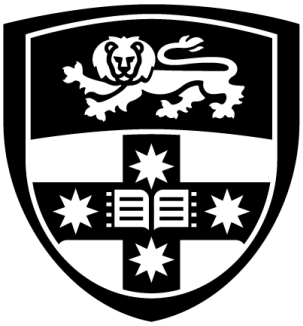


THE UNIVERSITY OF SYDNEY

MASTER THESIS

Longitudinal Chest X-ray Image Generation via Autoregression Model and Diffusion-based Model

Author:
Yiran WANG

Supervisor:
Associate Professor Luping ZHOU

Co-Supervisor:
Associate Professor Dong YUAN

A thesis submitted in fulfillment of the requirements for the degree of Master of Philosophy

in the

School of Electrical and Computer Engineering
Faculty of Engineering

2025

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signed:

Date: March 23, 2026

Attribution Statement

Chapter 4 of this thesis has been submitted to MICCAI 2026, and Chapter 5 of this thesis is to be submitted to TMI. I designed the study, analyzed the data, and wrote the manuscript. In addition to the authorship attribution 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.

In addition to the authorship attribution 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: Yiran Wang

Student Signed:

Date: March 23, 2026

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

Supervisor Name: Associate Professor Luping Zhou

Supervisor Signed:

Date: March 23, 2026

Generative AI Statement

During the preparation of the thesis, the author, Yiran WANG, used ChatGPT for purposes such as text enhancement, including paraphrasing, refining sentence structure, and correcting spelling and grammar. All AI-assisted outputs were carefully reviewed by the author to identify and correct any potential errors, inaccuracies, or biases.

The author takes full responsibility for the submitted thesis and ensures the work is their own and used generative AI within the parameters of use (refer to the University of Sydney generative AI guide for researches).

Student Signed:

Date: March 23, 2026

Abstract of thesis entitled

Longitudinal Chest X-ray Image Generation via Autoregression Model and Diffusion-based Model

Submitted by

Yiran WANG

for the degree of Master of Philosophy

at The University of Sydney

in March, 2026

Longitudinal chest X-ray (CXR) analysis is central to clinical follow-up of pulmonary diseases, where radiologists compare prior and current scans to assess subtle, localized lesion changes while preserving global thoracic anatomy. This thesis studies the problem of *longitudinal CXR generation*: synthesizing a plausible follow-up radiograph conditioned on a reference CXR and a textual description of disease progression.

Recent advances in transformer architectures have made them a dominant backbone for image generation, spanning both autoregressive image token modeling and diffusion transformer (DiT) frameworks. Despite their strong capacity for global context modeling, we observe a systematic *corner/edge bias* in transformer attention, where attention mass disproportionately drifts toward non-informative image boundaries. This behavior is particularly problematic for longitudinal CXR generation: Clinically meaningful changes are often small and spatially confined, yet the generated image must remain anatomically consistent and semantically aligned with the progression description.

To address these challenges, I propose *Gaussian-Biased Causal Attention (GBCA)*, a lightweight attention modulation module that injects lesion-centric Gaussian spatial priors into selected transformer layers to reduce attention drift toward non-informative regions (e.g., corners/edges) and to improve lesion-aligned control. GBCA is designed to be architecture-agnostic: it can be integrated into autoregressive image token generators and diffusion transformer (DiT) backbones without modifying the original model parameters. In the autoregressive setting, I integrate GBCA into a decoder-only multimodal autoregressive transformer (Emu3) for longitudinal CXR generation, and further validate its generality on a second autoregressive editing backbone (EditAR) by freezing the base generator and training only the GBCA module. In the diffusion setting, I extend GBCA to DiT-based longitudinal generation by identifying structurally critical (*vital*) layers and injecting the spatial prior into these layers to better balance global structure preservation and local lesion editing.

Extensive experiments on longitudinal CXR datasets demonstrate that GBCA consistently improves both image fidelity and clinical faithfulness. Beyond standard perceptual measures, I introduce and report lesion-aware attention and localization metrics (including Attn-IoU, Corner Bias Index, and Edge Activation Ratio) to quantify whether the model attends to clinically relevant regions. Results show that GBCA improves lesion-aligned attention, reduces corner/edge bias, and yields more anatomically consistent follow-up synthesis with better alignment to progression text. Overall, this thesis provides a practical and general mechanism for spatially grounded control in transformer-based medical image generation, enabling more reliable longitudinal CXR synthesis for follow-up modeling and clinical decision support.

Longitudinal Chest X-ray Image Generation via Autoregression Model and Diffusion-based Model

by

Yiran WANG
B.E. *North China Electric Power University*

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

at

University of Sydney
March, 2026

For my loving family, partner, and friends

Acknowledgements

With my deepest gratitude, I would like to sincerely thank everyone who has supported me throughout the entire journey of my Master of Philosophy. This thesis would not have been possible without the invaluable academic guidance, institutional support, and personal encouragement that I have received over the past years.

At the very beginning, I would like to express my sincere and deepest gratitude to my supervisors, A/Prof. Luping Zhou and A/Prof. Dong Yuan, for their invaluable guidance, continuous support, and great patience throughout my Master's study and the completion of this thesis. Their profound expertise, rigorous academic standards, and generous mentorship have guided me through every critical stage of my research. In particular, A/Prof. Zhou, with her insightful suggestions, rigorous academic attitude, and constant encouragement, has played an essential role in shaping both my research direction and my overall academic development. Her dedication to research excellence and her persistent support have been a constant source of motivation for me.

My deepest thanks also go to my beloved family and partner for their unconditional love, understanding, and unwavering support throughout my entire academic journey. Their patience, encouragement, and trust have given me the strength to overcome difficulties and pursue my goals with confidence. Without their support, this thesis would not have been possible.

I would also like to sincerely thank all my colleagues, peers, and seniors for their helpful discussions, constructive feedback, and warm support during my study. The inspiring academic environment and the spirit of collaboration I experienced have greatly enriched my research experience and broadened my academic perspective.

Furthermore, I would like to express my sincere gratitude to my friends for their companionship, encouragement, and emotional support during this important stage of my life. Their understanding and encouragement have helped me maintain balance, optimism, and perseverance throughout my postgraduate journey.

Once again, I would like to extend my heartfelt thanks to everyone who has accompanied and supported me along this unforgettable academic journey.

Yiran WANG
University of Sydney
March 23, 2026

List of Publications

JOURNALS:

[1] Xinyu Chen, **Yiran Wang**, Gaoyang Pang, Jiafu Hao, Chentao Yue, Luping Zhou, Yonghui Li. "Medical Referring Image Segmentation via Next-Token Mask Prediction", *IEEE Transactions on Medical Imaging (TMI)* (Under Review)

[2] **Yiran Wang**, and Luping Zhou. "Longitudinal Chest X-ray Generation via Diffusion-based Model", *IEEE Transactions on Medical Imaging (TMI)* (To be submitted)

CONFERENCES:

[1] **Yiran Wang**, Xiaoyu Yue, Haimei Zhao, Xinyu Chen, Luping Zhou. "Guided Longitudinal CXR Generation via Autoregression Model", *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026* (Under review)

[2] Qingcheng Lyu, Tong Chen, **Yiran Wang**, Erjian Guo, Luping Zhou. "WiD-PET: PET Image Reconstruction from Low-Dose Data Using a Wavelet-Informed Diffusion Model with Fast Inference", International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2025

Contents

Statement of Originality i

Attribution Statement iii

Generative AI Attribution Statement v

Acknowledgements xi

List of Publications xiii

List of Figures xix

List of Tables xxiii

List of Abbreviations xxv

List of Symbols xxvii

1 Introduction 1
- 1.1 Problem Statement . . . 1
- 1.2 Challenges and Motivation . . . 3
- 1.3 Thesis Outline and Contributions . . . 5

2 Literature Review 7
- 2.1 Prior Work on Medical Image Generation . . . 8
  - 2.1.1 Motivation and clinical use cases . . . 8
  - 2.1.2 Model families: from GAN/VAE to diffusion . . . 9
  - 2.1.3 Temporal and longitudinal medical image generation beyond CXR 9
  - 2.1.4 Implication for Longitudinal CXR Generation . . . 9
- 2.2 Chest X-ray Generation . . . 10
  - 2.2.1 Scalability, multimodality, and practical constraints . . . 10
  - 2.2.2 Datasets and longitudinal pair construction . . . 10
  - 2.2.3 Multimodal foundations in radiology . . . 11
  - 2.2.4 Text-conditioned CXR synthesis . . . 11
  - 2.2.5 CXR editing and counterfactual generation . . . 12
  - 2.2.6 Longitudinal CXR generation: formulations and notation . . . 12
  - 2.2.7 From Single-Time CXR Generation to Longitudinal CXR Generation 13

2.3 Longitudinal Chest X-ray Generation: Problem Setup and Challenges . . 13
2.3.1 Clinical motivation and longitudinal settings 13
2.3.2 Problem settings: forecasting, imputation, and progression editing 14
2.3.3 Key challenges in longitudinal CXR generation 14
2.3.4 Evaluation protocols and metrics for longitudinal CXR generation 14
2.3.5 Representative methods and positioning 15
2.4 Longitudinal CXR Generation via Autoregressive Models 16
2.5 Longitudinal CXR Generation via Diffusion-based Models 17
Controllability via structural conditions 17
Transformer-based diffusion backbones 17
Diffusion for longitudinal and counterfactual CXR generation . . 18
Limitations and connection to lesion-centric priors 18
2.6 Attention Manipulation and Spatial Prior Injection 18
2.6.1 Inference-Time Attention Manipulation 18
2.6.2 Layout and Coordinate-Guided Generation 19
2.6.3 Position-Dependent Attention Biases 19
2.7 Summary and Motivation 20

