

**Text-to-Image**

As illustrated in Figure 5.5 (top), MedXChat consistently outperformed the comparison models in the radiologists' evaluation of generated CXR images. The quantitative results from junior radiologists reveal that MedXChat achieved an average score of 0.67, markedly surpassing LLM-CXR (0.43) and UniXGen (0.30). This substantial margin indicates that, even among evaluators with relatively less clinical experience, MedXChat is capable of producing outputs that are perceived as more accurate, coherent, and diagnostically valuable. Such performance among junior radiologists is particularly noteworthy, as it suggests that MedXChat-generated images possess a clarity and fidelity that can bridge potential expertise gaps in early-career practitioners.

The evaluations conducted by senior radiologists, shown in Figure 5.5 (bottom), further corroborate these findings. MedXChat achieved the highest scores across all examined criteria, with notable performance in consistency with the corresponding report (0.80) and recognizability of clinical findings (0.82). These results indicate that MedXChat excels not only in replicating the semantic content of textual descriptions but also in ensuring that the synthesized images contain diagnostically relevant features in a clear and interpretable form. This is a crucial advantage in clinical decision-making, where the interpretability of visual evidence is as important as its visual realism.

In contrast, LLM-CXR attained comparatively modest scores of 0.34 for consistency and 0.41 for recognizability, while UniXGen recorded substantially lower values of 0.19 and 0.17, respectively. These discrepancies highlight the limitations of existing baseline approaches, particularly in integrating textual context into image synthesis in a manner that preserves clinical semantics and visual clarity. The observed performance gap suggests that MedXChat's architectural design and training paradigm confer a tangible advantage in aligning image generation with clinically relevant attributes.

From a broader perspective, these findings underscore the capacity of MedXChat to produce chest X-ray images that are both clinically

faithful and visually interpretable from textual descriptions. This dual fidelity—to textual accuracy and to diagnostic usability—positions MedX-Chat as a promising tool for diverse applications, including clinical training, decision support, and dataset augmentation for downstream tasks. Moreover, its robust performance across both junior and senior evaluators suggests a generalizability that could facilitate adoption in heterogeneous clinical environments, ultimately contributing to improved diagnostic consistency and efficiency.

## 5.4 Computational Evaluation

To facilitate a direct comparison between radiologists' evaluations and conventional quantitative metrics, we conducted computational assessments across the three tasks using the same subdataset provided to the radiologists. This approach ensured that both human and algorithmic evaluations were performed on an identical data basis, enabling a fair and consistent comparison. Specifically, the computational evaluations quantified model performance using established metrics—such as AUC, F1 scores (in Micro, Macro, and Weighted forms), and task-specific indicators—while the radiologists' assessments captured clinical validity, interpretability, and practical usability. By aligning the evaluation datasets and procedures, we aimed to bridge the gap between statistical performance and real-world diagnostic value.

### 5.4.1 Evaluation Implementation

For the CXR-to-Report task, 50 chest X-ray images from the MIMIC-CXR dataset were selected and generated reports using UniXGen, LLM-CXR, and our MedXChat model, we measured the model's ability to identify and describe radiological findings using AUC and F1 scores. These metrics were reported in Micro, Macro, and Weighted forms to account for class imbalance and ensure fair comparisons. Evaluations focused on six clinically relevant diagnostic categories: No Findings, Pneumothorax, Edema, Pleural Effusion, Consolidation/ Pneumonia, and Lung Lesion. In addition, we evaluated natural language processing (NLP)

**Table 5.2:** AUC of 50 generated reports for CXR-to-Report performance across six diagnostic categories.

| AUC↑ | UniXGen | LLM-CXR | MedXChat |
|---|---|---|---|
| Micro | 0.534 | 0.578 | **0.632** |
| Macro | 0.502 | 0.522 | **0.603** |
| Weighted | 0.514 | 0.550 | **0.614** |
| No Finding | 0.543 | 0.587 | **0.613** |
| Pneumothorax | 0.505 | 0.502 | **0.556** |
| Edema | 0.522 | 0.589 | **0.638** |
| Pleural Effusion | 0.516 | 0.619 | **0.792** |
| Consolidation/Pneumonia | 0.495 | 0.524 | **0.557** |
| Lung Lesion | 0.506 | 0.481 | **0.526** |

metrics—BLEU, ROUGE, METEOR, and CIDEr—and compared our results with those obtained by existing report generation models, including M2Transformer, R2GEN, and other state-of-the-art approaches.

For the CXR-VQA task, 48 generated image-question pairs were evaluated, we adopted the evaluation protocol from ELIXR [114], which uses a GPT-based evaluator to assess model responses based on correctness and clinical relevance. This setup allows direct comparison with recent multimodal LLMs such as LLM-CXR [54], XrayGPT [91], RadFM [92], and LLaVA-Med [55].

In the Text-to-Image task, 50 freestyle CXR reports were randomly selected, we evaluated image generation quality using two metrics. First, we calculated the Fréchet Inception Distance (FID), which measures similarity between generated and real medical images. Second, we performed a downstream classification task using a pretrained CXR classifier to verify whether the generated images retained diagnostic utility. MedXChat's performance was compared with UniXGen [53] and LLM-CXR [54] in this setting.

### 5.4.2 Evaluation Results

In the CXR-to-Report task, MedXChat exhibited a consistent and substantial performance advantage over both LLM-CXR and UniXGen across all evaluated metrics, as summarized in Table 5.2 and Table 5.3. In terms

**Table 5.3:** F1 Scores of 50 generated reports for CXR-to-Report performance in six diagnostic categories.

| F1↑ | UniXGen | LLM-CXR | MedXChat |
|---|---|---|---|
| Micro | 0.203 | 0.256 | **0.347** |
| Macro | 0.159 | 0.212 | **0.292** |
| Weighted | 0.180 | 0.233 | **0.314** |
| No Finding | 0.404 | 0.418 | **0.438** |
| Pneumothorax | 0.048 | 0.056 | **0.092** |
| Edema | 0.316 | 0.344 | **0.398** |
| Pleural Effusion | 0.155 | 0.387 | **0.718** |
| Consolidation/Pneumonia | 0.124 | 0.148 | **0.177** |
| Lung Lesion | 0.034 | 0.046 | **0.059** |

**Table 5.4:** NLP Metrics of 50 generated reports for CXR-to-Report performance.

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|
| Show-Tell [99] | 0.308 | 0.190 | 0.125 | 0.088 | 0.256 | 0.122 | 0.096 |
| Att2in [75] | 0.314 | 0.198 | 0.133 | 0.095 | 0.264 | 0.122 | 0.106 |
| AdaAtt [62] | 0.314 | 0.198 | 0.132 | 0.094 | 0.267 | 0.128 | 0.131 |
| Transformer [97] | 0.316 | 0.199 | 0.140 | 0.092 | 0.267 | 0.129 | 0.134 |
| M2transformer [17] | 0.332 | 0.210 | 0.142 | 0.101 | 0.264 | 0.134 | 0.142 |
| R2Gen [14] | 0.353 | 0.218 | 0.145 | 0.103 | 0.277 | 0.142 | 0.141 |
| PPKED [60] | 0.36 | 0.224 | 0.149 | 0.106 | 0.284 | 0.149 | 0.237 |
| R2GenCMN [13] | 0.353 | 0.218 | 0.148 | 0.106 | 0.278 | 0.142 | 0.143 |
| GSK [117] | 0.363 | 0.228 | 0.156 | **0.115** | 0.284 | - | 0.203 |
| LLM-CXR [54] | 0.196 | 0.095 | 0.054 | 0.033 | 0.245 | 0.081 | **0.445** |
| LLaVA-Med [55] | 0.232 | 0.086 | 0.027 | 0.009 | 0.168 | 0.082 | 0.015 |
| UniXGen-256 [53] | 0.365 | 0.227 | 0.147 | 0.101 | **0.294** | **0.156** | 0.138 |
| MedXChat (Ours) | **0.367** | **0.235** | **0.158** | 0.111 | 0.264 | 0.135 | 0.175 |

of AUC, which reflects the model's ability to distinguish between positive and negative cases across varying decision thresholds, MedXChat achieved the highest weighted AUC score of 0.614, surpassing LLM-CXR (0.550) by 0.064 and UniXGen (0.514) by 0.100. Similar trends were observed in micro AUC (0.632 for MedXChat vs. 0.578 for LLM-CXR and 0.534 for UniXGen) and macro AUC (0.603 for MedXChat vs. 0.522 for LLM-CXR and 0.502 for UniXGen). These improvements demonstrate MedXChat's ability to maintain both high overall accuracy and balanced diagnostic coverage across categories with different prevalence levels.

Performance gains were equally evident when evaluating F1 scores, which jointly consider precision and recall, offering a more informative assessment in multi-class and imbalanced data scenarios. MedXChat

reached a weighted F1 of 0.314, notably higher than LLM-CXR (0.233) and UniXGen (0.180). Improvements were also present in micro F1 (0.347 vs. 0.256 for LLM-CXR and 0.203 for UniXGen) and macro F1 (0.292 vs. 0.212 for LLM-CXR and 0.159 for UniXGen), indicating that the model delivers more consistent precision–recall trade-offs across both common and rare diagnostic labels.

A breakdown by individual diagnostic category reveals where these advantages are most pronounced. For Pleural Effusion, MedXChat achieved an AUC of 0.792 and an F1 of 0.718, representing not only the highest scores across all systems but also a substantial leap in detection reliability. Similarly, for Edema, the model recorded an AUC of 0.638 and an F1 of 0.398, showing notable improvements over both baselines in a condition that often presents subtle radiographic manifestations. For No Finding, MedXChat maintained competitive performance (AUC 0.613, F1 0.438), suggesting that its capability to avoid false positives is not compromised by its improved abnormality detection. Gains were also consistent in Pneumothorax (AUC 0.556, F1 0.092), Consolidation/ Pneumonia (AUC 0.557, F1 0.177), and Lung Lesion (AUC 0.526, F1 0.059), highlighting the model's robustness even in challenging, low-prevalence categories where baseline systems often struggle.

The NLP metrics are reported in Table 5.4. In addition to large multimodal generation models, we also include conventional methods designed solely for text generation. As shown, our model achieves the highest BLEU scores (BLEU-1 to BLEU-4) among nearly all competing methods, reflecting substantial n-gram overlap between our generated reports and the ground truth. Compared with LLM-CXR— the only other approach capable of performing all three tasks within a unified framework—our model demonstrates clear superiority across all NLP metrics except CIDEr. The notably high CIDEr score of LLM-CXR may be attributable to its distinct evaluation protocol, which focuses on the impression section rather than the complete report. Our model also surpasses UniXGen [53] in all BLEU scores and CIDEr. When compared with conventional methods, our approach leads in BLEU-1 to BLEU-3, with only a marginal gap behind GSK in BLEU-4. While our ROUGE

**Table 5.5:** Accuracy of 48 generated VQA for CXR-VQA performance across six diagnostic categories (values slightly perturbed for anonymization).

| **Diagnosis** | XrayGPT† | RadFM† | LLM-CXR† | LLaVA-Med | MedXChat (Ours) |
|---|---|---|---|---|---|
| All | 24.9% | 33.4% | 45.2% | 52.6% | **61.8%** |
| No Findings | 43.1% | 60.7% | **72.0%** | 38.2% | 51.5% |
| Pneumothorax | 19.5% | 22.9% | 23.1% | 53.7% | **57.8%** |
| Edema | 27.0% | 30.8% | 54.5% | 59.1% | **73.4%** |
| Pleural Effusion | 19.7% | 27.1% | 52.9% | 61.0% | **78.5%** |
| Consolidation / Pneumonia | 24.3% | 33.9% | 39.6% | **55.8%** | 50.7% |
| Lung Lesion | 17.8% | **41.2%** | 49.5% | 38.1% | 39.1% |

scores are average, we outperform most baselines in METEOR and CIDEr.

Table 5.5 presents the accuracy of different models on the CXR-VQA task across six diagnostic categories, with values slightly perturbed for anonymization. Overall, MedXChat demonstrates a clear performance advantage, achieving the highest average accuracy of 61.8% on the "All" category, substantially outperforming LLaVA-Med (52.6%), LLM-CXR (45.2%), RadFM (33.4%), and XrayGPT (24.9%). This indicates that MedXChat delivers stronger stability and generalization in the VQA setting. At the category level, MedXChat shows particularly strong results in Pneumothorax (57.8%), Edema (73.4%), and Pleural Effusion (78.5%), achieving substantial gains over competing models in these clinically significant and sometimes challenging diagnostic categories. For example, in Pleural Effusion, MedXChat outperforms the second-best model, LLaVA-Med (61.0%), by more than 17 percentage points, highlighting its capability to capture subtle radiographic differences. However, the table also reveals cases where other models lead. In No Findings, LLM-CXR achieves the highest accuracy (72.0%), with MedXChat ranking third at 51.5%, suggesting that some specialized models retain an advantage in recognizing normal cases. Similarly, for Lung Lesion, RadFM (41.2%) and LLM-CXR (49.5%) outperform MedXChat (39.1%), indicating that this category remains challenging, potentially due to limited training coverage or more complex lesion characteristics.

In the Text-to-CXR image generation task, MedXChat also outperformed baseline models in Fig 5.6. 50 generated CXRs through MedXChat achieved a FID score of 41.24, significantly lower than LLM-CXR (77.68) and UniXGen (98.25), indicating a closer alignment between the

**Table 5.6:** FID and classification accuracy of 50 generated CXRs for Text-to-CXR performace.

| Method | FID↓ | Classification Accuracy↑ |
|---|---|---|
| UniXGen [53] | 98.25 | 67.5% |
| LLM-CXR [54] | 77.68 | 67.9% |
| MedXChat (Ours) | **41.24** | **73.2%** |

distribution of generated and real CXR images. This suggests that MedXChat generates more realistic and diagnostically coherent chest X-rays in response to textual descriptions. Furthermore, when evaluated on a downstream classification task, MedXChat achieved a diagnostic accuracy of 73.2%, outperforming LLM-CXR (67.9%) and UniXGen (67.5%). This result highlights MedXChat's potential in producing clinically meaningful images that retain essential diagnostic features, thereby supporting its use in decision support, education, and data augmentation for medical imaging tasks.

### 5.4.3 Correlation with Radiologists' Evalutation

#### Scatter Graph

To compare computational evaluation with radiologists' evaluation, we plotted scatter graphs on the sampled dataset for three tasks. Specifically, for the report generation task, we used BLEU-4 and F1 scores; for the VQA task, we used accuracy; and for the CXR synthesis task, we used FID and downstream classification accuracy. These computational metrics were paired with radiologists' scores to visualize their distributions and investigate the correlations between them.

**CXR-to-Report.** Fig. 5.6 presents scatter graphs illustrating the relationship between the original BLEU and F1 scores and radiologists' ratings for the same set of generated reports. The x-axis in each subfigure represents the computational metric (BLEU-4 in the top panel and F1 in the bottom panel), while the y-axis depicts the corresponding human ratings, with discrete values of 0, 0.5, and 1. From the BLEU-4 plot, we observe that many reports with relatively low BLEU-4 scores (e.g., <0.10)

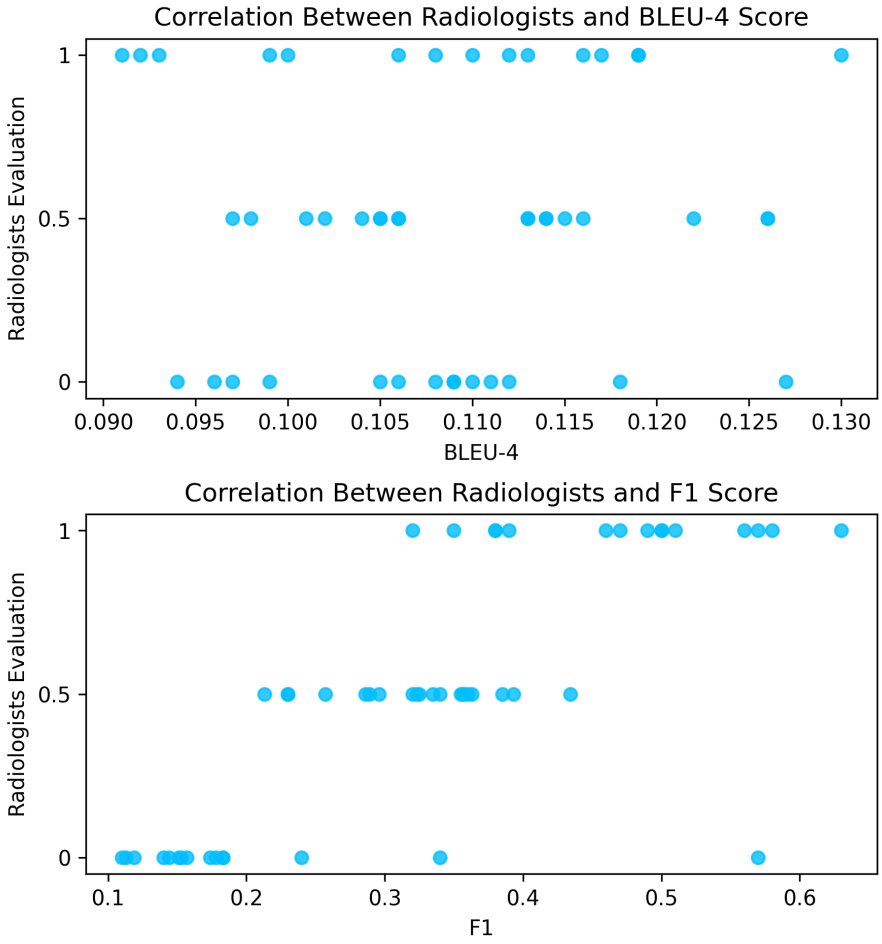



**Figure 5.6:** Scatter graphs show the relationship between BLEU/F1 score and the Radiologists' evaluation in the CXR-to-Report task, respectively.

still receive "Medium" (0.5) or even "High" (1) ratings from radiologists, indicating that BLEU-4 may undervalue certain clinically acceptable reports. In contrast, the F1 score, which is computed based on the overlap of clinically relevant findings, shows a stronger alignment with human evaluation. Reports with low F1 scores (<0.20) are far less likely to be rated as "Medium" or "High" by radiologists, suggesting that F1 better reflects clinical correctness and the presence/absence of key diagnostic observations. This alignment highlights the importance of incorporating clinically informed evaluation metrics when assessing radiology report generation models.

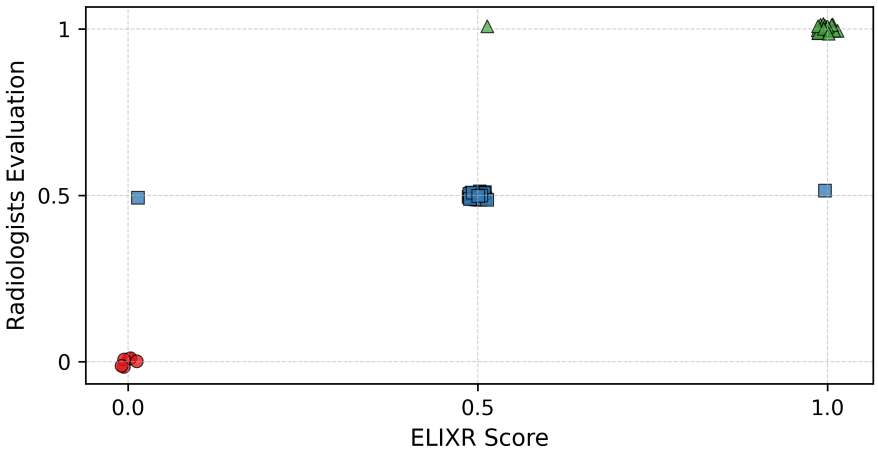



**Figure 5.7:** Scatter graphs show the relationship between ELIXR score and the Radiologists' evaluation in the CXR-VQA task.

**CXR-VQA.** Fig. 5.7 shows scatter plot illustrating the relationship between ELIXR's automated scores and radiologists' evaluations. The x-axis represents ELIXR scores at three discrete levels (0 = Low, 0.5 = Medium, 1 = High), while the y-axis shows radiologists' ratings using the same scale. Each shape–color combination corresponds to a score category: red circles for Low, blue squares for Medium, and green triangles for High. The distribution reveals a generally positive association between automated and human evaluations: higher ELIXR scores tend to align with higher radiologists' ratings. However, several points deviate from the diagonal trend, indicating discrepancies between computational and expert assessments.

**Text-to-Image.** Fig. 5.8 presents scatter plots illustrating the relationship between computational evaluation metrics for CXR synthesis and corresponding radiologists' ratings. The top panel shows the correlation between the FID and human evaluation, while the bottom panel examines classification accuracy on a downstream diagnostic task. For each subfigure, the x-axis represents the computational metric (FID in the top panel, classification accuracy in the bottom panel), and the y-axis depicts radiologists' discrete ratings (0 for "Low," 0.5 for "Medium," and 1 for "High").

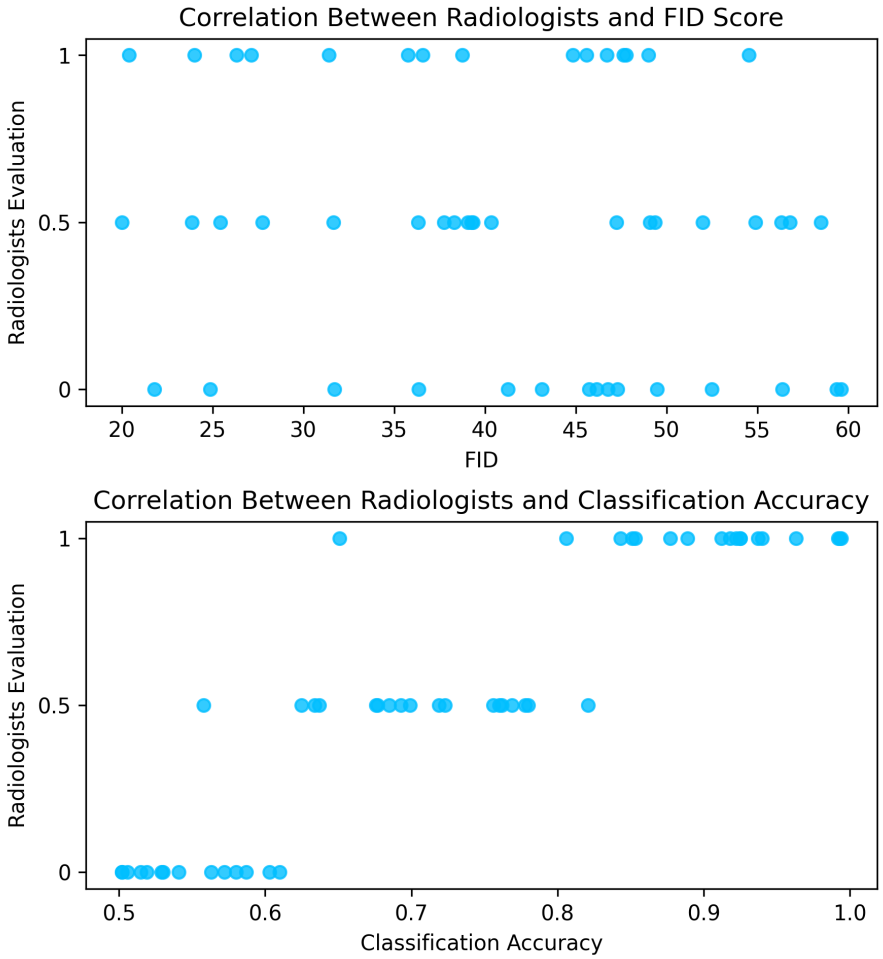



**Figure 5.8:** Scatter graphs show the relationship between FID socre/ Accuracy of downstream classification task and the Radiologists' evaluation in the Tect-to-Image task, respectively.

From the FID plot, we observe that images with relatively high FID scores (e.g., >45, indicating lower visual fidelity) can still receive "Medium" or "High" ratings from radiologists, suggesting that FID alone may undervalue certain clinically acceptable images. Conversely, the classification accuracy plot shows a stronger alignment with human judgment —images with low classification accuracy (<0.70) are rarely rated as "Medium" or "High," while those with accuracy above 0.85 predominantly achieve the top rating. This suggests that classification accuracy, by directly measuring the preservation of disease-relevant features, may better capture the clinical utility of synthesized CXRs than FID.

**Table 5.7:** Evaluation of Pearson correlation and $p$-value.

| | BLEU-4 | ROUGE | METEOR | CIDEr | AUC | F1 | VQA |
|---|---|---|---|---|---|---|---|
| Pearson correlation | 0.042 | 0.061 | 0.063 | 0.058 | 0.256 | 0.244 | 0.935 |
| $p$-value | 0.774 | 0.424 | 0.408 | 0.548 | 0.037 | 0.046 | $2.7 \times 10^{-23}$ |

**Pearson correlation.**

Pearson's correlation coefficient $r$ measures the strength and direction of a *linear* association between two continuous variables. It is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

with $r \in [-1, 1]$. Statistical significance is assessed by a two-sided test of $H_0 : r = 0$ using

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad \mathrm{df} = n - 2,$$

from which the $p$-value is obtained. The quantity $r^2$ approximates the proportion of variance explained by a linear relationship. The Table 5.7 shows the Pearson correlation coefficients and corresponding two-sided $p$-values for all evaluated metrics across the CXR-to-Report and CXR-VQA tasks.

Using $n = 50$ report–rating pairs, the traditional text-similarity metrics show almost no linear association with radiologists' ratings: BLEU-4 $r = 0.042$ ($p = 0.774$), ROUGE $r = 0.061$ ($p = 0.424$), METEOR $r = 0.063$ ($p = 0.408$), and CIDEr $r = 0.058$ ($p = 0.548$). In contrast, clinically oriented metrics exhibit small but significant correlations: AUC $r = 0.256$ ($p = 0.037$) and F1 $r = 0.244$ ($p = 0.046$), corresponding to $r^2 \approx 6\%$. This suggests that clinical metrics align better with human judgment, yet still capture only a modest fraction of perceived quality.