3 Background Knowledge **23**
3.1 Generative Models in Medical Imaging 23
3.2 Autoregressive Generative Models 24
3.2.1 Sequence Modeling and Next-Token Prediction 25
3.2.2 Transformers and Causal Self-Attention 25
3.3 Diffusion Probabilistic Models 26
3.3.1 Forward and Reverse Processes 26
The Forward Process (Diffusion) 26
The Reverse Process (Denoising) 27
Training Objective 27
3.3.2 Practical advances: fast sampling, guidance, and flow-matching . 27
3.3.3 Conditioning mechanisms for controllable diffusion 27
3.3.4 Latent Diffusion and DiT Architectures 28
Latent Diffusion Models (LDMs) 28
Diffusion Transformers (DiT) 28
3.4 Vision-Language Models (VLMs) as Spatial Priors 29
3.4.1 The Evolution from Contrastive to Generative VLMs 29
Contrastive Dual-Encoder Models 30
Generative Multimodal Large Language Models (MLLMs) 30
3.4.2 Visual Grounding and Coordinate Prediction 30
3.4.3 From coordinates to Gaussian spatial priors 30
3.4.4 VLMs in Medical Imaging 31
3.5 Summary 31

4 Longitudinal Chest X-ray Generation via Autoregression Model **33**

4.1 Motivations and Contributions . . . 33
4.2 Methodology . . . 34
4.2.1 Layer-wise Attention Bias Profiling . . . 35
4.2.2 2D Gaussian Spatial Prior from VLM . . . 39
4.2.3 Gaussian-Biased Causal Attention . . . 39
4.2.4 Training and Inference Pipelines . . . 40
4.3 Experiments . . . 41
4.3.1 Dataset . . . 41
4.3.2 Evaluation Metrics . . . 41
4.3.3 Implementation Details . . . 43
4.3.4 Results . . . 43
4.3.5 Ablation Study . . . 44
4.4 Generality of GBCA Across Autoregressive Backbones . . . 47
4.5 Visualization of Diverse VLM-Predicted Spatial Priors . . . 48
4.6 Prompts Design for VLM Point Annotation . . . 48

5 Longitudinal Chest X-ray Generation via Diffusion-based Model 53
5.1 Motivations and Contributions . . . 53
5.2 Methodology . . . 53
5.2.1 Phase 1: Offline Structural Calibration (Vital Layer ID) . . . 54
5.2.2 Phase 2: Semantic-to-Spatial Mapping (Gaussian Bias) . . . 55
Coordinate Prediction via VLM . . . 56
Latent Space Transformation and Gaussian Modeling . . . 56
Sequence Alignment and Bias Injection . . . 57
5.2.3 Phase 3: Dual-Path Guided Generation (Inference) . . . 58
Latent Initialization via DDIM Inversion . . . 58
Synchronized Denoising Trajectories . . . 58
Feature Injection and Attention Modulation . . . 59
5.3 Experiments . . . 60
5.3.1 Dataset and Implementation Details . . . 60
5.3.2 Evaluation Metrics . . . 61
5.3.3 Quantitative Results . . . 61
5.3.4 Ablation Study . . . 62
Quantitative Component Analysis . . . 62
Visual Inspection and Artifact Analysis . . . 63

6 Conclusion and Future Work 67
6.1 Conclusion . . . 67
6.2 Future Work . . . 68

Bibliography 71

List of Figures

1.1 The Overview of the Task. The model takes an initial reference image and a text prompt describing the clinical evolution as inputs. Through a generative model, it synthesizes a corresponding follow-up image that visually demonstrates the radiographic features, such as the improved right pleural effusion and bilateral pulmonary edema. 3

1.2 (a) The reference image. The boxes correspond to the area specified in the prompt. (b) The Gaussian spatial prior overlaps with the reference image. (c) The attention map without additive bias in the last layer of the model. (d) Additive bias helps the model focus on lesion areas. (e) The generated follow-up image. 4

2.1 Overview of the related-work structure in Chapter 2 Literature Review. . 8

4.1 Inference-time overview of LeGend. Given a reference CXR I_A and a progression description D, a vision–language model (VLM) predicts sparse lesion coordinates on the reference image. These coordinates are converted into a 2D Gaussian prior, which is injected as an additive bias into the decoder's causal self-attention logits (GBCA), steering attention toward clinically relevant reference regions while preserving causality. The decoder then autoregressively predicts the follow-up token sequence $\hat{X}_B$; after each step, the newly generated token is fed back as part of the visible prefix for subsequent decoding, and the final sequence is detokenized into the generated follow-up image $\hat{I}_B$. Ground-truth follow-up tokens X_B are not used in this inference pipeline. 35

4.2 The layer-wise analysis of causal attention distribution with Attn-IoU, CBI, and EAR for the vanilla autoregressive model. 36

4.3 Teacher-forcing setup used for attention-bias profiling. The decoder input is the concatenated sequence $\{X_A, X_D, X_B\}$, where X_A denotes reference-image tokens, X_D denotes progression-text tokens, and X_B denotes the ground-truth follow-up tokens. When analyzing the query corresponding to the n-th target token in X_B, the causal mask allows attention only to X_A, X_D, and the preceding ground-truth prefix $X_{B,<n}$, while future follow-up tokens remain invisible. The heatmap is obtained by aggregating attention from follow-up queries to reference-image keys. 37

4.4 Visual comparison of follow-up CXR generation. Our LeGend shows higher similarity with regard to the ground truth image. 41
4.5 Attention profiling across layers highlights the mid-layer region where GBCA most effectively improves lesion focus. 42
4.6 Ablation of different VLMs for spatial-prior generation. This comparison evaluates the robustness of GBCA across alternative VLM-derived spatial priors. 47
4.7 Layer-wise attention metrics of EditAR under the *no_gaussian* setting. We report Attn-IoU, CBI, and EAR across transformer layers. The marked EAR peak around the 18th layer indicates a strong edge-attention tendency, motivating GBCA injection at this middle layer. 48
4.8 Visualization of spatial priors predicted by two VLMs (Qwen2.5-VL and Llama-3.2-Vision). The two models produce different lesion-coordinate sets, illustrating the natural variability in VLM-based localization. The corresponding quantitative results in Table 4.6 evaluate the robustness of GBCA across these alternative VLM-derived priors. 49

5.1 **Overview of the proposed LeGend-Diffusion framework.** The pipeline consists of three phases: **Phase 1** identifies Vital Layers ($\mathcal{V}$) offline via ablation analysis to lock anatomical structures. **Phase 2** leverages a Vision-Language Model (VLM) to map the progression description D_P into a Gaussian spatial prior B_{gaussian}. **Phase 3** performs dual-path inference, where the Edit Path generates the follow-up image I_B by retrieving structural keys/values (K_{src}, V_{src}) from the Source Path and integrating the Gaussian bias (B_{gaussian}) specifically at the identified Vital Layers. 54
5.2 Workflow for Offline Structural Calibration (Phase 1). 56
5.3 **Visualization of Semantic-to-Spatial Mapping and Generation Results.** This figure illustrates the intermediate spatial guidance signals and the final output. **(1) Reference Image:** The baseline CXR input (I_A). **(2) Points on Reference Image:** Discrete lesion coordinates (red circles) predicted by the VLM based on the progression description, identifying the anatomical ROI (e.g., left lower lobe). **(3) Gaussian Map:** The continuous Gaussian spatial prior (B_{gaussian}) derived from the discrete points, which serves as the attention bias B_{gaussian} injected into the DiT. **(4) Ours:** The final follow-up image generated by our full framework (LeGend-Diffusion), showing precise lesion synthesis. **(5) Ground Truth:** The real follow-up exam. Comparing the Gaussian Map with the generated outcome confirms that our GBCA mechanism effectively translates semantic coordinates into accurate visual pathology. 57

5.4 **Qualitative comparison of longitudinal CXR generation.** The top row displays results from baseline methods, which exhibit various degrees of structural distortion or blurring. The bottom row compares our previous AR model (LeGend), our proposed method, and the Ground Truth. Our Diffusion method (LeGend-Diffusion) achieves the best trade-off, preserving the precise anatomical structure of the reference image while faithfully generating the progression lesion. 63
5.5 Visual ablation of framework components. The text prompt describes progression in the left lung. (a) Base DiT introduces unnecessary changes in non-target upper-chest content while also failing to synthesize the requested pathology faithfully. (b) + Vital Layer improves preservation of the overall image context, but suppresses the requested pathological change. (c) + Vital Layer + GBCA (LeGend-Diffusion) successfully generates the lesion progression while keeping non-target regions largely stable, closely matching (d) the Ground Truth. 64

List of Tables

1.1 Compact comparison of thoracic imaging modalities for longitudinal monitoring. 2

2.1 Four-column taxonomy of representative medical image generation models (CXR-centric), categorized by model family, modality, and task. . . . 13

3.1 Representative task-level works for longitudinal CXR generation. 24
3.2 Common building blocks for AR/diffusion/flow-based image generation. 24
3.3 VLMs and whether they provide explicit visual grounding outputs. . . . 29

4.1 Performance comparison of image generation quality and downstream classification. 41
4.2 Last-layer attention statistics under different GBCA injection locations. . 43
4.3 Performance across different GBCA injection locations. 44
4.4 Effect of injecting GBCA into different mid-depth layer combinations. . 46
4.5 The ablation of Gaussian σ's ratio. 46
4.6 Ablation of different VLMs for spatial-prior generation. 46
4.7 Performance comparison of EditAR with/without GBCA. 48

5.1 Performance comparison of image generation quality and downstream classification. Best results are highlighted in **bold**. 62
5.2 Ablation study on the two core components. **Trend Analysis:** The Baseline suffers from structural collapse. Vital Layer injection recovers structure but limits lesion synthesis due to reference suppression. The addition of GBCA achieves the best trade-off, maximizing both structural fidelity and clinical accuracy. 64

List of Abbreviations

AR	Autoregressive
AUC	Area Under the ROC Curve
CBI	Corner Bias Index
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-ray
DDIM	Denoising Diffusion Implicit Models
DDPM	Denoising Diffusion Probabilistic Models
DiT	Diffusion Transformer
EAR	Edge Attention Ratio
FID	Fréchet Inception Distance
FiLM	Feature-wise Linear Modulation
GAN	Generative Adversarial Network
GBCA	Gaussian-Biased Causal Attention
ICU	Intensive Care Unit
IoU	Intersection over Union
LDM	Latent Diffusion Model
LLM	Large Language Model
MLLM	Multimodal Large Language Model
MRI	Magnetic Resonance Imaging
MS-SSIM	Multi-Scale Structural Similarity
MSE	Mean Squared Error
PET	Positron Emission Tomography
PSNR	Peak Signal-to-Noise Ratio
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SPECT	Single-Photon Emission Computed Tomography
SSIM	Structural Similarity
TB	Tuberculosis
U-Net	U-shaped Convolutional Network
US	Ultrasound
VAE	Variational Autoencoder
VLM	Vision-Language Model

List of Symbols

Global notations

x_{ref}	Reference image / study in a longitudinal pair
x_{fu}	Target follow-up image / study in a longitudinal pair
Δ	Follow-up interval between the reference and target follow-up studies
y	Generic conditioning signal (e.g., text, structured attributes, or spatial priors)
H, W	Spatial height and width of an image or feature grid
θ	Learnable parameters of a neural network / generative model
$\mathbf{I}$	Identity matrix
σ	Standard deviation controlling the spatial spread of a Gaussian kernel

Chapter 2 symbols

$x_{\text{obs},i}$	The i-th observed study in an irregular longitudinal sequence
$p_\theta(\cdot)$	Conditional data distribution parameterized by θ

Chapter 3 symbols

$\mathbf{x} = (x_1, \ldots, x_N)$	Discrete token sequence representing an image or multimodal input
$X_{B,<n}$	Ground-truth follow-up prefix preceding position n
$\hat{X}_{B,<n}$	Generated follow-up prefix preceding position n at inference
N	Sequence length in autoregressive modeling
$\mathcal{L}_{AR}$	Autoregressive log-likelihood objective
Q, K, V	Query, key, and value matrices in attention
d_k	Dimensionality of the key vectors in scaled dot-product attention
M	Causal / attention mask
$q(\cdot)$	Forward diffusion process
x_τ	Noisy sample at diffusion timestep τ
τ	Diffusion timestep
K	Total number of diffusion steps
β_τ	Variance schedule at diffusion timestep τ
α_τ	Noise-retention coefficient, defined as $1 - \beta_\tau$
$\bar{\alpha}_\tau$	Cumulative product $\prod_{s=1}^{\tau} \alpha_s$
ϵ	Gaussian noise sample

$\epsilon_\theta(x_\tau, \tau)$	Predicted noise at diffusion timestep τ
$\mu_\theta(x_\tau, \tau)$	Predicted reverse-process mean
$\Sigma_\theta(x_\tau, \tau)$	Predicted / fixed reverse-process variance
$\mathcal{L}_{simple}$	Simplified diffusion denoising loss
$\mathcal{E}, \mathcal{D}$	Encoder and decoder of a latent diffusion model
z	Latent representation of an image
z_τ	Noisy latent representation at diffusion timestep τ
C	Channel dimension of the latent representation
$\mathcal{L}_{LDM}$	Latent diffusion training objective
$\gamma(\tau, y), \beta(\tau, y)$	Scale and shift parameters in adaptive LayerNorm conditioning
$\mathbf{b}^{\text{box}}$	Bounding box predicted by a VLM
μ	Center of a predicted bounding box
$G(u, v)$	Continuous 2D Gaussian prior defined over spatial coordinates (u, v)

Chapter 4 symbols

I_A	Reference chest X-ray image in the autoregressive framework
$\hat{I}_B$	Generated follow-up chest X-ray image in the autoregressive framework
D	Progression description / textual conditioning signal
X_A	Reference-image token sequence
X_D	Text-token sequence derived from the progression description
X_B	Ground-truth target follow-up token sequence used in teacher forcing and attention analysis
$\hat{X}_B$	Autoregressively generated follow-up token sequence at inference
S	Total input sequence length in the autoregressive decoder
$S^{(\ell)}$	Pre-softmax attention logits at decoder layer ℓ
Q_B	Index set of target follow-up token positions in X_B
K_A	Index set of reference-image token positions in X_A
$S^{(\ell)}_{B\to A}$	Attention-logit submatrix from follow-up queries to reference-image keys at layer ℓ
$m^{(\ell)}_{B\to A}(k)$	Average logit from target follow-up queries to the k-th reference token at layer ℓ
ϕ_A	Mapping from a reference-image token index to its 2D spatial location
$\mathcal{H}^{(\ell)}(i, j)$	Layer-wise spatial attention heatmap on the reference token grid
$\mathcal{C}$	Corner region used in the Corner Bias Index (CBI) computation
$\mathcal{E}$	Edge band used in the Edge Attention Ratio (EAR) computation
Ω	Binary lesion mask / lesion support region
$\mathbf{p}_k = (x_k, y_k)$	VLM-predicted lesion-relevant coordinate in the original pixel space
$(\tilde{x}_k, \tilde{y}_k)$	Projected lesion coordinate on the autoregressive visual-token lattice

N_p	Number of VLM-predicted lesion points
$R_{\text{tok}}(i,j)$	Gaussian prior map defined on the autoregressive reference-token grid
$\mathbf{r}$	Flattened token-grid Gaussian prior vector
B_{gaussian}	Gaussian bias tensor injected into autoregressive attention logits
s	Learnable scalar controlling the strength of Gaussian bias injection
$\mathcal{L}_{\text{CE}}$	Next-token cross-entropy loss for autoregressive training

Chapter 5 symbols

$\mathcal{F}$	Full pre-trained DiT / FLUX backbone
$\mathcal{F}_{\text{skip-}l}$	Modified model with the l-th layer skipped
z_i	Initial Gaussian noise for the i-th sample in vitality analysis
p_i	Text prompt for the i-th sample in vitality analysis
I_i	Image generated by the full model for the i-th sample
$I_i^{(l)}$	Image generated when layer l is skipped for the i-th sample
$\Phi(\cdot)$	Frozen DINOv2 feature extractor
$\text{Sim}(U,V)$	Cosine similarity between DINOv2 features of images U and V
$v(l)$	Vitality score of layer l
τ_{vit}	Threshold for defining vital layers
$\mathcal{V}$	Set of vital layers selected for feature injection
D_A	Source text conditioning the source path
D_P	Progression text conditioning the edit path
$\mathbf{p}_k = (x_k, y_k)$	VLM-predicted coordinate in the original pixel space
N_p	The number of VLM-predicted coordinates in the original pixel space
$(x_k^{\text{lat}}, y_k^{\text{lat}})$	Projected lesion coordinate on the latent spatial grid
h, w	Spatial height and width of the latent grid
f	Downsampling factor from image space to latent space
$R_{\text{lat}}(i,j)$	Gaussian prior map defined on the latent spatial grid
$\mathbf{r}_{\text{img}}$	Flattened latent-image bias vector
$\mathbf{r}_{\text{seq}}$	Sequence-aligned bias vector after zero-padding text positions
N_{txt}	Number of text tokens in the DiT input sequence
N_{img}	Number of image / latent tokens in the DiT input sequence
L_{seq}	Total DiT input sequence length
$\mathcal{T}_{\text{src}}$	Source denoising path
$\mathcal{T}_{\text{edit}}$	Edit denoising path
$z_K^{(\text{src})}, z_K^{(\text{edit})}$	Shared inverted terminal latent for the source and edit paths
$h_\tau^{(\text{src},l)}$	Hidden state of the source path at diffusion timestep τ and layer l
$h_\tau^{(\text{edit},l)}$	Hidden state of the edit path at diffusion timestep τ and layer l
$\text{Attn}_{\text{edit}}^{(l)}$	Edit-path attention output at layer l
$Q_{\text{edit}}^{(l)}$	Query matrix of the edit path at layer l
$K_{\text{src}}^{(l)}, V_{\text{src}}^{(l)}$	Key and value matrices of the source path at layer l

$s(\tau)$	Diffusion-step-dependent scalar controlling Gaussian guidance strength

Chapter 1

Introduction

This chapter first introduces the importance of longitudinal chest X-rays (CXR) in disease diagnosis and treatment monitoring, and highlights the key problems and challenges faced in applying generative models to assist with follow-up image synthesis. It then analyzes the structural attention bias in mainstream generative models (including autoregressive models and diffusion models with Transformer backbones), and elaborates on the research motivation—namely, proposing a Gaussian-biased attention mechanism to explicitly guide the model's focus toward lesion regions and improve the clinical credibility of generated follow-up images.

1.1 Problem Statement

Thoracic imaging modalities for diagnosis and follow-up. Clinical thoracic assessment relies on multiple imaging modalities, each with different trade-offs in cost, accessibility, radiation exposure, and sensitivity to subtle pathology. Table 1.1 summarizes the key strengths and limitations from a longitudinal monitoring perspective.

Why chest radiography (CXR) is particularly suitable for longitudinal monitoring. Among these modalities, CXR is often the frontline exam for a broad spectrum of respiratory complaints because it is rapid, widely available (including portable imaging in wards/ICU), and relatively low-cost, making repeated follow-up feasible in routine clinical workflows. Clinical guidelines and appropriateness criteria commonly recommend CXR as the initial imaging test, while escalating to chest CT when radiographs are negative/inconclusive or when complications are suspected. [33] In tuberculosis (TB) programs, chest radiography is also recommended as an essential tool for detection/screening and is frequently used in longitudinal assessment across treatment episodes. [48]

Clinical importance of longitudinal CXR comparison. For conditions such as pneumonia and tuberculosis, patients often undergo serial CXR examinations during treatment. Clinicians compare the latest radiograph with prior scans to identify changes such as the emergence, enlargement, or resolution of lesions, which directly informs treatment

Table 1.1: Compact comparison of thoracic imaging modalities for longitudinal monitoring.