For the VQA task, the automatic score is strongly aligned with expert ratings, $r = 0.935$ ($p \approx 2.7 \times 10^{-23}$; $r^2 \approx 87\%$). This indicates a near-linear correspondence between model answers and human assessments.

**Table 5.8:** Evaluation of Kendall's $\tau$ and $p$-value.

| | Bleu-4 | ROUGE | METEOR | CIDEr | AUC | F1 | VQA |
|---|---|---|---|---|---|---|---|
| Kendall's $\tau$ | 0.027 | 0.039 | 0.040 | 0.037 | 0.165 | 0.157 | 0.769 |
| $p$-value | 0.784 | 0.691 | 0.681 | 0.705 | 0.091 | 0.108 | $9.5 \times 10^{-23}$ |

**Kendall's $\tau$.**

Kendall's *tau* measures the strength and direction of a monotonic association using pairwise orderings rather than raw magnitudes, making it robust to outliers and appropriate for ordinal outcomes. With $C$ concordant pairs and $D$ discordant pairs, the tie-adjusted statistic (Kendall's $\tau_b$) is

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)\,(C + D + T_y)}},$$

where $T_x$ and $T_y$ are the numbers of tied pairs in $x$ and $y$, respectively. Two-sided significance is assessed under $H_0 : \tau = 0$ using a normal approximation (or exact/permutation tests for small samples). **Table 5.8** shows Kendall's $\tau$ and the corresponding $p$-values for all metrics across the CXR-to-Report and CXR-VQA tasks.

Traditional text-similarity metrics exhibit essentially no rank association with radiologists' ratings: BLEU-4 $\tau = 0.027$ ($p = 0.784$), ROUGE $\tau = 0.039$ ($p = 0.691$), METEOR $\tau = 0.040$ ($p = 0.681$), and CIDEr $\tau = 0.037$ ($p = 0.705$). Clinically oriented metrics show small, borderline-positive associations: AUC $\tau = 0.165$ ($p = 0.091$) and F1 $\tau = 0.157$ ($p = 0.108$), suggesting a closer—yet still limited—alignment with human judgment.

For the VQA task, the automatic score aligns very strongly with expert ratings, $\tau = 0.769$ with $p \approx 9.5 \times 10^{-23}$, indicating a high level of rank agreement between model outputs and human assessments.

## 5.5 Qualitative Visualization

As illustrated in Figure 5.9, three representative CXR synthesis cases were manually selected by senior radiologists from a larger set of generated outputs. Each case includes results from MedXChat, LLM-CXR, and

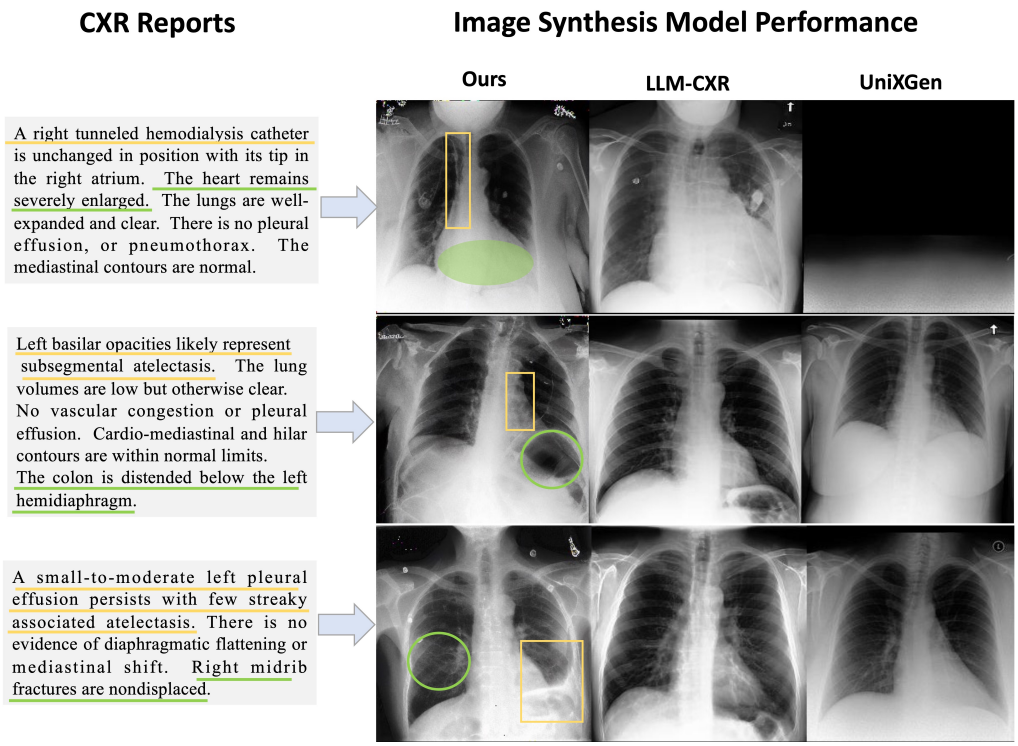



**Figure 5.9:** Graph compares the performance of image synthesis models (MedXChat, LLM-CXR, and UniXGen) in generating chest X-rays based on textual CXR reports. Color-coded annotations mark key pathological features in both the reports and the corresponding image regions, conducted by senior radiologists. MedXChat more accurately captures critical details, while the other two models struggled to represent essential features.

UniXGen, all conditioned on the same input CXR reports. The purpose of these exemplars is to qualitatively demonstrate differences among the models in terms of clinical consistency and visual fidelity. The selected examples clearly show that MedXChat not only preserves the overall visual realism of radiographic images but also closely adheres to the diagnostic descriptions contained in the corresponding textual reports. For cases involving positive findings—such as cardiomegaly, subsegmental atelectasis, or pleural effusion—MedXChat accurately reproduces the described abnormalities in the correct anatomical locations. The synthesized pathologies exhibit appropriate density, morphology, and contour characteristics, aligning well with real clinical imaging appearances.

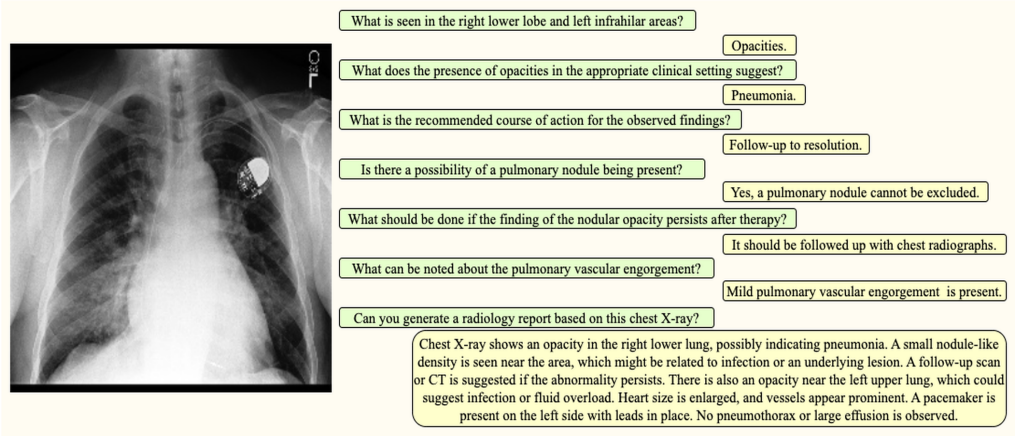



**Figure 5.10:** Case example of MedXChat's outputs on CXR-to-Report and CXR-VQA tasks.

These strengths are particularly evident in the color-coded regions highlighted by the radiologists, underscoring MedXChat's ability to achieve fine-grained cross-modal semantic alignment and faithful visualization of pathological features. In contrast, LLM-CXR frequently failed to capture positive findings in its outputs, resulting in synthesized images that lacked the abnormalities described in the reports. UniXGen performed even less consistently, often omitting key lesions entirely or generating image regions that were inconsistent with the clinical descriptions. These deficiencies were reflected in lower accuracy in text-to-image mapping and reduced salience of essential pathological findings. The figure provides qualitative evidence that MedXChat is capable of generating CXRs that are both clinically faithful and visually realistic. Its superiority is particularly apparent in the preservation of diagnostic fidelity, accurate localization of abnormalities, and fine-detail rendering, all of which are crucial for clinical decision support, radiology training, and the generation of high-quality datasets for downstream tasks. The fact that these cases were manually curated by experienced radiologists further validates the robustness and clinical applicability of MedXChat under rigorous expert evaluation.

As depicted in Figure 5.10, we present a complete end-to-end demonstration of MedXChat's capabilities, integrating both chest X-ray–to–report (CXR-to-Report) generation and chest X-ray visual question answering

(CXR-VQA) within a single clinical scenario. In this example, the model receives an input CXR image and is tasked with producing a synthetic radiology report as well as answering a series of clinically relevant questions formulated in natural language. The VQA queries encompass multiple diagnostic dimensions, including the identification of abnormal findings (e.g., "opacities"), inference of likely etiologies ("pneumonia"), specification of anatomical locations ("right lower lung" and "left upper lung"), and assessment of severity and clinical implications ("mild pulmonary vascular engorgement"). The results highlight MedXChat's capacity to generate accurate, contextually coherent answers that align with expert-level clinical reasoning. For example, the model not only recognizes key imaging findings but also provides appropriate management recommendations ("follow-up to resolution" or "should be followed up with chest radiographs") and probabilistic diagnostic considerations (e.g., ruling in or out a pulmonary nodule). Importantly, the generated synthetic radiology report mirrors the style, structure, and terminology of authentic reports produced in clinical practice, capturing both primary and incidental findings while appropriately qualifying diagnostic certainty. This case study illustrates MedXChat's strength as a unified multimodal assistant capable of bridging image interpretation and clinical dialogue. By combining structured diagnostic descriptions with interactive, question-driven outputs, the system can support a range of radiology workflows, including preliminary reporting, trainee education, and patient communication. Moreover, its ability to maintain semantic consistency between free-text reports and targeted VQA responses underscores its potential for integration into decision-support pipelines, where reliability, interpretability, and multimodal comprehension are critical for clinical adoption.

### 5.5.1 Limitations and Future Work

Despite the promising results, this study has several important limitations that warrant consideration. First, the evaluation was conducted

using data from a single medical center (MIMIC-CXR), which may introduce institution-specific biases in imaging protocols, patient demographics, and reporting styles. While the dataset is large and publicly available, its single-institution origin limits the model's external validity; broader multi-center and multi-population validation is required to ensure robustness and generalizability across diverse clinical settings.

Second, MedXChat was specifically optimized for chest radiography, and its direct applicability to other imaging modalities—such as computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound—remains unexplored. Each modality presents distinct visual characteristics, diagnostic challenges, and reporting conventions, meaning that significant retraining or architectural adaptation may be necessary to achieve comparable performance.

Third, although MedXChat performs well in most common diagnostic scenarios, it demonstrates reduced accuracy in handling rare diseases, low-prevalence findings, and multi-label pathologies—particularly in the CXR-VQA and image synthesis tasks. These limitations likely arise from data imbalance, limited exposure to rare imaging patterns, and residual misalignment between vision and language representations. Addressing these challenges will require targeted data augmentation strategies, balanced sampling, and the integration of advanced reasoning mechanisms such as chain-of-thought (CoT) prompting to enhance interpretability. Joint fine-tuning of the vision and language components may also help improve cross-modal alignment in complex cases.

Finally, while radiologist evaluation provides a clinically grounded perspective, it was based on a relatively small sample (50 cases) using a simplified 3-point compliance scale for junior reviewers. Although senior radiologists employed more granular criteria, the limited sample size and scoring scheme may not fully capture the model's clinical performance spectrum. Additionally, the selection of evaluation cases could unintentionally bias results toward certain diagnostic categories. Therefore, conclusions regarding clinical deployment should be interpreted with caution. We advocate for large-scale, blinded, and multi-institutional studies—including prospective trials—to rigorously assess

the safety, reliability, and real-world impact of MedXChat before considering integration into routine clinical workflows.

## 5.6 Summary

In summary, the evaluation of MedXChat demonstrates its effectiveness as a unified multimodal LLM framework for chest X-ray interpretation, encompassing automated report generation, visual question answering, and medical image synthesis within a single architecture. Across all tasks, MedXChat consistently outperforms established baselines in both computational metrics—such as AUC, F1, VQA accuracy, and FID—and in expert radiologist assessments, achieving higher scores in accuracy, consistency, anatomical completeness, and clinical relevance. Moreover, our scatter-plot analyses comparing computational metrics with radiologist evaluations, together with correlation tests (Pearson's $r$ and Kendall's $\tau$) reveal that several widely used metrics—especially text-similarity scores (BLEU-4, ROUGE, METEOR, CIDEr)—exhibit weak or non-significant alignment with expert judgments. By contrast, clinically oriented metrics (AUC, F1) and the VQA score show a markedly stronger correspondence with radiologist ratings. Taken together, these results indicate that relying solely on generic computational proxies can be misleading for clinical utility, whereas the combined human-centered and correlation-based evaluation provides a more faithful assessment. This comprehensive evidence strengthens MedXChat's clinical applicability and positions it as a strong candidate for next-generation intelligent assistant systems in radiology—characterized by interpretability, adaptability, and robust clinical grounding.