Modality	Strengths	Limitations	Dose (typ.)
CXR	Fast, ubiquitous, low-cost; portable bedside imaging; feasible for frequent follow-up.	2D projection with anatomical overlap; limited sensitivity for subtle lesions.	∼0.1 mSv
CT	High sensitivity; detailed lesion characterization; mitigates overlap.	Higher cost and radiation; less suitable for very frequent routine follow-up.	∼6 mSv
MRI	No ionizing radiation; strong soft-tissue contrast.	Less standard for routine lung surveillance; practical constraints.	0
US	Portable and radiation-free; useful for pleura/peripheral findings.	Operator-dependent; limited by aerated lung; incomplete deep parenchyma view.	0
PET/CT	Functional assessment; oncology-specific value.	High dose/cost; not routine for frequent follow-up.	varies (high)

Note: Doses are representative adult effective doses; protocols vary.

decisions and prognosis assessment. However, manual multi-timepoint comparison is time-consuming and subjective, increasing radiologist workload and inter-reader variability. Therefore, leveraging artificial intelligence to automatically analyze follow-up changes—and even generate predictive future CXRs—has become a practical and clinically meaningful direction in intelligent medical imaging.

In recent years, various medical image generative models, such as GAN-based methods like PIE [40], diffusion-based approaches like CXR-IRGen [57] and BioMedJourney [24], have been applied to simulate disease progression scenarios to support clinical decisions. However, current approaches exhibit a major limitation: insufficient attention to lesion regions and lack of fine-grained depiction. Many generic image generation models optimize overall visual realism and can reproduce normal structures and backgrounds with high fidelity, yet they often fail to accurately replicate lesion morphology or dynamics. In some cases, they even hallucinate anatomical structures or distort clinical content [24, 57]. This is a severe shortcoming in medical imaging, as subtle lesion changes often carry significant clinical meaning. If generated follow-up images fail to highlight lesion evolution, the practical diagnostic value of these models is greatly diminished.

In this thesis, we formulate the task as *progression-conditioned follow-up CXR synthesis*: given a reference CXR image at an earlier time point and a textual description detailing the expected disease progression, the goal is to synthesize a plausible future follow-up CXR that (1) conforms precisely to the semantics of the progression text, including lesion location, extent, and severity; (2) retains the anatomical layout and structural characteristics of the reference image; and (3) exhibits high visual fidelity without introducing clinically implausible artifacts. Importantly, although the training data are longitudinal reference/follow-up pairs, the elapsed follow-up interval Δ between the two studies is *not explicitly encoded* as a model input in Chapters 4 and 5. Instead,

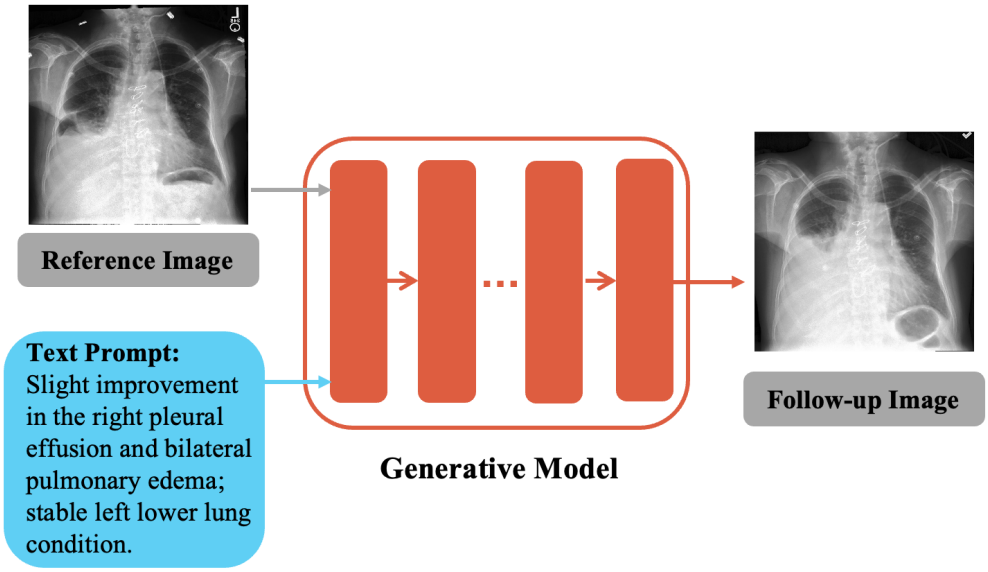


Figure 1.1: The Overview of the Task. The model takes an initial reference image and a text prompt describing the clinical evolution as inputs. Through a generative model, it synthesizes a corresponding follow-up image that visually demonstrates the radiographic features, such as the improved right pleural effusion and bilateral pulmonary edema.

temporal information is conveyed only implicitly through the progression description. Therefore, the present thesis should be interpreted as addressing *progression-conditioned follow-up synthesis* rather than fully *time-specific longitudinal forecasting*. Explicit interval-aware prediction—for example, answering clinically specific queries such as "how will the disease appear after two months?"—is left for future work.

1.2 Challenges and Motivation

Thesis statement. This thesis improves the clinical credibility of longitudinal CXR synthesis by mitigating lesion-under-attention in transformer-based generative models, via a lesion-conditioned Gaussian attention bias that steers generation toward pathology while preserving global thoracic anatomy.

One fundamental reason for the above limitation is the structural attention bias in existing generative models. In particular, autoregressive generation models—which synthesize images pixel-by-pixel or token-by-token—tend to exhibit biased attention patterns. For example, autoregressive Transformers generate tokens sequentially, and thus often place excessive attention on corner and edge of the image (See Figure 1.2), leading to spatial artifacts or fixed-mode patterns. This "corner bias" manifests as overly concentrated attention on peripheral areas at early generation stages, while clinically relevant central lesion regions receive insufficient attention. Combined with the causal nature of attention—where each token only attends to previously generated

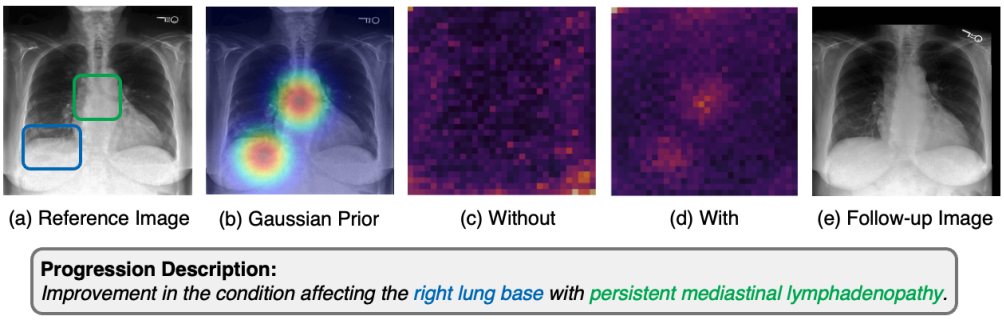


Figure 1.2: (a) The reference image. The boxes correspond to the area specified in the prompt. (b) The Gaussian spatial prior overlaps with the reference image. (c) The attention map without additive bias in the last layer of the model. (d) Additive bias helps the model focus on lesion areas. (e) The generated follow-up image.

tokens—these models struggle to naturally attend to local lesion areas without explicit guidance. Such bias impairs local fidelity in generated images and weakens the model's ability to represent lesion-specific details [43, 66].

Furthermore, similar challenges appear in diffusion-based generative models, especially those that adopt Transformer-based backbones. Recently proposed models such as Diffusion Transformer (DiT) [50] replace traditional U-Net architectures in diffusion models with global self-attention modules, significantly enhancing their representational power and generation quality. However, due to architectural similarities with autoregressive Transformers, these models also inherit similar attention distribution issues—namely, the lack of explicit focus on lesion-critical regions.

To address this challenge, we propose a unified solution by introducing the attention bias correction technique originally developed for autoregressive models into diffusion models like DiT. This approach aims to improve the spatial alignment and lesion-awareness of both model types during medical image synthesis.

The core motivation of this study is to introduce a Gaussian-biased causal attention mechanism, which guides generative models to focus more effectively on lesion regions. Unlike classical position bias techniques such as ALiBi [52] and T5's learned positional embeddings [54], GBCA introduces a dynamic, lesion-conditioned spatial prior. Unlike post-hoc diffusion editing techniques such as Attend-and-Excite [11], GBCA integrates lesion guidance during generation without breaking autoregressive consistency. Specifically, we inject a 2D Gaussian bias—centered on known lesion coordinates—into the causal self-attention logits during generation. This spatial bias increases the attention weights for regions near lesions at each generation step. The mechanism explicitly steers attention toward clinically relevant regions, correcting for spatial imbalance. Prior research has shown that incorporating distance-aware attention bias can enhance local structure modeling in sequence generation tasks. In the context of longitudinal CXR synthesis, we expect that Gaussian-biased attention will help the model capture lesion evolution with higher accuracy and generate clinically interpretable follow-up

images. This technique contributes to the advancement of medical image generation by improving spatial precision and clinical fidelity in synthesized radiographs.

Clinical faithfulness and privacy as deployment constraints. For medical image generation, clinical validity is often determined by subtle local structures; therefore, perceptual realism alone (e.g., FID) can be insufficient, and task-aware evaluation (structure similarity, downstream performance, and expert review) is typically required. [58] In addition, synthetic images do not automatically guarantee privacy. Prior studies show that diffusion models can memorize training samples under certain regimes, creating potential data-extraction and re-identification risks. [7, 16] These constraints motivate model designs that improve lesion-grounded generation (reducing spurious attention patterns) while maintaining rigorous evaluation and privacy-aware release practices.

1.3 Thesis Outline and Contributions

This thesis systematically addresses the challenges of longitudinal medical image synthesis by identifying structural attention deficiencies in existing generative models and proposing a unified, lesion-guided attention mechanism. The primary contributions and the organizational structure of this work are outlined below:

1. Summary of Main Contributions

- **Diagnosing peripheral-attention failure in medical image generation.** We identify and quantify a recurrent failure mode where transformer-based generators over-attend to image corners/edges while under-attending to clinically relevant lesion regions. We propose two simple diagnostics, the Corner Bias Index (CBI) and Edge Attention Ratio (EAR), to measure this effect.
- **A lesion-conditioned attention prior for faithful progression synthesis.** We introduce a lesion-guided attention modulation mechanism (Gaussian-Biased Causal Attention, GBCA) that injects a 2D Gaussian spatial prior into attention logits to explicitly increase focus on pathology regions derived from progression descriptions.
- **An autoregressive instantiation for longitudinal CXR generation.** We build an autoregressive follow-up generator (LeGend) and show that applying GBCA at the semantic-formation stage of the transformer substantially improves lesion alignment and reduces peripheral artifacts, without sacrificing global anatomical consistency.
- **A diffusion-transformer instantiation with structure-preserving guidance.** We further adapt GBCA to diffusion models with transformer backbones by integrating it into a structure-preserving inference pipeline (dual-path inference and vital-layer structure locking inspired by Stable Flow [1]). This enables localized pathological editing while keeping patient-specific anatomy stable.

2. Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2: Literature Review** surveys medical image generation with a deliberate progression from general to specific. It first reviews key model families and clinical use cases across medical modalities, then focuses on CXR generation by discussing scalability constraints, dataset construction for longitudinal pairing, and multimodal radiology foundations. Building on this, it formalizes longitudinal CXR generation (forecasting, imputation, and progression editing), summarizes task-specific evaluation protocols, and reviews representative longitudinal solutions under two paradigms: autoregressive token generation and diffusion-based synthesis (including transformer backbones). Finally, it connects recent advances in attention manipulation and spatial prior injection to motivate our lesion-centric Gaussian prior and the proposed Gaussian-Biased Causal Attention (GBCA).
- **Chapter 3: Background Knowledge** establishes the theoretical foundations underpinning this research. It details the mathematical formulation of Autoregressive Next-Token Prediction and Denoising Diffusion Probabilistic Models (DDPM), with a specific focus on the isotropic architecture of Diffusion Transformers (DiT). Additionally, it introduces Vision-Language Models (VLMs) and their role in Visual Grounding, which serves as the core engine for our semantic-to-spatial mapping strategy.
- **Chapter 4: Longitudinal Chest X-ray Generation via Autoregression Model** presents the **LeGend** framework. This chapter details the diagnostic profiling of the "corner-edge attention bias" using the proposed CBI and EAR metrics. It elaborates on the implementation of the GBCA mechanism within a decoder-only Transformer and presents comprehensive experiments demonstrating that mid-layer bias injection significantly improves lesion alignment (Attn-IoU) and clinical interpretability compared to standard fine-tuning approaches.
- **Chapter 5: Longitudinal Chest X-ray Generation via Diffusion-based Model** introduces the **LeGend-Diffusion** framework. It describes a three-phase methodology designed to overcome the structural instability of DiT architectures: (1) **Offline Structural Calibration** to identify "Vital Layers"; (2) **Semantic-to-Spatial Mapping** via VLM-predicted coordinates; and (3) **Dual-Path Guided Generation**. In this final phase, we describe how we adapt the Stable Flow [1] inference strategy by incorporating GBCA into the dual-path pipeline, ensuring that pathological changes are accurately synthesized within the locked Vital Layers.
- **Chapter 6: Conclusion and Future Work** summarizes the key findings of this thesis and discusses the broader implications of the GBCA mechanism. It outlines potential future research directions, including the extension of this framework to 3D medical modalities and its integration into real-time clinical decision-support workflows.

Chapter 2

Literature Review

This chapter reviews prior work on medical image generation with a focus on chest X-ray (CXR) and its longitudinal setting. We first summarize major generative model families in medical imaging and discuss the unique multimodal properties of CXR in radiology workflows. We then formulate longitudinal CXR generation settings and challenges, and review two representative generative paradigms: autoregressive token generation and diffusion-based synthesis (including Transformer backbones). Finally, we summarize related work on attention manipulation and spatial prior injection, which motivates our Gaussian-Biased Causal Attention (GBCA) mechanism. Figure 2.1 provides a visual roadmap of the related work reviewed in this chapter, organizing prior studies from general medical image generation to longitudinal CXR generation, and positioning our GBCA framework within both autoregressive and diffusion-based paradigms.

Clinical imaging spans heterogeneous modalities with distinct physics, dimensionality, and clinical roles. Anatomical imaging (e.g., radiography and CT) primarily depicts tissue density and morphology, whereas functional imaging (e.g., PET/SPECT) reflects metabolic or perfusion signals and is often interpreted jointly with anatomical context. Modalities also differ in data structure: radiography is a 2D projection of 3D anatomy, while CT/MRI are volumetric (3D) and may extend to 4D with time or contrast phases; ultrasound is operator-dependent with real-time acquisition; ophthalmic imaging (fundus/OCT) and digital pathology have their own characteristic textures and scale.

These differences directly affect generative modeling. Projection images (CXR) require preserving global anatomical layout under superposition, while volumes (CT/MRI) demand 3D consistency and substantially higher computational budgets. Functional imaging further introduces cross-modality alignment and calibration issues. Therefore, a practical strategy in this thesis is to start from broad medical image generation, narrow down to chest imaging where longitudinal monitoring is routine, and finally focus on longitudinal CXR generation with two representative generative paradigms: autoregressive token generation and diffusion-based synthesis.

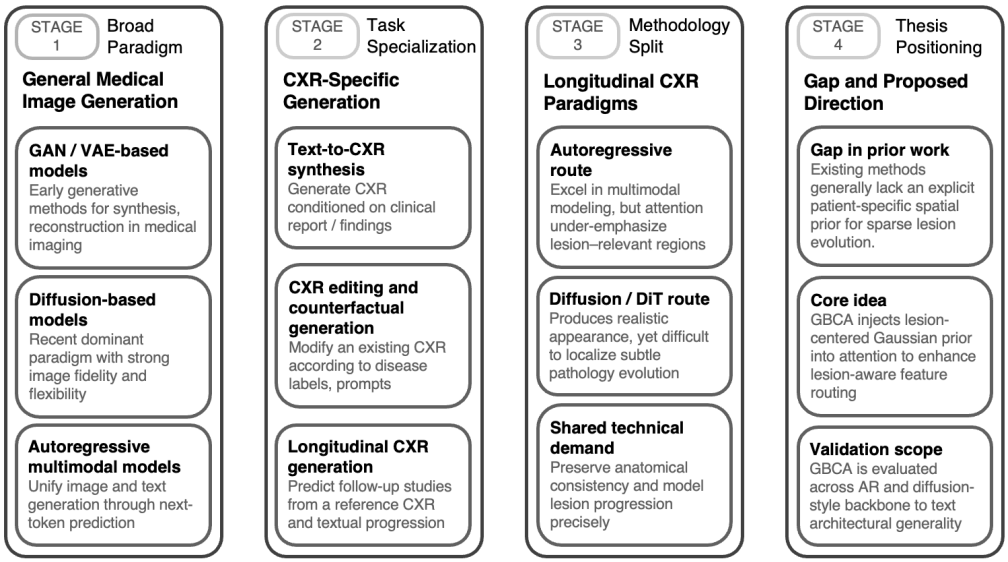


Figure 2.1: Overview of the related-work structure in Chapter 2 Literature Review.

2.1 Prior Work on Medical Image Generation

2.1.1 Motivation and clinical use cases

Medical image generation has been studied across a wide spectrum of modalities (radiography, CT, MRI, ultrasound, PET/SPECT, ophthalmic imaging, and digital pathology), primarily in response to several persistent constraints in clinical imaging: (i) limited and imbalanced labeled cohorts, especially for rare diseases and long-tail findings; (ii) acquisition cost, safety, and operational constraints (e.g., radiation exposure and scanner availability); (iii) missing or unavailable modalities in real-world care (e.g., MR-only radiotherapy planning or incomplete multi-sequence MRI); and (iv) privacy barriers to data sharing and federated development. As summarized in influential reviews, generative modeling has therefore been used for *single-modality synthesis* (data augmentation, denoising, super-resolution, artifact reduction) and *cross-modality synthesis* (image-to-image translation such as MR→CT or MR→PET), often with explicit goals of preserving anatomy while synthesizing modality-specific contrast [73, 17].

Importantly, the clinical value of generation is modality- and task-dependent. For volumetric imaging (CT/MRI), faithful 3D consistency and scanner/protocol variability are major constraints; for functional imaging (PET/SPECT), cross-modality alignment and calibration become critical; and for projection radiography (CXR), the dominant requirement is preserving global anatomical layout under superposition while allowing subtle, localized changes. Representative cross-modality work demonstrates that conditional adversarial learning can translate between modalities while encouraging realistic target-domain appearance and sharpness [46]. These modality-specific requirements motivate the thesis strategy adopted here: starting from broad medical image generation, then narrowing to chest imaging where follow-up monitoring is routine,

and finally focusing on longitudinal CXR generation where subtle disease evolution must be modeled under strong anatomical preservation.

2.1.2 Model families: from GAN/VAE to diffusion

Historically, GANs and VAEs were widely adopted due to their single-pass sampling and flexible conditioning mechanisms. However, in medical imaging they exhibit well-known limitations, including mode collapse, training instability, and the difficulty of calibrating the fidelity–diversity trade-off for rare or heterogeneous pathologies [73]. Large-scale empirical evaluations further caution that visually convincing synthetic images do not necessarily translate into improved downstream performance; the benefit depends strongly on modality, task definition, and evaluation protocol [58].

In recent years, diffusion models have emerged as a major backbone for medical image generation, largely due to their stable optimization behavior and strong mode coverage, often yielding higher-fidelity and more diverse samples (at the cost of iterative sampling) [37]. Kazerouni *et al.* [37] provide a comprehensive survey of diffusion variants across medical synthesis, translation, editing, and reconstruction. In parallel, diffusion has also been reviewed specifically for medical image reconstruction and inverse imaging settings, highlighting its suitability as a conditional generative prior under physics- or acquisition-constrained measurements [68].

2.1.3 Temporal and longitudinal medical image generation beyond CXR

While early medical generative models largely focused on static images, multiple recent works explicitly target *longitudinal* settings, where the goal is to model patient-specific temporal evolution, impute missing follow-up scans, or forecast future states. For example, diffusion-based frameworks have been proposed for longitudinal MRI imputation and completion by leveraging adjacent time points and enforcing temporal consistency [77, 61]. More broadly, conditional diffusion formulations have also been developed to handle practical longitudinal challenges such as irregular follow-up intervals and heterogeneous imaging formats, e.g., conditional latent diffusion [44] with temporal fusion for irregularly spaced radiological data, as well as diffusion-based forecasting in ophthalmology fundus sequences with explicit continuous-time modules and population-memory retrieval [19]. These works collectively suggest that temporal generation requires (i) mechanisms to preserve stable anatomy, (ii) explicit modeling of time gaps and irregular sampling, and (iii) inductive biases that focus modeling capacity on clinically meaningful change regions rather than background.