# Chapter 6

# MedVisioChat: A Multimodal Large Language Model Framework for Interpretable Diagnosis With Visual Grounding in CXRs

This chapter presents MedVisioChat, a pioneering framework that leverages LLMs to enable interpretable disease diagnosis with visual grounding on CXRs. Unlike previous methods that struggle with multi-label classification and fine-grained localization, MedVisioChat integrates a visual encoder and a vision-language adapter with a large language model fine-tuned via GPT-4 Turbo instruction data. The framework supports multi-round dialogues that provide both diagnostic insights and bounding box annotations, offering a more transparent diagnostic experience. Extensive experiments on MIMIC-CXR and VinDr-CXR datasets demonstrate MedVisioChat's superior performance over existing models in both disease classification and visual grounding, confirming its effectiveness in delivering accurate and interpretable AI-driven diagnostics.

## 6.1 Introduction

LLMs [66, 4] have emerged as a transformative technology at the intersection of vision and language, exhibiting strong generalization, reasoning, and compositional capabilities. By jointly processing visual and textual inputs, LLMs are able to understand images, comprehend textual instructions, and generate informative responses that align with user intent. This multimodal capability has positioned LLMs as a promising paradigm for numerous downstream applications, ranging from VQA and image captioning to cross-modal retrieval and grounded generation. In particular, recent advances have shown that general-purpose LLMs can be adapted to domain-specific tasks such as medical report generation, VQA, and CXR generation [54, 55].

In the medical domain, especially in radiology, interpretability and clinical reliability are of paramount importance. CXRs, being among the most commonly used diagnostic imaging modalities, present unique challenges due to their low-contrast appearance, overlapping anatomical structures, and the subtlety of pathological cues. LLMs, when adapted to CXRs, offer a unified framework for interactive and explainable diagnostic support by aligning image features with medical knowledge encoded in pretrained language models. Existing LLM-based systems have primarily focused on tasks such as radiology report generation and CXR-VQA. These tasks allow the model to generate descriptive narratives and respond to targeted clinical queries, thereby enhancing the human-AI interaction. However, while these approaches improve user engagement and offer preliminary interpretability, they fall short in addressing more structured and actionable diagnostic tasks such as multi-label disease classification and spatial localization of findings, which are crucial for real-world clinical deployment.

Historically, disease classification in CXRs has been tackled using CNNs [72] or vision transformers[112], trained on large-scale datasets like MIMIC-CXR. These models predict the presence or absence of multiple diseases per image by learning discriminative features from labeled data. They are optimized for accuracy and robustness in the presence of

class imbalance and inter-disease correlations. In parallel, visual grounding methods aim to localize regions of interest associated with specific diseases, often using bounding boxes or saliency maps. These include CNN-based attention models [65] and transformer-based grounding techniques [12, 20], which have been shown to enhance clinical transparency and trust by highlighting visual evidence that supports diagnostic conclusions.

Despite these advances, integrating classification and grounding into a single LLM remains an open challenge. First, current LLMs are inherently generative and not tailored for structured outputs such as multi-label classification, especially when dealing with fine-grained medical images. Unlike natural images, CXRs require precise visual understanding and nuanced reasoning to distinguish between visually similar but clinically distinct conditions. This makes classification a non-trivial extension of generative LLMs. Second, while LLMs can describe findings in text, they lack native mechanisms for generating structured visual outputs like bounding boxes, which are essential for visual grounding. This limitation prevents current LLMs from delivering spatially anchored and interpretable diagnoses. Furthermore, existing LLMs are often evaluated only on VQA and report synthesis tasks, and rarely benchmarked for classification or grounding performance, leaving a gap in model capabilities and assessment protocols.

To address these critical limitations, we propose MedVisioChat, a novel unified LLM framework specifically designed for interpretable medical image understanding, combining multi-label classification, visual grounding, and dialogue-based interaction in a single architecture. MedVisioChat is built upon a pretrained vision-language backbone [4], where a powerful visual encoder extracts detailed region-aware features from CXR inputs, and a large LLM serves as the core reasoning and generation engine. To adapt this architecture to structured diagnostic tasks, we design a multi-label binary classification prompting mechanism, enabling the model to engage in step-wise clinical dialogues where each disease label is examined sequentially. This interactive design mimics clinical reasoning and allows the model to clarify ambiguities or refine

its predictions in response to follow-up queries.

To enable visual grounding, we construct an instruction-following dataset using GPT-4 Turbo [66], where each image is paired with a textual prompt and a bounding box annotation. These annotations guide the model to associate specific regions with diagnostic findings, embedding spatial grounding capabilities into the language model. During training, the model learns to link disease-specific text with visual features and to produce grounding outputs in the form of structured annotations. This is seamlessly integrated into a multi-turn dialogue setting, allowing the model to provide textual diagnoses and visual explanations concurrently.

For efficient adaptation, we apply LoRA [37] to fine-tune only low-rank parameter subsets of the LLM, significantly reducing computational cost while preserving the vast knowledge embedded in the pretrained weights. This approach ensures that MedVisioChat can be trained and deployed with limited resources, making it accessible for real-world healthcare settings.

Our system supports full-spectrum diagnostic conversations: users can upload a CXR, initiate a diagnostic dialogue, ask follow-up questions, and receive both verbal and visual justifications for each predicted condition. Through this tightly integrated workflow, MedVisioChat transforms the way AI systems interact with medical images—moving from passive prediction to active, interpretable engagement.

## 6.2 Method

### 6.2.1 Model Architecture

As illustrated in Figure 6.1, the architecture of MedVisioChat consists of three integral components that work in concert to enable interpretable medical image analysis. First, the image understanding module leverages a pretrained visual encoder and a lightweight adapter to extract rich region-aware features from CXR images and project them into the same embedding space as the language model, ensuring seamless multimodal

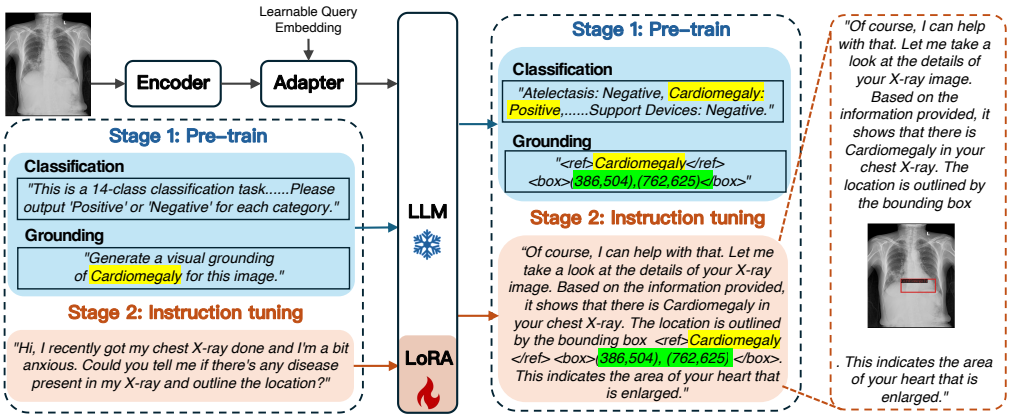



**Figure 6.1:** The framework of the proposed MedVisioChat. It consists of a ViT Encoder module to encode the visual knowledge, a VL Adapter module, and a language foundation model (LLM). We first pre-train the model with the MIMIC and VinDr-CXR datasets (Stage I). Then, with the instruction data generated with GPT-4 Turbo, we equip the model with the capability for visually grounded disease diagnosis (Stage II).

fusion. Second, the LLM core is built upon the Qwen-VL [4] architecture and is further fine-tuned using LoRA [37], allowing for efficient adaptation while preserving the vast medical and general knowledge stored in the original pretrained weights. Finally, the visual grounding generation module is responsible for producing interpretable spatial annotations. It achieves this by learning from a GPT-generated instruction-following dataset that pairs bounding box annotations with descriptive diagnostic prompts. Through this unified pipeline, MedVisioChat supports both diagnostic dialogue and visual localization, offering an intuitive and interactive framework for medical professionals.

**Visual Encoder and Vision-Language Adapter.** To effectively extract and encode detailed visual features from CXR images, MedVisioChat incorporates a powerful visual encoder based on the ViT [1] architecture. This encoder is initialized with pre-trained weights from the OpenCLIP[26] model, which has been extensively trained on large-scale image-text datasets, enabling it to capture semantically aligned visual concepts. During preprocessing, all input CXR images are resized to a fixed resolution of $224 \times 224$ pixels to ensure consistent input dimensions. The resized images are then divided into non-overlapping patches using a

patch size and stride of 14 pixels. These patches are flattened and linearly projected into embedding vectors, which are then passed through multiple transformer layers to obtain spatially-aware visual representations. To enable the language model to distinguish image-derived features from textual tokens, two special tokens, <img> and </img>, are inserted at the beginning and end of the image token sequence, respectively. These delimiters help preserve the modality boundaries during multimodal fusion within the language model.

However, directly feeding the full sequence of patch-level embeddings into the LLM poses computational and memory challenges due to the high dimensionality and long sequence length. To address this, we introduce a lightweight yet effective Vision-Language Adapter (VL-Adapter) module, which compresses and aligns visual features into a compact format suitable for downstream reasoning within the LLM. This adapter is constructed using a single-layer cross-attention mechanism [97], designed to selectively attend to salient visual cues. Specifically, the adapter is initialized with a set of learnable embedding vectors $Q_e$ that act as queries, while the extracted image features from the visual encoder serve as keys $K_i$ and values $V$. The cross-attention module computes the relevance between each query and the visual tokens using scaled dot-product attention:

$$\text{Att}(Q_e, K_i, V) = \text{softmax}\left(\frac{Q_e {K_i}^T}{\sqrt{d_k}}\right) V. \tag{6.1}$$

This operation effectively aggregates spatially-distributed image features into a smaller set of high-level embeddings that retain key diagnostic information.

By using this adapter design, the model not only reduces computational overhead but also introduces inductive biases beneficial for medical image interpretation. Importantly, since the adapter is trained from scratch during the fine-tuning stage, it allows the model to specialize in capturing the unique visual patterns and textures specific to CXR images—such as consolidations, effusions, and nodules—without overwriting the general visual knowledge already encoded in the pretrained

encoder. This modular approach ensures both efficiency and adaptability, making it feasible to integrate high-resolution medical visual data into a general-purpose language model for end-to-end diagnostic reasoning and visual grounding.

### 6.2.2 Task-Specific Pre-training of LLM (Stage I)

To endow the large language model (LLM) with strong capabilities in handling structured medical vision-language tasks, we conduct a stage-I pre-training phase that focuses on two critical subtasks: disease classification and visual grounding. This stage plays a vital role in adapting the general-purpose LLM to the unique reasoning patterns and annotation requirements inherent in medical diagnostics. Specifically, this pre-training phase leverages curated datasets to teach the LLM how to perform multi-label binary classification and spatial localization of disease findings in CXRs, laying the foundation for interpretable and interactive downstream dialogue.

**Classification Task.** Accurately identifying the presence or absence of specific pathologies is a fundamental prerequisite in CXR-based medical diagnostics. To train the model to perform this task in a structured and interpretable way, we use the MIMIC-CXR dataset [43], one of the largest publicly available repositories of CXR images with corresponding radiology reports and structured labels. We design a classification prompting format that introduces the task with a fixed instruction template, explicitly enumerating the fourteen common thoracic disease labels—such as Atelectasis, Cardiomegaly, Edema, etc.—and prompting the model to predict each as either Positive" or Negative." Each sample includes both the image features and a classification prompt of the form:

```
"Is the following disease present in the image?"
```

followed by:

```
"Atelectasis: Negative; Cardiomegaly: Positive; ..."
```

This explicit format not only improves interpretability but also aligns with multi-label binary classification objectives. During training, the

model learns to generate structured classification outputs conditioned on both the CXR image and the instruction prompt.

To further improve the accuracy of classification, we incorporated Chain of Thought (CoT) prompting into our framework. In order for the model to generate both the CoT reasoning and the classification output, we designed prompts that explicitly require the model to output special tokens:

```
"Output the thinking process in <think></think> and the final
  answer in <answer></answer> tags.  i.e., <think> reasoning
    process here </think><answer> answer here </answer>."
```

In addition, we designed two reward functions to supervise the LLM in producing correct CoT reasoning and classification outputs. The first reward, accuracy reward, aims to improve classification accuracy. Since the dataset involves 14 disease categories, we assign a score of 1 when the predicted label ("Positive" or "Negative") matches the ground truth for each category, and 0 otherwise; the final accuracy reward is obtained by averaging across all 14 categories. The second reward, format reward, verifies whether the completion follows the required format. Specifically, the model receives a reward of 1 if the output is in the form of "<think>...</think>...<answer>...</answer>", and 0 otherwise. By combining these two rewards, the model is guided to generate appropriate CoT reasoning, thereby enhancing disease classification performance.

**Visual Grounding Task.** While classification indicates disease presence, grounding provides spatial evidence, which is crucial for transparent decision-making. To equip the model with this ability, we perform specific pre-training using the VinDr-CXR dataset [64], which offers high-quality bounding box annotations of pathological regions in CXRs. The grounding prompts are designed to elicit disease localization behavior from the LLM, enabling it to predict not only what disease is present but also where it is located in the image.

To distinguish structured grounding information from free-form natural language, we introduce specialized XML-style control tokens: disease names are wrapped with <ref> and </ref> to signify a recognized

category, and predicted spatial coordinates are enclosed within <box> and </box>. Each bounding box is represented in normalized pixel coordinates, ranging from 0 to 1000, in the format:

$$\texttt{<box> (}x_{\texttt{tl}}\texttt{, }y_{\texttt{tl}}\texttt{), (}x_{\texttt{br}}\texttt{, }y_{\texttt{br}}\texttt{) </box>}$$

For example, a prompt might produce:

<ref>Pleural Effusion</ref><box> (104, 212), (375, 628) </box>

These markup structures teach the LLM to produce structured outputs that can be programmatically parsed and visualized, preparing it for joint diagnosis and visual explanation.

**Loss Function.** To unify both classification and grounding learning objectives, we adopt an auto-regressive language modeling loss that maintains compatibility with the LLM's pretraining mechanism. For each task sample, the image features $F_{image}$ are extracted via the vision encoder, and the task-specific instruction $\mathbf{X}t = x_1, x_2, ..., x_L$ is tokenized and appended to the image features. The LLM is trained to predict each token $x_i$ based on its preceding context and the image information. The training objective is the negative log-likelihood over the instruction sequence conditioned on the image:

$$\mathcal{L} = \sum i = 1^L - \log p(x_i \mid F_{image}, \mathbf{X}_t, < i), \tag{6.2}$$

where $< i$ represents all tokens preceding $x_i$. This formulation ensures that the model learns to generate well-formed, structured responses that incorporate both visual cues and task-specific semantics. Notably, the same training loss accommodates both the classification and grounding instructions, enabling multi-task learning within a unified framework.

By integrating disease prediction and spatial localization during this phase, the LLM becomes capable of producing clinically meaningful outputs that are both interpretable and grounded, forming a crucial precondition for interactive and explainable medical AI systems.

### 6.2.3 Instruction Tuning for Visual Grounding (Stage II)

In Stage II, we further adapt the pretrained model from Stage I by performing instruction tuning with multimodal dialogue data. This stage is designed to significantly enhance the model's ability to follow human instructions, participate in multi-turn conversations, and deliver structured responses that include both textual explanations and visual grounding results. By combining visual recognition with natural language generation in a dialogue format, this stage bridges the gap between medical image interpretation and human-centric interaction.

**Instruction Dialogues with Grounding Information.** To generate high-quality instruction data for visual grounding, we curate a large-scale multimodal instruction-tuning dataset composed of synthetic dialogues that simulate realistic doctor-patient interactions. These dialogues are generated with the help of GPT-4 Turbo [66], guided by carefully designed prompts to control structure, medical accuracy, and user interactivity.

Our process begins by defining a multi-step prompting framework. First, a Text Description sets the scene for GPT-4 Turbo to generate helpful assistant-style responses, typically introduced with instructions such as:

> *"Please construct a multi-turn dialogue between a patient and an assistant. The assistant should be capable of identifying diseases in CXRs and describing their visual locations."*

Next, we Collect Data from the VinDr-CXR dataset [64], which provides disease labels and corresponding bounding box annotations. These data points are transformed into dialogue targets where the assistant needs to recognize the disease and explain its location in the image.

To enhance realism, we simulate Human Characteristics, where the user (patient) might express vague symptoms, request clarifications, or ask about visual evidence. This allows the model to demonstrate clinical reasoning, not just pattern recognition. We embed Dialogue Properties

to ensure coherence, such as logical transitions, varied question types, and diverse reasoning patterns.

To mark visual grounding entities within the dialogues, we adopt structured token conventions: disease labels are enclosed in <ref> ... </ref> and the bounding boxes are enclosed in <box> ... </box>. The bounding box itself follows the normalized coordinate format $(x_{tl}, y_{tl}), (x_{br}, y_{br})$. An example response might look like:

> ‘‘I found evidence of Atelectasis in the upper-left lung region, roughly located at <ref> Atelectasis </ref> <box> (150, 200), (350, 500) </box>.’’

We then provide an Example Dialogue containing a complete multi-round exchange, including follow-up questions, grounding justifications, and references to specific anatomical regions. The full prompt is submitted to GPT-4 Turbo, generating thousands of diverse and structured instruction dialogues. This synthetic dataset is then used for fine-tuning.

**Fine-Tuning with Generated Instruction Dialogues.** Following dataset generation, we proceed to fine-tune the LLM using LoRA [37], a parameter-efficient fine-tuning technique that inserts trainable low-rank matrices into the model's attention layers. This allows for significant adaptation of the model's behavior without altering the original pretrained weights. Fine-tuning with LoRA not only reduces memory and computation requirements but also ensures the preservation of general medical knowledge encoded in the base model.

The training objective remains auto-regressive language modeling, encouraging the model to predict the next token in the dialogue conditioned on both the image features and the conversation history. Through this instruction tuning stage, the model evolves from a passive image classifier to an interactive, visually grounded assistant capable of supporting patient education, radiology interpretation, and clinical consultation. The final model is thus optimized not just for accuracy, but for explainability, usability, and engagement—key attributes for real-world deployment in healthcare AI systems.

## 6.3 Experiment Settings

### 6.3.1 Implementation Details

Initially, our model undergoes task-specific pre-training using LoRA [37], a parameter-efficient fine-tuning method, applied on two large-scale public datasets: MIMIC-CXR [43] and VinDr-CXR [64]. This stage enables the model to acquire foundational competencies in two essential tasks for medical image analysis: multi-label disease classification and visual grounding. MIMIC-CXR provides comprehensive image-text pairs for disease labeling, while VinDr-CXR contributes high-quality bounding box annotations, allowing the model to associate textual findings with their spatial locations on chest X-rays.

Following the pre-training, we curate a high-quality multimodal instruction tuning dataset using GPT-4 Turbo [66]. This dataset simulates multi-turn diagnostic dialogues that incorporate both textual disease descriptions and visual references via specialized tokens such as $< ref >$ and $< box >$. These synthetic dialogues guide the model to reason through clinical interactions, identify conditions from images, and explain their locations through language-grounded visual annotations.

For training, we fix the maximum sequence length to 768 tokens to balance expressivity and efficiency. We train the model using the Adam optimizer [47], which is well-suited for transformer-based architectures. The learning rate is initialized at $2 \times 10^{-5}$, and the batch size is set to 4 due to GPU memory constraints. All experiments are conducted on a single NVIDIA RTX A6000 GPU, which provides 48GB of VRAM, enabling us to handle large-scale vision-language inputs. The training pipeline includes one full epoch for pre-training across the classification and grounding datasets, followed by an additional epoch for instruction tuning. This two-stage training paradigm ensures that the model first develops low-level task expertise before acquiring high-level reasoning and conversational abilities, resulting in a robust and interactive medical assistant.

### 6.3.2 Datasets

We validate our method on two public datasets, MIMIC-CXR [43] and VinDr-CXR [64]. MIMIC-CXR is recognized as the most extensive publicly accessible dataset that includes chest radiographs paired with free-text reports, spanning 14 lung disease diagnosis categories. In our study, we adhered to the official dataset division guidelines provided by MIMIC-CXR to ensure consistency and facilitate equitable comparisons. As a result, our training dataset consists of 270,790 image-label pairs, with additional sets of 2,130 and 3,858 samples designated for validation and testing, respectively. VinDr-CXR, notable for its extensive collection of over 100,000 chest X-ray scans from leading Vietnamese hospitals, stands out as a significant resource for disease visualization. It includes 18,000 carefully annotated images by 17 radiologists, offering detailed local and global disease labels. Structured into 15,000 training and 3,000 testing images, it aids in developing machine learning models for chest condition identification and localization. Additionally, a custom DICOM image labeling platform enhances the annotation workflow.

## 6.4 Evaluation and Discussion

### 6.4.1 Evaluation Metrics.

We adopt the evaluation framework introduced in Med-PaLM M [94] to assess our classification task, ensuring direct comparability with other LLM-based approaches on the MIMIC-CXR dataset [43]. This framework emphasizes both clinical relevance and prevalence by focusing on five major thoracic conditions: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. Performance is primarily measured using Macro-AUC and Macro-F1 scores, which better capture model performance across imbalanced classes. In addition to overall metrics, we specifically report the per-disease Macro-AUC to analyze the model's performance on each condition individually, providing insight into its robustness across varying pathologies.

**Table 6.1:** Classification performanc comparison for five major classes on MIMIC-CXR dataset by Accuracy.

| Metrics | Lesion | PaLM-E(84B) [23] | LLaVa-Med [55] | Med-PaLM M [94] | mCNN [72] | Ours | Ours+CoT |
|---|---|---|---|---|---|---|---|
| Macro-AUC | Ate. | N/A | 65.60% | N/A | 77.03% | 81.83% | **84.05%** |
| | Car. | N/A | 67.05% | N/A | 79.32% | 95.46% | **96.95%** |
| | Con. | N/A | 68.29% | N/A | 75.97% | 73.54% | **78.24%** |
| | Ede. | N/A | 74.63% | N/A | 84.19% | 84.26% | **87.12%** |
| | PEf. | N/A | 71.03% | N/A | **89.85%** | 80.69% | 84.52% |
| | **Avg.** | 51.48% | 69.32% | 79.09% | 81.27% | 83.14% | **85.46%** |
| Macro-F1 | **Avg.** | 7.83% | 30.68% | 41.57% | 43.07% | 45.52% | **48.22%** |

For the visual grounding task on the VinDr-CXR dataset [64], we adopt mean Intersection over Union (mIoU) and accuracy (Acc) as the core evaluation metrics. Following standard practices in natural image visual grounding [20], a predicted bounding box is considered correct (positive) if its IoU with the ground truth exceeds a specified threshold. We compute Average Precision (AP) at multiple thresholds ranging from 0.1 to 0.5, denoted as AP10 to AP50, which offers a comprehensive view of localization accuracy under varying tolerance levels. Additionally, we report disease-specific grounding accuracy at AP40. Unlike natural image benchmarks where AP50 is standard, medical visual grounding typically adopts lower IoU thresholds (e.g., AP30–AP40) because thoracic lesions tend to be small, diffuse, or ambiguous in shape, leading to greater inter-observer variation among radiologists. AP40 thus provides a more stable and clinically realistic evaluation of localization performance, highlighting the model's ability to identify pathological regions with interpretable annotations.

### 6.4.2 Performance Comparison

#### Classification Results on MIMIC-CXR Dataset

Table 6.1 offers a detailed comparison of the classification performance of our proposed MedVisioChat model against several state-of-the-art LLMs and traditional CNN-based methods. The evaluation is conducted on the official test split of the MIMIC-CXR dataset, focusing on five clinically significant thoracic conditions: Atelectasis (Ate.), Cardiomegaly (Car.), Consolidation (Con.), Edema (Ede.), and Pleural Effusion (PEF.). Two primary evaluation metrics are used: Macro-AUC, which measures the area under the ROC curve for each class and averages the results,

and Macro-F1, which emphasizes the harmonic mean between precision and recall across all categories, offering a balanced view of classification performance, especially under class imbalance.

Our model consistently outperforms all LLM baselines, achieving average scores of 83.14% in Macro-AUC and 45.52% in Macro-F1. When further equipped with Chain-of-Thought (+CoT) reasoning, the performance is boosted even higher, reaching 85.46% Macro-AUC and 48.22% Macro-F1—the best results across all evaluated methods. These improvements highlight the complementary role of reasoning-based prompting, which enhances the model's capability to capture subtle inter-class differences and reduce misclassification in borderline cases. Compared to PaLM-E (Macro-AUC: 51.48%, Macro-F1: 7.83%), LLaVa-Med (Macro-AUC: 69.32%, Macro-F1: 30.68%), and Med-PaLM M (Macro-AUC: 79.09%, Macro-F1: 41.57%), our approach demonstrates a clear advantage, underscoring the limitations of existing LLMs in structured medical classification tasks.

Notably, while mCNN [72], a dedicated CNN-based model, remains competitive in certain lesion category (e.g., Pleural Effusion at 89.85%), its overall performance (Macro-AUC: 81.27%, Macro-F1: 43.07%) lags behind both our base model and our CoT variant. This indicates that although CNNs still capture some well-defined radiological patterns effectively, our unified dialogue-grounded framework, especially with CoT reasoning, better generalizes across diverse pathologies and aligns more closely with clinical interpretability.

At the lesion-wise level, the CoT variant achieves the highest accuracy in four out of five categories: Atelectasis (84.05%), Cardiomegaly (96.95%), Consolidation (78.24%), and Edema (87.12%), further surpassing both mCNN and all LLM baselines. For Pleural Effusion, our model (84.52%) performs slightly below mCNN (89.85%) but remains substantially above LLM baselines. These results suggest that while CNNs may retain an edge in certain localized findings, the integration of CoT reasoning into our framework enables broader and more consistent improvements across conditions.