2.1.4 Implication for Longitudinal CXR Generation

Prior work across modalities suggests that longitudinal generation must simultaneously address two requirements: (1) preserving patient-specific anatomy over time, and (2)

modeling temporally evolving patterns that are often spatially sparse. These observations motivate longitudinal formulations that explicitly represent follow-up intervals and progression cues, which we introduce in Section 2.3. We defer the discussion of lesion-aware inductive biases to the chapter summary (Section 2.7), where we connect limitations of existing backbones to our proposed attention prior.

2.2 Chest X-ray Generation

2.2.1 Scalability, multimodality, and practical constraints

Chest X-ray (CXR) is one of the most frequently performed imaging examinations in routine care because it is fast, widely available, and can be acquired portably at bedside in emergency and intensive care settings. In many acute respiratory presentations, clinical imaging guidelines explicitly position chest radiography as a standard first-line modality, with chest CT typically reserved for complicated cases or when radiographs are negative/indeterminate despite persistent clinical concern [33]. Compared with cross-sectional imaging, CXR offers markedly lower radiation exposure, supporting its practical use in repeated follow-up scenarios [56].

From a data perspective, CXR is also *natively multimodal* in radiology workflows: each imaging study is accompanied by a free-text report and often includes multiple views (e.g., frontal and lateral), enabling scalable image–text conditioning and report-grounded supervision. Large public resources such as MIMIC-CXR [32] provide hundreds of thousands of radiographs paired with free-text reports (377,110 images across 227,835 studies), while its derived MIMIC-CXR-JPG release further lowers the barrier to model development by providing standardized JPG conversions and structured labels mined from reports. Complementary datasets such as CheXpert [9] and PadChest [6] further expand the spectrum of acquisition settings, label schemas, and linguistic/report styles.

At the same time, real-world CXR poses practical constraints that are especially salient for longitudinal generation. Substantial appearance variation arises from view position (PA/AP/lateral), portable bedside acquisition, patient posture and inspiration level, and device/site-specific processing; moreover, many large-scale labels are mined from reports and can be noisy or uncertain. These factors imply that successful longitudinal CXR generators must (i) preserve patient-specific global thoracic anatomy under acquisition shifts and (ii) allocate modeling capacity to subtle, localized pathological changes rather than spurious background variations.

2.2.2 Datasets and longitudinal pair construction

Public CXR datasets differ not only in scale and label schemas, but also in whether longitudinal follow-up can be reliably constructed. For longitudinal generation, a

typical pipeline builds training tuples from patient-level timelines:

$$(x_{\text{ref}}, x_{\text{fu}}, \Delta, y)\,, \tag{2.1}$$

where x_{ref} and x_{fu} are selected from the same patient under compatible acquisition settings whenever possible (e.g., matching view position and excluding severe projection mismatch).

A practical construction includes: **(i) Study linking and time ordering** using patient and study metadata; **(ii) View and quality filtering** to reduce spurious appearance shifts (PA/AP, portable studies, inspiration level); **(iii) Progression cue extraction** from free-text reports, such as report-difference summaries, entity-level changes, or instruction-style prompts derived from radiology narratives; **(iv) Optional spatial supervision** (lesion masks or coordinates) when available, used for evaluation or for lesion-aware conditioning.

This perspective clarifies why longitudinal generation is more challenging than single-study synthesis: the model must disentangle true disease evolution from acquisition-induced variations while maintaining patient-specific anatomy.

2.2.3 Multimodal foundations in radiology

Radiology is inherently multimodal: images are accompanied by free-text reports, clinical indications, and longitudinal follow-up context. This has driven rapid progress in medical vision–language modeling, where image–text data are used to learn representations that align visual findings with radiology terminology and support both recognition and grounding. For instance, BiomedCLIP [75] learns contrastive image–text representations from large-scale biomedical image–text pairs spanning diverse modalities, and demonstrates strong transfer and zero-shot performance across multiple biomedical tasks. Knowledge-enhanced pretraining such as MedKLIP [70] further injects domain knowledge into language–image pretraining to improve diagnostic grounding on X-ray scans. These multimodal encoders are not only useful for understanding tasks (e.g., classification and report generation), but also serve as effective conditioning modules for generative models by providing semantically dense, clinically meaningful text (or entity-level) embeddings.

2.2.4 Text-conditioned CXR synthesis

Building on radiology vision–language representation learning, text-to-CXR generation has emerged as a practical route to controllable synthesis and scalable data augmentation. RoentGen [8] adapts a text-conditioned latent diffusion model to the radiology domain using paired CXR–report data, demonstrating that free-form radiology prompts can steer salient appearances while maintaining plausible global structure; follow-up studies further report expert-based assessments and discuss potential use in augmentation

pipelines [4]. More recent efforts explore efficiency-oriented designs. For example, ChestDiffusion [31] proposes a lightweight diffusion framework with a domain-specific text encoder and a transformer-based denoiser to reduce computational cost while retaining fidelity for report-to-CXR generation. Overall, these pipelines establish viable mappings from radiology text to realistic single-study CXR images, but they are primarily designed for *single-time-point* synthesis and typically do not impose explicit temporal consistency constraints required by longitudinal follow-up generation.

2.2.5 CXR editing and counterfactual generation

Complementary to unconditional or text-conditioned synthesis, CXR editing and counterfactual generation aim to modify a given radiograph while preserving patient-specific anatomy. RadEdit [51] formulates text-guided diffusion image editing for chest X-rays to simulate controlled dataset shifts, using spatial constraints (e.g., masked regions) to localize changes and reduce editing artifacts. FastDiME [69] likewise leverages diffusion-based counterfactuals to add or remove spurious shortcut features, enabling systematic robustness assessment and facilitating mitigation studies by generating targeted variants. Importantly, these editing frameworks operationalize the core requirement shared with longitudinal progression modeling: clinically meaningful changes are often subtle and spatially localized, whereas the remaining anatomy should remain stable.

CXR-centric taxonomy. To provide a structured overview of representative medical image generative models while keeping CXR as the primary focus, Table 2.1 summarizes prior arts by *model family*, *imaging modality*, and *task type* (synthesis, translation/reconstruction, editing, and longitudinal prediction). This taxonomy will be referenced throughout the remainder of this chapter when discussing longitudinal CXR settings and backbone choices.

2.2.6 Longitudinal CXR generation: formulations and notation

We denote the reference chest radiograph as $x_{\text{ref}} \in \mathbb{R}^{H\times W}$, the target follow-up radiograph as $x_{\text{fu}} \in \mathbb{R}^{H\times W}$, and the follow-up interval as $\Delta > 0$. Longitudinal CXR generation is formulated as conditional synthesis of x_{fu} given a reference study and conditioning signal:

$$x_{\text{fu}} \sim p_\theta(x \mid x_{\text{ref}}, \Delta, y)\,, \tag{2.2}$$

where y represents progression cues. In radiology workflows, c may take different forms, including (i) a free-text progression description (e.g., report-difference or instruction), (ii) structured clinical attributes, or (iii) spatial priors such as lesion locations or masks when available.

Under this notation, we consider three common settings: **Forecasting**: generate x_{fu} from $(x_{\text{ref}}, \Delta, y)$; **Imputation**: generate missing scans given partial observations $\{x_{\text{obs},i}\}$ with irregular intervals; **Progression editing**: edit x_{ref} according to an instruction y while preserving patient-specific anatomy.

Model	Family	Modality	Task
RoentGen [8]	Diffusion	CXR (2D)	Synthesis
Chest-Diffusion [31]	Diffusion	CXR (2D)	Synthesis
CXR-IRGen [57]	Diffusion	CXR (2D) + Report	Synthesis
BiomedJourney [24]	Diffusion	CXR (2D) + Report	Longitudinal Prediction
PIE [40]	Diffusion	CXR (2D)	Longitudinal Prediction
ProgEmu [43]	Autoregression	CXR (2D) + Text	Longitudinal Prediction
RadEdit [51]	Diffusion	CXR (2D)	Editing
FastDiME [69]	Diffusion	CXR (2D)	Editing
CXR-Mpe [13]	Diffusion	CXR (2D)	Editing
Latent Drift [72]	Diffusion	CXR (2D)	Longitudinal Prediction
Spirit-Diffusion [15]	Diffusion	MRI (3D)	Reconstruction
LoCI-DiffCom [77]	Diffusion	Brain MRI (3D)	Longitudinal Prediction
SECONDGRAM [61]	Diffusion	Brain MRI	Longitudinal Prediction
SynDiff [49]	Diffusion (Adversarial)	MRI/CT	Translation / Synthesis

Table 2.1: Four-column taxonomy of representative medical image generation models (CXR-centric), categorized by model family, modality, and task.

2.2.7 From Single-Time CXR Generation to Longitudinal CXR Generation

Longitudinal CXR generation extends single-study synthesis and editing by introducing temporal dependencies: the model is required to preserve patient-specific thoracic anatomy while producing temporally consistent changes across follow-ups. Compared with single-time-point generation, longitudinal settings must account for irregular follow-up intervals and acquisition variability, which complicate the separation of true disease evolution from projection-induced appearance shifts. We formalize longitudinal problem settings and summarize their domain-specific challenges in Section 2.3.

2.3 Longitudinal Chest X-ray Generation: Problem Setup and Challenges

2.3.1 Clinical motivation and longitudinal settings

Longitudinal chest X-rays (CXRs) are routinely acquired to assess disease progression, treatment response, and potential complications. In everyday practice, clinicians compare studies across time points to track evolving pulmonary opacities, pleural effusions, lung nodules, post-operative states, or resolution of infection-related findings. This longitudinal comparison is particularly common in infectious and chronic respiratory diseases (e.g., tuberculosis and pneumonia) and in cancer surveillance, yet it remains time-consuming and experience-dependent.