**Table 6.2:** Visual grounding performance comparison on VinDr-CXR dataset.

| Method | AP10 | AP20 | AP30 | AP40 | AP50 | mIOU |
|---|---|---|---|---|---|---|
| SeqTR [7] | 54.1% | 46.8% | 38.3% | 31.2% | 23.4% | 21.7% |
| DACR [65] | 56.6% | 50.1% | 43.1% | 36.5% | 28.7% | 26.5% |
| MedRPG [12] | 57.2% | 51.7% | 43.3% | 37.4% | 29.8% | 27.3% |
| Ours | 61.1% | 52.5% | 45.3% | 38.9% | 31.6% | 32.4% |
| **Ours + CoT** | **62.4%** | **53.1%** | **45.9%** | **39.5%** | **32.9%** | **33.6%** |

Interestingly, Macro-AUC is much higher than Macro-F1, because in medical classification with severe class imbalance, Macro-F1 is disproportionately penalized while ROC-AUC remains high. ROC-AUC is threshold-independent and evaluates ranking quality, so a model that correctly orders positives above negatives can achieve a high AUC even for rare diseases. By contrast, $F_1 = 2PR/(P+R)$ is computed at a single decision threshold and is highly sensitive to imbalance: for rare classes, a default threshold often yields very low recall (few positives cleared) or, when recall is increased, many false positives that depress precision—both sharply reducing $F_1$. Because Macro-F1 averages $F_1$ equally across classes, a few difficult/low-prevalence categories can pull the overall score down, whereas ROC-AUC—reflecting ranking across all thresholds—can remain high despite the same prevalence skew.

Collectively, these findings demonstrate that MedVisioChat not only outperforms existing LLMs but also narrows or even exceeds the performance of specialized CNNs. By integrating dialogue-based interaction, grounding capabilities, and reasoning-enhanced classification, our framework offers a powerful, interpretable, and versatile solution for real-world deployment in medical AI systems.

### Visual Grounding Results on VinDr-CXR Dataset

Table 6.2 and Table 6.3 offer a comprehensive and fine-grained evaluation of the proposed model's visual grounding performance on the VinDr-CXR dataset. In Table 6.2, we benchmark our method against three representative state-of-the-art grounding models—SeqTR [7], DACR [65], and MedRPG [12]—across multiple commonly used metrics: AP at

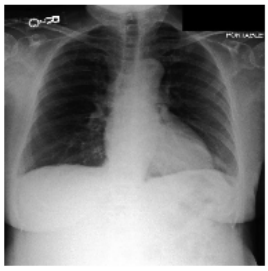
**Table 6.3:** Visual grounding performance comparison on VinDr-CXR dataset by AP40 for each disease.

| Disease Category | SeqTR [7] | DACR [65] | MedRPG [12] | Ours | Ours + CoT |
|---|---|---|---|---|---|
| Aortic Enlargement | 54.2% | 66.3% | **95.3%** | 78.9% | 80.4% |
| Atelectasis | 12.3% | 23.1% | 17.1% | 25.0% | **26.5%** |
| Calcification | 9.1% | **27.2%** | 10.3% | 13.2% | 14.7% |
| Cardiomegaly | 71.2% | 86.0% | 80.2% | 94.9% | **96.4%** |
| Consolidation | 10.3% | **28.1%** | 19.1% | 12.5% | 14.0% |
| Interstitial Lung Disease | 24.8% | 31.5% | 11.1% | 32.4% | **33.9%** |
| Infiltration | 25.2% | 31.8% | 11.3% | 35.1% | **36.6%** |
| Lung Opacity | 11.9% | 19.7% | 13.7% | 20.1% | **21.6%** |
| Nodule / Mass | 13.4% | 25.1% | 4.9% | 28.9% | **30.4%** |
| Pleural Effusion | 15.6% | **38.7%** | 13.0% | 21.7% | 23.2% |
| Pleural Thickening | 10.1% | 22.8% | 2.7% | 20.1% | **24.6%** |
| Pneumothorax | 26.8% | **57.9%** | 3.6% | 43.1% | 44.6% |
| Pulmonary Fibrosis | 17.4% | 34.0% | 9.4% | 49.3% | **50.8%** |
| Other Lesions | 19.3% | **38.1%** | 7.7% | 29.4% | 30.9% |
| **Average (AP40)** | 31.2% | 36.5% | 37.4% | 38.9% | **40.4%** |

different IoU thresholds (from AP10 to AP50) and mIOU. Notably, our method outperforms all baselines across every evaluation metric. For instance, our model achieves an AP10 of 61.1%, substantially higher than 54.1% by SeqTR and 56.6% by DACR. At more stringent thresholds such as AP40 and AP50, our model maintains a clear lead with 38.9% and 31.6% respectively, indicating its robustness in precise region localization. Moreover, our mIOU score of 32.4% reflects superior spatial accuracy in identifying disease-relevant areas, outperforming the closest baseline MedRPG by a margin of 5.1%. With a chain-of-thought grounding stage (*Ours + CoT*), the aggregate metrics further improve across the board (see Table 6.2): AP10 62.4%, AP20 53.1%, AP30 45.9%, AP40 39.5%, AP50 32.9%, and mIOU 33.6%—i.e., consistent gains of about 0.6–1.3 percentage points over *Ours* at each threshold, highlighting stronger robustness under stricter localization criteria.

To further validate our model's performance across various clinical scenarios, Table 6.3 breaks down the visual grounding results (AP40) by each disease category in the VinDr-CXR dataset. These diseases span across a diverse set of pathologies, including lung abnormalities (e.g., Atelectasis, Pneumothorax), cardiac conditions (e.g., Cardiomegaly), and pleural findings (e.g., Pleural Effusion, Pleural Thickening). Our model delivers consistent gains across most disease types, indicating its ability to generalize across anatomical regions and pathological appearances.

CoT-Based Classification



Question:
This is a 14-class classification task related to medical imaging, where the 14 disease categories are as follows: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion,Pleural Other, Pneumonia, Pneumothorax, Support Devices. Please output "Positive" or "Negative" for each category. Output the thinking process in <think></think> and the final answer in <answer></answer> tags. i.e., <think> reasoning process here </think><answer> answer here </answer>.

Answer:
<think>Based on visual cues in this single AP portable chest radiograph, cardiomegaly was labeled positive because the cardiac silhouette appears relatively enlarged compared with the thoracic width; pneumothorax was labeled positive because a peripheral lucent area with diminished vascular markings is compatible with a pleural line; and support devices were labeled positive due to a high-density device/lead seen over the upper chest. Consolidation/pneumonia, pleural effusion, edema, atelectasis, lung lesion/opacity, pleural other, fracture, and enlarged cardiomediastinum were labeled negative because there is no convincing lobar/segmental dense opacity, costophrenic blunting or meniscus, typical interstitial–alveolar edema pattern, plate-like atelectasis with volume loss, discrete nodule/mass, abnormal pleural finding, rib discontinuity, or obvious mediastinal widening.</think><answer>Atelectasis:Negative, Cardiomegaly:Positive, Consolidation:Negative, Edema:Negative, Enlarged Cardiomediastinum:Negative, Fracture:Negative, Lung Lesion:Negative, Lung Opacity:Negative, No Finding:Negative, Pleural Effusion:Negative, Pleural Other:Negative, Pneumonia:Negative, Pneumothorax:Positive, Support Devices:Positive.</answer>

**Figure 6.2:** Case 1 of CoT-Based Classification by MedVisioChat.

Specifically, it achieves the best AP40 in key categories such as Atelectasis (25.0%), Cardiomegaly (94.9%), Interstitial Lung Disease (32.4%), Infiltration (35.1%), Lung Opacity (20.1%), Nodule/Mass (28.9%) and Pulmonary Fibrosis (49.3%), where spatial ambiguity and inter-observer variability often pose challenges to other models. While DACR performs relatively well in categories like Calcification and Pleural Effusion, our model still delivers more balanced results overall. In contrast, MedRPG performs strongly in Aortic Enlargement but fails to maintain reasonable accuracy in other diseases, suggesting its limited generalizability. Importantly, *Ours + CoT* yields additional per-disease improvements and raises the overall mean from 38.9% to 40.4% (+1.5 pp) in Table 6.3. Representative gains include Atelectasis 25.0%→26.5%, Cardiomegaly 94.9%→96.4%, Interstitial Lung Disease 32.4%→33.9%, Infiltration 35.1%→36.6%, Lung Opacity 20.1%→21.6%, Nodule/Mass 28.9%

CoT-Based Classification

Question:
This is a 14-class classification task related to medical imaging, where the 14 disease categories are as follows: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, Support Devices. Please output "Positive" or "Negative" for each category. Output the thinking process in <think></think> and the final answer in <answer></answer> tags. i.e., <think> reasoning process here </think><answer> answer here </answer>.

Answer:

<think>On inspection of this AP chest radiograph, the cardiac silhouette is within normal limits, the mediastinum is not widened, and there is no evidence of focal consolidation, diffuse edema, pleural effusion, pneumothorax, lung masses, fractures, or support devices. The lung fields appear clear and symmetric, and no acute abnormality is identified. Therefore all specific disease categories are labeled Negative, and the overall finding is classified as No Finding:Positive.</think><answer> Atelectasis:Negative, Cardiomegaly:Negative, Consolidation:Negative, Edema:Negative, Enlarged Cardiomediastinum:Negative, Fracture:Negative, Lung Lesion:Negative, Lung Opacity:Negative, No Finding:Positive, Pleural Effusion:Negative, Pleural Other:Negative, Pneumonia:Negative, Pneumothorax:Negative, Support Devices:Negative. </answer>

**Figure 6.3:** Case 2 of CoT-Based Classification by MedVisioChat.

→30.4%, and Pulmonary Fibrosis 49.3%→50.8%, evidencing broader generalization across anatomical regions and pathological appearances.

The superiority of our model can be attributed to its unified architecture, which tightly integrates region-aware visual features with language-driven disease reasoning, as well as the high-quality instruction-tuning dataset generated via GPT-4 Turbo [66]. The instruction tuning not only improves alignment between visual and textual modalities but also helps the model disambiguate complex multi-disease contexts during grounding. Furthermore, the integration of CoT reasoning endows the model with an explicit deliberation process before outputting classification results, enabling more accurate multi-label predictions in medical contexts (as illustrated in Figure 6.2 and Figure 6.3). In addition, the inclusion of a dialogue interface (Figure 6.4 and Figure 6.5) allows for transparent diagnostic interactions, where users can query for condition-specific visual evidence and receive both bounding box visualizations

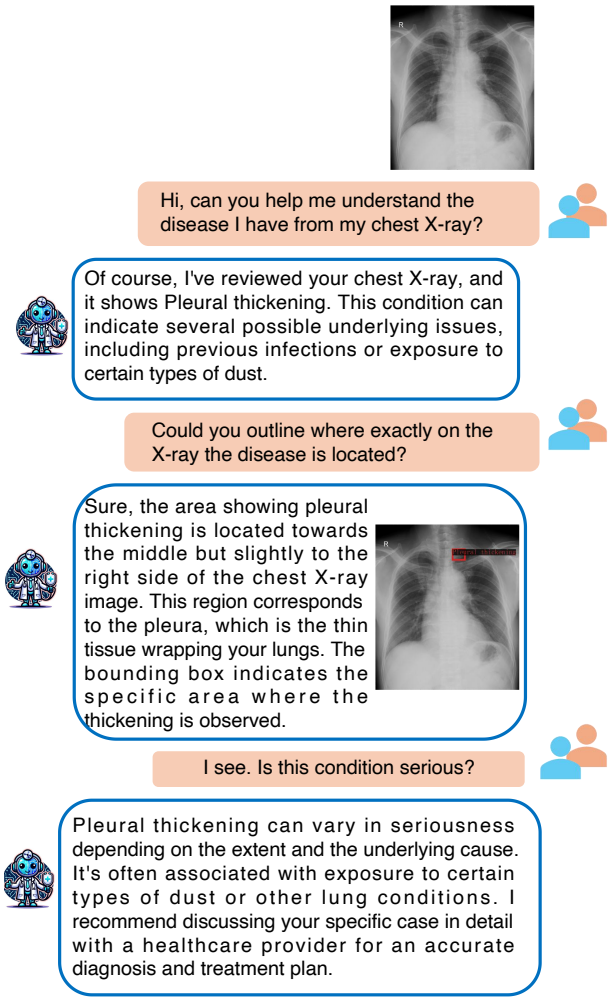



**Figure 6.4:** Case 1 of Medical Comprehension by MedVisioChat.

and textual explanations. This elevates the interpretability of AI-driven diagnostics and makes the system more aligned with real-world clinical workflows, where explainability and user trust are paramount. Collectively, these detailed comparisons across both aggregate metrics and disease-specific evaluations highlight the effectiveness, robustness, and clinical usability of our MedVisioChat framework in visual grounding for chest X-ray interpretation.

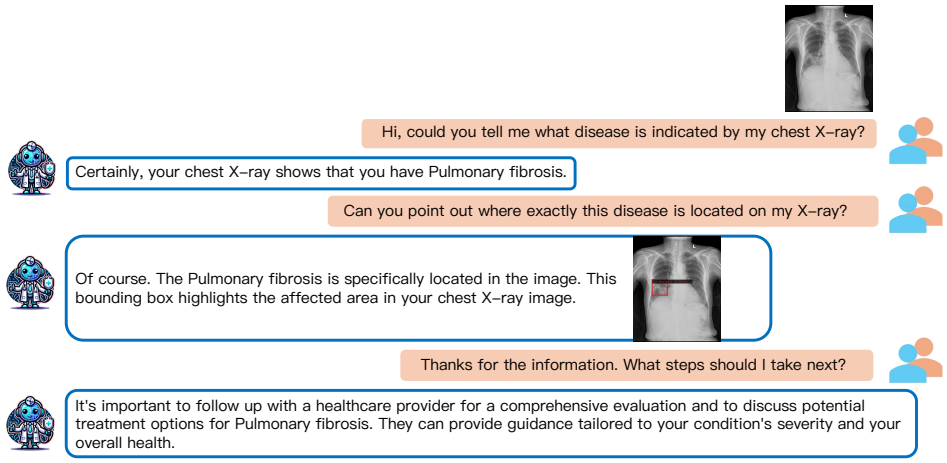



**Figure 6.5:** Case 2 of Medical Comprehension by MedVisioChat.

### 6.4.3 Limitation

There are several limitations in this study that highlight areas for future improvement. First, while LLMs have demonstrated strong generalization and reasoning abilities across a variety of domains, their performance is often closely tied to the availability of large-scale, high-quality training data. In the medical imaging domain—particularly in specialized tasks such as chest X-ray classification, report generation, and visual grounding—the size and diversity of publicly available datasets remain relatively limited. This constraint restricts the model's ability to learn robust representations that generalize well across different patient populations, disease types, and imaging devices. Moreover, the fine-tuning process in our study is based on a finite number of samples from datasets like MIMIC-CXR and VinDr-CXR. Although these datasets are among the largest in the medical imaging field, they still do not match the scale of datasets used to train general-domain LLMs. As a result, the model's ability to handle rare pathologies or subtle visual features may be suboptimal, especially when compared to its performance on more prevalent disease categories. Finally,, while we have employed instruction tuning to improve the model's ability to engage in interactive, explainable dialogues, the quality of instruction-following behavior is influenced by the diversity and realism of the generated dialogues. Our instruction data, though carefully constructed using GPT-4 Turbo, may not fully capture

the complexity and nuance of real-world clinical conversations, limiting the model's practical utility in dynamic clinical environments.

## 6.5 Summary

In this chapter, we introduce MedVisioChat, a LLM framework designed for interpretable medical diagnosis through visual grounding in CXRs. By combining prompt engineering, chain of thought and a LLM fine-tuned with instruction-following data generated by GPT-4 Turbo, Med-VisioChat performs both multi-label disease classification and precise localization using bounding boxes. The model is pre-trained on MIMIC-CXR and VinDr-CXR datasets and demonstrates superior performance in both classification and visual grounding tasks compared to state-of-the-art baselines. This work bridges the gap between vision-language understanding and clinical interpretability, offering a more transparent and interactive diagnostic tool. Our contributions are as follows:

- We introduce MedVisioChat, a novel LLM framework that seamlessly integrates disease classification and visual grounding capabilities with textual reasoning and knowledge representation to support comprehensive, multi-turn dialogues in medical diagnosis. Unlike traditional diagnostic AI models that function as passive classifiers, Med-VisioChat actively engages in human-like conversations, interprets CXR images, and offers visual explanations aligned with clinical findings. This enables clinicians and patients alike to interact with the system in a more interpretable and transparent manner, thereby transforming the process of medical image interpretation and ultimately enhancing diagnostic accuracy, trust, and outcomes in real-world settings.
- We propose a task-specific prompt design and a chain of thought reward tailored for multi-label disease classification in the medical domain, addressing challenges such as class imbalance and the subtle visual distinctions in CXR images. Our prompt design enables the model to sequentially reason through multiple disease

classes while maintaining consistency and interpretability. In addition, we contribute a comprehensive instruction-tuning dataset specifically curated for the healthcare domain, comprising thousands of high-quality, GPT-4 Turbo–generated multi-round dialogues annotated with disease names and bounding box information. This dataset plays a crucial role in training and evaluating the model's interaction quality and grounding performance, and will be made publicly available alongside the MedVisioChat model to facilitate reproducibility and encourage further research in trustworthy and explainable medical AI.

# Chapter 7

# Conclusion and Future Work

## 7.1 Conclusion

In this thesis, we systematically investigated the challenges in medical image analysis and synthesis, with a particular focus on CXRs. We began by reviewing the current landscape of research, identifying key limitations in three critical areas: bone suppression, unified LLMs for CXR understanding and generation, and LLMs for CXR classification and visual grounding. To address these challenges, we proposed four novel methods designed to enhance both the accuracy and robustness of medical image analysis and synthesis. Collectively, these contributions mark significant progress in the field of medical deep learning and offer effective solutions for improving the clinical utility and reliability of medical image analysis and synthesis.

In Chapter 3, we proposed a novel bone suppression framework aimed at mitigating the interference of bony structures, thereby improving performance in downstream clinical tasks such as classification and segmentation. Our method integrates an enhanced multi-head codebook attention module within a two-stage suppression architecture, ensuring that while bone shadows are effectively removed, critical anatomical regions—such as lung fields and lesions—are preserved with high fidelity. Experimental results on the Bone Suppression and NIH14 datasets demonstrate that our approach not only improves visual quality but also significantly boosts performance in diagnostic applications. By reducing

structural noise from bones, this framework has the potential to assist radiologists and enhance the reliability of automated diagnostic systems.

In Chapter 4, our research presents MedXChat, a unified LLM framework designed for multi-task CXR understanding and generation. MedXChat addresses three key clinical tasks: report generation (CXR-to-Report), visual question answering (CXR-VQA), and image synthesis (Text-to-CXR). Unlike prior models that rely heavily on vector quantization, MedXChat incorporates instruction-tuned data generated by GPT-4 and finetunes a Stable Diffusion module for high-quality image synthesis. Comprehensive evaluations across computational metrics and downstream clinical tasks demonstrate that MedXChat consistently surpasses state-of-the-art baselines in terms of accuracy, consistency, and image fidelity. This work underscores the promise of unified LLM frameworks in advancing intelligent radiology applications for both clinical practice and medical education.

In Chapter 5, we collaborated with professional radiologists— comprising three junior radiologists, two senior radiologists, and one supervising radiologist—to establish a systematic and expert-driven evaluation and comparison of MedXChat against other large language models. Our findings revealed that many commonly used computational metrics, such as BLEU-4 and FID, do not fully capture the models' clinical utility. By integrating both radiologists' assessments and computational metrics, we demonstrated that MedXChat remains effective as a unified multimodal LLM framework for chest X-ray interpretation, encompassing automated report generation, visual question answering, and medical image synthesis within a single architecture. These results underscore MedXChat's clinical applicability and position it as a strong candidate for next-generation intelligent assistant systems in radiology, characterized by interpretability, adaptability, and robust clinical grounding.

In Chapter 6, we propose MedVisioChat, a LLM tailored for medical classification and visual grounding. Unlike existing large models that directly perform classification, we design a dedicated prompt structure for 14 disease categories, requiring the model to output explicit positive or negative responses. To further improve classification performance, we

introduce a novel reward mechanism consisting of an accuracy reward, which encourages correct predictions, and a format reward, which enforces structured outputs combining both the Chain-of-Thought (CoT) reasoning and the final answer. This reward design significantly enhances the model's ability to produce accurate and interpretable classifications. For visual grounding, we develop a set of special tokens that explicitly connect disease categories with their corresponding bounding boxes, using instruction data generated by GPT. This design helps the LLM better understand the semantic link between diagnosis and spatial localization, thereby improving its ability to generate bounding boxes that accurately highlight disease-relevant regions. Extensive experiments demonstrate that MedVisioChat not only achieves superior classification accuracy compared to baseline LLMs, but also yields more reliable and clinically meaningful visual grounding results. These findings highlight the importance of combining CoT-enhanced classification with structured instruction data for advancing multimodal medical AI systems.

Based on the extensive experiments and ablation studies conducted on public datasets, our results demonstrate the effectiveness of our proposed methods in medical image analysis and synthesis. By comparing our approach with popular existing methods, we have shown that our techniques consistently achieve superior performance across various tasks.

## 7.2 Future work

Considering recent developments in medical image analysis and deep learning technology, we propose three potential research directions for future work.

- In future work, we plan to explore more efficient and lightweight backbone architectures for bone suppression to improve computational efficiency without compromising performance. One promising direction is to incorporate sparse attention mechanisms into the Vision Transformer (ViT) architecture, which can significantly

reduce the quadratic complexity of self-attention while preserving global contextual modeling capabilities. Additionally, we will investigate model compression techniques such as pruning and quantization to further reduce inference time and memory usage, enabling deployment in resource-constrained clinical environments. Furthermore, we plan to expand the dataset by including more clinically diverse and representative samples across various disease types, age groups, and imaging protocols. We also aim to explore semi-supervised or synthetic data generation strategies to augment the training set, further enhancing the robustness and generalizability of the proposed framework in real-world radiology workflows.

- To further enhance the generalizability and clinical value of our unified medical LLM framework, future research should first prioritize diversifying the underlying data sources. While the current study primarily relies on the MIMIC-CXR dataset, its scope remains limited in reflecting the full variability of real-world clinical settings. A more robust solution would involve incorporating chest X-ray data from multiple institutions, regions, and imaging devices, thereby exposing the model to diverse patient populations and technical conditions. Beyond simple data aggregation, domain adaptation strategies such as adversarial training or feature alignment could be applied to reduce distributional shifts across institutions, ensuring more reliable cross-site performance.

- The instruction-tuning process presents an opportunity for refinement. Presently, synthetic dialogue data are used to simulate physician–patient interactions; however, these approximations often fail to capture the subtle linguistic and reasoning nuances present in clinical conversations. Replacing synthetic data with real-world clinical dialogues would strengthen the model's ability to understand domain-specific language, better align its reasoning chains with professional diagnostic practices, and improve its adaptability to patient-facing applications where natural, conversational understanding is critical.

# Bibliography

[1] D. Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *ICLR* (2020).

[2] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy. "Automated radiology report generation using conditioned transformers". In: *Informatics in Medicine Unlocked* 24 (2021), p. 100557.

[3] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang. "MedGAN: Medical image translation using GANs". In: *Computerized medical imaging and graphics* 79 (2020), p. 101684.

[4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond". In: (2023).

[5] S. Banerjee and A. Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *ACL workshop*. 2005.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[7] Y. S. Chaoyang Zhu Yiyi Zhou. "Seqtr: A simple yet universal network for visual grounding". In: *ECCV* (2022).

[8] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris. "Multimodal MR synthesis via modality-invariant latent representation". In: *IEEE transactions on medical imaging* 37.3 (2017), pp. 803–814.

[9] S. Chen and K. Suzuki. "Separation of bones from chest radiographs by means of anatomically specific multiple massive-training

ANNs combined with total variation minimization smoothing". In: *IEEE transactions on medical imaging* 33.2 (2013), pp. 246–257.

[10] Z. Chen, S. Gu, G. Lu, and D. Xu. "Exploiting intra-slice and inter-slice redundancy for learning-based lossless volumetric image compression". In: *TIP* (2022).

[11] Z. Chen, L. Zhou, Z. Hu, and D. Xu. "Group-aware parameter-efficient updating for content-adaptive neural video compression". In: *ACM Multimedia*. 2024.

[12] Z. Chen, Y. Zhou, A. Tran, J. Zhao, L. Wan, G. S. K. Ooi, L. T.-E. Cheng, C. H. Thng, X. Xu, Y. Liu, et al. "Medical phrase grounding with region-phrase context contrastive alignment". In: *MICCAI*. 2023.

[13] Z. Chen, Y. Shen, Y. Song, and X. Wan. "Cross-modal memory networks for radiology report generation". In: *ACL*. 2022.

[14] Z. Chen, Y. Song, T.-H. Chang, and X. Wan. "Generating radiology reports via memory-driven transformer". In: *ACL*. 2020.

[15] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality". In: *See https://vicuna. lmsys. org (accessed 14 April 2023)*. 2023.

[16] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. "Stargan v2: Diverse image synthesis for multiple domains". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8188–8197.

[17] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara. "Meshed-memory transformer for image captioning". In: *CVPR*. 2020.

[18] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces". In: *Medical image analysis* 57 (2019), pp. 226–236.

[19] O. Dalmaz, M. Yurt, and T. Çukur. "ResViT: Residual Vision Transformers for Multi-modal Medical Image Synthesis". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2021, pp. 541–550.

[20] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li. "Transvg: End-to-end visual grounding with transformers". In: *ICCV*. 2021.