Importantly, longitudinal CXR follow-up is not only a pragmatic habit but also supported by clinical guidance in specific scenarios. For example, British Thoracic Society

guidance for community-acquired pneumonia recommends arranging a repeat chest radiograph after about 6 weeks for patients with persistent symptoms/signs or those at higher risk of underlying malignancy [41]. Similarly, primary-care guidance also describes 6-week follow-up CXR in higher-risk pneumonia patients to avoid missing occult malignancy [59]. In tuberculosis control, WHO guidance recognizes chest radiography as an essential tool for TB detection and programmatic approaches, and national public-health guidelines likewise emphasize CXR as an important component in assessing TB contacts and suspected disease [48]. These guideline-backed practices underscore the clinical relevance of longitudinal CXR comparison and motivate AI-assisted longitudinal analysis and generative forecasting.

2.3.2 Problem settings: forecasting, imputation, and progression editing

We consider longitudinal CXR generation as conditional synthesis with patient-specific anatomical preservation. Common formulations include: (i) **Forecasting** (future follow-up prediction): generating a follow-up CXR given a reference CXR and a temporal/context condition (e.g., time gap or clinical progression description); (ii) **Imputation** (missing time-point completion): generating missing intermediate or follow-up scans from partially observed longitudinal trajectories; (iii) **Progression editing / counterfactual generation**: modifying a reference CXR according to a textual progression instruction while preserving patient identity, enabling "what-if" simulation.

2.3.3 Key challenges in longitudinal CXR generation

Longitudinal CXR generation poses several domain-specific challenges. First, **anatomical consistency** must be maintained: global thoracic structure should remain stable despite acquisition shifts (PA/AP, portable imaging, inspiration level) and across time. Second, clinically meaningful changes are often **subtle and localized**, requiring the model to allocate sufficient capacity to small pathology regions while avoiding spurious background variation. Third, real-world follow-ups exhibit **irregular time intervals** and heterogeneous clinical contexts, making temporal conditioning non-trivial. Finally, evaluation must go beyond visual realism, emphasizing **clinical faithfulness** and **structure preservation** under controlled change.

2.3.4 Evaluation protocols and metrics for longitudinal CXR generation

Evaluation for longitudinal CXR generation should go beyond generic visual realism (e.g., FID/SSIM) and reflect clinical requirements. We summarize three complementary axes:

(i) Structural preservation. The generated follow-up should maintain patient-specific thoracic anatomy. Practical measurements include global similarity under robust

alignment (e.g., registration-aware differences), anatomy-focused similarity computed on lung/heart regions, and identity preservation assessed by feature consistency from radiology encoders.

(ii) Change correctness. Clinically meaningful evolution is often sparse and localized. Evaluation should therefore examine whether changes occur in correct regions and directions, e.g., consistency with lesion masks when available, agreement of difference maps, and stability of non-target regions.

(iii) Clinical consistency. Generated images should be compatible with the intended clinical narrative. This can be assessed by report consistency (e.g., a report generator applied to synthetic images), vision–language alignment using radiology VLMs, and expert-based rating on realism, faithfulness, and diagnostic plausibility.

In this thesis, we use these axes to interpret prior methods and motivate lesion-aware inductive biases that better allocate modeling capacity to sparse disease regions.

2.3.5 Representative methods and positioning

Recent methods address longitudinal CXR generation by combining a reference study with progression cues (often derived from reports) and optimizing for follow-up realism under patient-specific anatomical preservation. Below we summarize representative approaches with a consistent lens: what they solve, how they implement conditioning, where they remain limited, and how these limitations motivate lesion-centric priors.

PIE [40]. *Goal/setting:* simulate disease progression through staged and controllable edits on CXR. *Method:* progressive editing conditioned on disease-related descriptions, aiming to generate step-wise evolution rather than a single jump. *Limitations:* when progression signals are subtle, purely description-driven edits may yield dispersed changes or inconsistent localization under acquisition variability. *Connection:* this highlights the need for a patient-specific mechanism that concentrates generation capacity on evolving regions while leaving the remaining anatomy stable.

BioMedJourney [24]. *Goal/setting:* counterfactual follow-up generation from longitudinal "patient journeys" using multimodal supervision. *Method:* construct training triples (prior image, progression description, follow-up image) and train a latent diffusion model conditioned on both the reference image and the progression text. *Limitations:* text cues can be semantically rich yet spatially under-specified; as a result, the model may align to the narrative without reliably placing changes in clinically correct regions. *Connection:* a lesion-aligned spatial prior can complement text conditioning by providing explicit, patient-specific localization signals.

CXR-IRGen [57]. *Goal/setting:* improve semantic alignment between generated CXR and clinical narratives by jointly leveraging image and report information from the reference study. *Method:* extract multimodal embeddings from the reference study and use them as conditioning for image synthesis (and report generation), strengthening cross-modal consistency. *Limitations:* embedding-level conditioning improves global semantics but

does not explicitly specify *where* clinically meaningful changes should occur in the image, which is critical for longitudinal progression. *Connection:* lesion-centric priors can act as a spatial complement to semantic conditioning.

ProgEmu [43]. *Goal/setting:* formulate longitudinal CXR generation as multimodal autoregression to jointly generate follow-up images and explanatory text. *Method:* tokenize images and texts into a unified discrete space and train a decoder-only Transformer with causal factorization. *Limitations:* causal decoding introduces order-dependent inductive biases tied to token traversal; this can dilute attention on small, localized progression regions under fixed generation order. *Connection:* a lightweight spatial prior injected into attention can mitigate such bias by prioritizing clinically relevant regions during decoding.

Diffusion-based counterfactual strategies [20]. *Goal/setting:* separate anatomy preservation from pathology synthesis for controlled counterfactual generation. *Method:* combine deterministic reconstruction for normal regions with stochastic sampling for abnormal regions (or analogous region-dependent strategies) to localize edits. *Limitations:* these approaches still depend on reliable region definitions and may not generalize when lesion locations are uncertain or vary across patients. *Connection:* a soft, learnable lesion-aligned prior offers a more flexible alternative to hard region partitioning.

Overall, these methods indicate that multimodal conditioning (reference image plus progression cues) is necessary but not sufficient for longitudinal CXR generation: the key difficulty lies in allocating modeling capacity to sparse, clinically meaningful change regions under realistic acquisition variability. This motivates spatially grounded, patient-specific priors that directly influence attention allocation, which we develop in later chapters.

2.4 Longitudinal CXR Generation via Autoregressive Models

Autoregressive (AR) generative models factorize the distribution of an image into a sequence of conditional predictions. Classical AR models such as PixelRNN [64] and PixelCNN [62] demonstrate strong likelihood-based modeling capacity, but their pixel-by-pixel decoding leads to high inference latency. With the rise of Transformers, token-based AR models have re-emerged as powerful tools for image and multimodal generation by operating on discrete visual tokens, enabling long-range dependency modeling via causal self-attention. Emu3 [66] exemplifies a unified next-token prediction framework that tokenizes images, text, and videos into a shared discrete space and trains a single decoder-only Transformer on mixed multimodal sequences; it reports strong performance across diverse generation and perception tasks, supporting the scalability and generality of next-token prediction beyond language.

In the medical imaging domain, ProgEmu [43] introduces a multimodal autoregressive framework for longitudinal CXR generation. Given a reference CXR and a textual progression description, ProgEmu generates a follow-up CXR alongside an explanatory caption. By tokenizing both modalities and modeling them within a unified causal sequence, ProgEmu enforces consistency between visual progression and textual narrative, highlighting the potential of AR models for interpretable clinical generation.

Despite these strengths, token-based AR image Transformers inherit order-dependent inductive biases from causal decoding under a chosen token traversal scheme. In raster-order generation, spatially adjacent regions may become distant in the 1D token sequence, which can weaken local coherence and introduce artifacts at row boundaries. In longitudinal CXR, where progression signals are often subtle and localized, such sequence-induced bias can further dilute modeling emphasis on clinically relevant regions. This motivates complementary mechanisms that modulate attention allocation with patient-specific spatial cues, which we review in Section 2.6 and revisit in the chapter summary (Section 2.7).

2.5 Longitudinal CXR Generation via Diffusion-based Models

Diffusion models generate samples by learning to reverse a gradual noising process, and have become a strong backbone for high-fidelity image synthesis in both natural and medical imaging. In practice, Latent Diffusion Models (LDMs) improve efficiency by performing denoising in a compressed latent space while retaining high-frequency details, which enables high-resolution generation and flexible conditioning (e.g., text, masks, and multi-modal inputs) at feasible compute cost.

Controllability via structural conditions

For longitudinal CXR, controllability is critical because the generator must preserve patient-specific global anatomy while applying localized, clinically meaningful changes. ControlNet [74] augments a pre-trained diffusion model with an external condition pathway (e.g., edges, depth, segmentation), providing a practical mechanism to impose spatial constraints on the denoising trajectory without fully retraining the backbone. Although ControlNet is not specific to radiology, its design motivates analogous uses of lesion masks or anatomy priors to constrain where changes are allowed during generation or editing.

Transformer-based diffusion backbones

Beyond U-Net denoisers, Transformer diffusion backbones are increasingly explored due to their global receptive field and scalability. U-ViT [3] treats the noisy image, timestep, and conditioning as tokens and uses long skip connections to stabilize training. DiT (Diffusion Transformer) [50] replaces the U-Net denoiser with a ViT-style backbone

operating on latent patches and demonstrates strong scaling behavior with improved sample quality as model capacity increases. These Transformer-based diffusion backbones are appealing for medical images with complex global structure, but they also raise challenges in allocating attention to subtle pathology regions when the change is spatially sparse.

Diffusion for longitudinal and counterfactual CXR generation

Diffusion-based longitudinal CXR generators typically condition on a reference radiograph together with progression cues (often derived from reports), and learn to synthesize follow-up images through iterative denoising in pixel or latent space. Representative CXR-specific systems and their positioning are summarized in Section 2.3 (see also Table 2.1). Here we emphasize diffusion-specific advantages for this setting: (i) stable optimization and strong mode coverage for high-fidelity synthesis, and (ii) flexible conditioning interfaces (text, masks, or structure priors) that naturally support counterfactual editing and localized progression simulation.