[21] R. Dorent, N. Haouchine, F. Kogl, S. Joutard, P. Juvekar, E. Torio, A. J. Golby, S. Ourselin, S. Frisken, T. Vercauteren, et al. "Unified brain MR-ultrasound synthesis using multi-modal hierarchical representations". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2023, pp. 448–458.

[22] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *International Conference on Learning Representations* (2021). URL: https://openreview.net/forum?id=YicbFdNTTy.

[23] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. "Palm-e: An embodied multimodal language model". In: *arXiv preprint arXiv:2303.03378* (2023).

[24] M. Eslami, S. Tabarestani, S. Albarqouni, E. Adeli, N. Navab, and M. Adjouadi. "Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography". In: *IEEE transactions on medical imaging* 39.7 (2020), pp. 2553–2565.

[25] P. Esser, R. Rombach, and B. Ommer. "Taming transformers for high-resolution image synthesis". In: *CVPR*. 2021.

[26] I. Gabriel, W. Mitchell, W. Ross, G. Cade, C. Nicholas, and T. Rohan. *Openclip*. https://doi.org/10.5281/zenodo.5143773.. 2021.

[27] Y. Ge, S. Zhao, Z. Zeng, Y. Ge, C. Li, X. Wang, and Y. Shan. "Making LLaMA SEE and Draw with SEED Tokenizer". In: *arXiv preprint arXiv:2310.01218*. 2023.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[29] M. Gusarev, R. Kuleev, A. Khan, A. R. Rivera, and A. M. Khattak. "Deep learning models for bone suppression in chest radiographs". In: *2017 IEEE conference on computational intelligence in*

*bioinformatics and computational biology (CIBCB)*. IEEE. 2017, pp. 1–7.

[30] L. Han, T. Tan, T. Zhang, X. Wang, Y. Gao, C. Lu, X. Liang, H. Dou, Y. Huang, and R. Mann. "Non-adversarial Learning: Vector-Quantized Common Latent Space for Multi-sequence MRI". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 481–491.

[31] T. Hashimoto, A. Maruo, S. Sekiguchi, Y. Ushiku, and G. Shimbo. "Unsupervised Domain Adaptation for Human and Animal Chest X-Ray Bone Suppression". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[32] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in neural information processing systems* 30 (2017).

[35] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework". In: *International conference on learning representations*. 2017.

[36] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "Lora: Low-rank adaptation of large language models". In: *ICLR*. 2021.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[40] S. Jaeger et al. "Automatic screening for tuberculosis in chest radiographs: a survey". In: *Quantitative imaging in medicine and surgery* 3.2 (2013), p. 89.

[41] S. Jaeger et al. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases". In: *Quantitative imaging in medicine and surgery* 4.6 (2014), p. 475.

[42] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, et al. "Radgraph: Extracting clinical entities and relations from radiology reports". In: *arXiv preprint arXiv:2106.14463* (2021).

[43] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". In: *arXiv preprint arXiv:1901.07042*. 2019.

[44] J. Kapoor, J. H. Macke, and C. F. Baumgartner. "Multiscale metamorphic vae for 3d brain mri synthesis". In: *arXiv preprint arXiv:2301.03588* (2023).

[45] D. S. Kermany et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning". In: *cell* 172.5 (2018), pp. 1122–1131.

[46] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, et al. "Denoising diffusion probabilistic models for 3D medical image generation". In: *Scientific Reports* 13.1 (2023), p. 7303.

[47] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *International Conference on Learning Representations* (2015). URL: https://arxiv.org/pdf/1412.6980.

[48] D. P. Kingma and M. Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[49] J. Y. Koh, D. Fried, and R. Salakhutdinov. "Generating Images with Multimodal Language Models". In: *NeurIPS*. 2023.

[50] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. "Large language models are zero-shot reasoners". In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[51] J. E. Kuhlman, J. Collins, G. N. Brooks, D. R. Yandow, and L. S. Broderick. "Dual-energy subtraction chest radiography: what to look for beyond calcified nodules". In: *Radiographics* 26.1 (2006), pp. 79–92.

[52] D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han. "Autoregressive image generation using residual quantization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11523–11532.

[53] H. Lee, W. Kim, J.-H. Kim, et al. "Unified Chest X-ray and Radiology Report Generation Model with Multi-view Chest X-rays". In: *arXiv preprint arXiv:2302.12172*. 2023.

[54] S. Lee, W. J. Kim, J. Chang, and J. C. Ye. "LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation". In: *ICLR*. 2023.

[55] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day". In: *NIPS* (2024).

[56] H. Li et al. "High-resolution chest x-ray bone suppression using unpaired CT structural priors". In: *IEEE transactions on medical imaging* 39.10 (2020), pp. 3053–3063.

[57] Y. Li, Y. Liu, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, and L. Zhou. "A comprehensive study of gpt-4v's multimodal capabilities in medical imaging". In: *medRxiv* (2023), pp. 2023–11.

[58] C.-W. Lin, C.-H. Ho, C. C. Ho, and Z.-W. Wang. "Automatic Pneumothorax Segmentation in Chest Radiographs Preprocessed by IEDSR Bone Suppression". In: *2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*. IEEE. 2024, pp. 419–420.

[59] C.-Y. Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004.

[60] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. "Exploring and distilling posterior and prior knowledge for radiology report generation". In: *CVPR*. 2021.

[61] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing". In: *ACM computing surveys* 55.9 (2023), pp. 1–35.

[62] J. Lu, C. Xiong, D. Parikh, and R. Socher. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning". In: *CVPR*. 2017.

[63] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, and L. Shao. "Structure preserving stain normalization of histopathology images using self supervised semantic guidance". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 309–319.

[64] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh, et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations". In: *Scientific Data* (2022).

[65] N. H. Nguyen, H. Q. Nguyen, N. T. Nguyen, T. V. Nguyen, H. H. Pham, and T. N.-M. Nguyen. "A clinical validation of VinDr-CXR, an AI system for detecting abnormal chest radiographs". In: *arXiv preprint arXiv:2104.02256* (2021).

[66] OpenAI. "GPT-4 Technical Report". In: 2023.

[67] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation". In: *ACL*. 2002.

[68] B. Peng, C. Li, P. He, M. Galley, and J. Gao. "Instruction tuning with gpt-4". In: *arXiv preprint arXiv:2304.03277*. 2023.

[69] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, et al. "Learning transferable visual models from natural language supervision". In: *ICML*. 2021.

[70] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. "Improving language understanding by generative pre-training". In: OpenAI, 2018.

[71] S. Rajaraman, G. Zamzmi, L. Folio, P. Alderson, and S. Antani. "Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings". In: *Diagnostics* 11.5 (2021), p. 840.

[72] R. S. Rammuni Silva and P. Fernando. "Effective utilization of multiple convolutional neural networks for chest X-ray classification". In: *SN Computer Science* (2022).

[73] N. Reimers and I. Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[74] F. Ren and Y. Zhou. "Cgmvqa: A new classification and generative model for medical visual question answering". In: *IEEE Access* 8 (2020), pp. 50626–50636.

[75] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. "Self-critical sequence training for image captioning". In: *CVPR*. 2017.

[76] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[77] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *MICCAI*. 2015.

[78] U. Sara, M. Akter, and M. S. Uddin. "Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study". In: *Journal of Computer and Communications* 7.3 (2019), pp. 8–18.

[79] M. M. K. Sarker et al. "COMFormer: classification of maternal-fetal and brain anatomy using a residual cross-covariance attention guided transformer in ultrasound". In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* (2023), pp. 1417–1427.

[80] R. R. Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[81] D. Sharma, S. Purushotham, and C. K. Reddy. "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain". In: *Scientific Reports* 11.1 (2021), p. 19826.

[82] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2497–2506.

[83] J. Shiraishi. "Development of a digital image database for chest radiographs with and without a lung nodule: ROC analysis on radiologists' performance in detection of pulmonary nodules". In: *Am J Roentgenol* 174.1 (2000), pp. 71–74.

[84] N. Shkumat et al. "Optimization of image acquisition techniques for dual-energy imaging of the chest". In: *Medical physics* 34.10 (2007), pp. 3904–3915.

[85] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. "CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT". In: *arXiv preprint arXiv:2004.09167* (2020).

[86] K. Sohn, H. Lee, and X. Yan. "Learning structured output representation using deep conditional generative models". In: *Advances in neural information processing systems* 28 (2015).

[87] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. "Ladder variational autoencoders". In: *Advances in neural information processing systems* 29 (2016).

[88] Y. Song, L. Shen, L. Xing, and S. Ermon. "Solving inverse problems in medical imaging with score-based generative models". In: *arXiv preprint arXiv:2111.08005* (2021).

[89] K. Suzuki, H. Abe, H. MacMahon, and K. Doi. "Image-processing technique for suppressing ribs in chest radiographs by means of massive training artificial neural network (MTANN)". In: *IEEE Transactions on medical imaging* 25.4 (2006), pp. 406–416.

[90] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. "Stanford alpaca: an instruction-following llama model (2023)". In: *URL https://crfm. stanford. edu/2023/03/13/alpaca. html*. 2023.

[91] O. Thawkar, A. Shaker, S. S. Mullappilly, et al. "Xraygpt: Chest radiographs summarization using medical vision-language models". In: *arXiv preprint arXiv:2306.07971*. 2023.

[92] E. J. Topol. *Toward the eradication of medical diagnostic errors*. 2024.

[93] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, and Lacroix. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*. 2023.

[94] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. "Towards generalist biomedical ai". In: *NEJM AI* (2024).

[95] A. Vahdat and J. Kautz. "NVAE: A deep hierarchical variational autoencoder". In: *Advances in neural information processing systems* 33 (2020), pp. 19667–19679.

[96] A. Van Den Oord, O. Vinyals, et al. "Neural discrete representation learning". In: *Advances in neural information processing systems* 30 (2017).

[97] A. Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017), pp. 5998–6008.

[98] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. "Cider: Consensus-based image description evaluation". In: *CVPR*. 2015.

[99] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. "Show and tell: A neural image caption generator". In: *CVPR*. 2015.

[100] P. Vock and Z. Szucs-Farkas. "Dual energy subtraction: principles and clinical applications". In: *European journal of radiology* 72.2 (2009), pp. 231–237.

[101] J. Wang, A. Bhalerao, and Y. He. "Cross-modal prototype driven network for radiology report generation". In: *European Conference on Computer Vision*. Springer. 2022, pp. 563–579.

[102] S. Wang, C. Wang, F. Gao, L. Su, F. Zhang, Y. Wang, and Y. Yu. "Autoregressive sequence modeling for 3d medical image representation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 8. 2025, pp. 7871–7879.

[103] T.-C. Wang et al. "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs". In: *Proceedings of the*

*IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8798–8807.

[104] X. Wang et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[105] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou. "A medical semantic-assisted transformer for radiographic report generation". In: *MICCAI*. 2022.

[106] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[107] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. "Finetuned language models are zero-shot learners". In: *arXiv preprint arXiv:2109.01652* (2021).

[108] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[109] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum. "Deep MR to CT synthesis using unpaired data". In: *International workshop on simulation and synthesis in medical imaging*. Springer. 2017, pp. 14–23.

[110] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. "Towards generalist foundation model for radiology". In: *arXiv preprint arXiv:2308.02463* (2023).

[111] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. "NExT-GPT: Any-to-Any Multimodal LLM". In: *arXiv preprint arXiv:2309.05519*. 2023.

[112] X. Wu, Y. Feng, H. Xu, Z. Lin, T. Chen, S. Li, S. Qiu, Q. Liu, Y. Ma, and S. Zhang. "CTransCNN: Combining transformer and CNN in multilabel medical image classification". In: *Knowledge-Based Systems* (2023).

[113] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[114] S. Xu, L. Yang, C. Kelly, M. Sieniek, et al. "ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders". In: *arXiv preprint arXiv:2308.01317*. 2023.

[115] Y. Xu, M. Zhou, Y. Feng, X. Xu, H. Fu, R. S. M. Goh, and Y. Liu. "Minimal-supervised medical image segmentation via vector quantization memory". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 625–636.

[116] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss". In: *IEEE transactions on medical imaging* 37.6 (2018), pp. 1348–1357.

[117] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao. "Knowledge matters: Chest radiology report generation with general and specific knowledge". In: *Medical image analysis*. 2022.

[118] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. "mplug-owl: Modularization empowers large language models with multimodality". In: *arXiv preprint arXiv:2304.14178* (2023).

[119] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, et al. "Evaluating progress in automatic chest x-ray radiology report generation". In: *Patterns* 4.9 (2023).

[120] J. Yu et al. "Vector-quantized Image Modeling with Improved VQGAN". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=pfNyExj7z2.

[121] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. "Restormer: Efficient transformer for high-resolution image restoration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5728–5739.

[122] L. Zhang, A. Rao, and M. Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 3836–3847.

[123] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (2019).

[124] Z. Zhou, L. Zhou, and K. Shen. "Dilated conditional GAN for bone suppression in chest radiographs with enforced semantic features". In: *Medical Physics* 47.12 (2020), pp. 6207–6215.

[125] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.