Limitations and connection to lesion-centric priors

While diffusion models offer strong fidelity and flexible conditioning interfaces, longitudinal CXR generation remains challenging when pathological changes occupy only a small portion of the image. In such cases, denoising capacity and attention may be allocated diffusely, leading to unstable localization of clinically meaningful changes. We therefore turn to a complementary line of research on attention manipulation and spatial prior injection (Section 2.6), and consolidate the shared gap across paradigms in Section 2.7.

2.6 Attention Manipulation and Spatial Prior Injection

Beyond backbone architectures, a complementary line of research investigates modifying or guiding attention distributions to improve spatial fidelity during generation. These methods can be broadly grouped into inference-time attention manipulation, conditional-branch priors, and position-dependent attention biasing.

2.6.1 Inference-Time Attention Manipulation

Several approaches adjust attention maps *post hoc* during the inference stage while keeping the pre-trained model parameters fixed. These methods intervene in the cross-attention or self-attention layers to steer the generative process.

Prompt-to-Prompt (P2P) [25] enables text-driven editing by directly manipulating the cross-attention layers. Based on the observation that the spatial structure of a generated image is largely determined by the cross-attention maps between pixels and text tokens, P2P injects the attention maps from a reference generation into the target

generation. This allows for localized semantic edits (e.g., changing "cat" to "dog") while strictly preserving the original spatial layout and background composition.

Attend-and-Excite [11] addresses the issue of "catastrophic neglect," where generative models fail to render one or more subjects specified in the prompt. It introduces an optimization loop during inference that calculates a loss based on the maximum attention values of specific subject tokens. By iteratively updating the latent representations to minimize this loss, it explicitly "excites" the attention maps for neglected terms, ensuring that all semantically meaningful regions are represented in the final output.

Self-Attention Guidance (SAG) [29] improves sample quality by refining the internal self-attention mechanism. It operates by masking regions with high attention scores and computing the discrepancy between the model's output on the masked versus unmasked inputs. This difference serves as an adversarial gradient signal to guide the denoising process, effectively sharpening the attention on salient objects and reducing background artifacts without requiring external supervision.

While these inference-time interventions are effective for general artistic control and semantic alignment, they introduce no medically grounded inductive bias. They optimize for general text-image correspondence rather than the precise, anatomical constraints required for longitudinal monitoring, and thus cannot enforce consistent lesion-centric focus across diverse patient samples.

2.6.2 Layout and Coordinate-Guided Generation

A more direct form of control involves injecting spatial coordinates (bounding boxes or keypoints) into the generation process. GLIGEN (Grounded Language-to-Image Generation) [39] extends pre-trained diffusion models to condition on grounding inputs like bounding boxes. By freezing the original weights and injecting "grounding tokens" via gated self-attention layers, GLIGEN enables precise open-set object layout control without retraining the entire model. Similarly, methods like BoxDiff [71] utilize layout constraints to guide the attention maps of diffusion models, ensuring that objects appear in user-specified regions. These approaches demonstrate that explicit coordinate-based guidance is highly effective for spatial control. However, they typically require training additional adapter layers (as in GLIGEN) or complex iterative optimization (as in BoxDiff). Our proposed GBCA draws inspiration from these coordinate-based priors but simplifies the integration by injecting the prior directly as a bias term, avoiding the need for additional gated layers.

2.6.3 Position-Dependent Attention Biases

Orthogonal to the above strategies, several architectures incorporate explicit bias terms directly into the attention logits to encode spatial or sequential relationships. T5 [54] introduces relative position biases by adding a learnable scalar to the attention scores based on the distance between the query and key tokens. This allows the model to

generalize beyond fixed sequence lengths but relies on static, learned tables shared across all attention heads. Swin Transformer [42] extends this concept to 2D vision by utilizing a relative position bias table for window-based self-attention. This mechanism effectively captures local spatial dependencies within image patches but treats all image content uniformly regardless of semantic significance. ALiBi (Attention with Linear Biases) [52] imposes a non-learnable, head-specific linear penalty to attention scores as a function of the distance between tokens. While ALiBi enables robust length extrapolation and training stability, its rigid geometric decay assumes that relevance is strictly a function of proximity.

Crucially, while these methods introduce stable and interpretable inductive biases, they remain fundamentally *input-agnostic*—the bias values are determined solely by relative coordinates, independent of the actual image content. Consequently, they cannot dynamically adapt their attention patterns to patient-specific lesion distributions or pathological abnormalities, which often require long-range dependencies that defy simple geometric proximity.

2.7 Summary and Motivation

This chapter reviewed medical image generation with a focus on CXR and its longitudinal setting, covering two representative paradigms: multimodal autoregressive token generation and diffusion-based synthesis (including Transformer backbones). Across these lines of work, a consistent requirement emerges for longitudinal CXR: the generator must preserve patient-specific thoracic anatomy while producing temporally coherent changes that are typically subtle and spatially sparse.

Shared limitation across paradigms. Despite substantial progress, existing longitudinal CXR generators often lack an explicit, patient-specific mechanism to concentrate modeling capacity on evolving disease regions. Autoregressive Transformers may inherit order-dependent biases from causal tokenization, which can dilute attention on small change regions under fixed traversal schemes. Diffusion-based approaches, while flexible in conditioning, can still allocate denoising capacity diffusely when the clinically relevant signal occupies only a small fraction of the field-of-view. As a result, models may produce samples that are visually plausible yet less reliable in lesion localization and progression faithfulness under realistic acquisition variability.

Why existing controllability tools are insufficient. Prior efforts to improve controllability can be grouped into three categories. First, inference-time attention interventions (e.g., cross-attention control or self-attention guidance) can steer generation without retraining, but they primarily optimize generic text–image correspondence or sample quality, rather than enforcing patient-specific, clinically grounded focus. Second, conditional control branches (e.g., ControlNet-style pathways) can impose structural constraints but typically require additional control signals and condition-specific parameters, increasing system complexity and limiting applicability when reliable spatial

supervision is unavailable. Third, position-dependent attention biases (e.g., relative position bias tables or distance-based linear biases) provide stable inductive structure but remain input-agnostic, and therefore cannot adapt to patient-specific lesion distributions.

Motivation for GBCA. As summarized in Figure 2.1, although prior methods differ in backbone and conditioning strategy, both autoregressive and diffusion-based approaches still lack an explicit patient-specific spatial prior for sparse lesion evolution. Motivated by this gap, we propose **Gaussian-Biased Causal Attention (GBCA)**, a lightweight mechanism that injects a lesion-aligned Gaussian spatial prior directly into Transformer attention logits. By providing a content-aware, patient-specific bias with minimal architectural overhead, GBCA is compatible with both autoregressive decoding and diffusion-based denoising, aiming to improve lesion-centric focus while preserving global anatomy. A

Chapter 3

Background Knowledge

This chapter provides the theoretical foundation for the methodologies presented in this thesis. We begin with an overview of generative modeling in medical imaging, tracing the evolution from Generative Adversarial Networks (GANs) [22] to modern likelihood-based and score/flow-based models. Subsequently, we provide a detailed technical introduction to Autoregressive Models and Diffusion Probabilistic Models, focusing on the Transformer architectures that underpin our proposed methods. Finally, we discuss Vision-Language Models (VLMs), which serve as the core engine for our proposed spatial guidance mechanism.

To facilitate a quick comparison, Table 3.1 summarizes representative task-level longitudinal CXR generation methods by paradigm and conditioning. Table 3.2 lists common building blocks across AR/diffusion/flow pipelines by stage and purpose. Table 3.3 compares representative VLMs by whether they provide explicit visual grounding outputs, which underpins our coordinate-based spatial priors in Chapter 4.

3.1 Generative Models in Medical Imaging

Generative modeling aims to learn the underlying data distribution $p_{\text{data}}(x)$ from a set of observed samples, enabling the synthesis of new, plausible data points. In the medical domain, this capability is particularly valuable for tasks such as data augmentation, anomaly detection, and longitudinal prediction.

Early approaches largely relied on Generative Adversarial Networks (GANs) [22]. GANs employ a minimax game between a generator, which synthesizes images, and a discriminator, which distinguishes real from fake samples. While GANs have demonstrated success in tasks like cross-modality translation (e.g., CycleGAN [76] for CT-to-MRI synthesis), they notoriously suffer from training instability and mode collapse, where the model fails to capture the full diversity of the data distribution. Furthermore, GANs typically lack a tractable likelihood measure, making it difficult to rigorously evaluate density estimation performance.

Beyond the original GAN formulation, practical GAN variants and translation pipelines (e.g., CycleGAN [76]) and high-fidelity generators (e.g., StyleGAN [35]) have

Work	Paradigm	Conditioning
PIE [40]	Edit-based	Reference CXR + lesion edit specification
BioMedJourney [24]	Latent diffusion	Two time-point CXRs + progression text
CXR-IRGen [57]	Diffusion fine-tuning	Reference-study CLIP embeddings
ProgEmu [43]	AR transformer	Image tokens (ref/follow-up) + progression text tokens

Table 3.1: Representative task-level works for longitudinal CXR generation.

Technique	Stage	Purpose
VQ-VAE [63]	Tokenization	Map images to discrete latent tokens
MaskGIT [10]	Token modeling	Iterative masked-token generation
DDIM [60]	Sampling	Reduce diffusion sampling steps
CFG [28]	Guidance	Strengthen conditional generation
LDM [55]	Diffusion	Generate in compressed latent space
ControlNet [74]	Control	Inject spatial conditions via an auxiliary branch
Rectified Flow [18]	Flow	Fewer-step generation via flow matching

Table 3.2: Common building blocks for AR/diffusion/flow-based image generation.

been adopted as baselines in medical image synthesis. Recent surveys further highlight the rapid adoption of diffusion-based methods across medical imaging tasks due to their stable optimization and strong controllability [36].

To address the limitations of GANs, research has shifted toward likelihood-based or score/flow-based paradigms, notably Autoregressive Models (AR) and Diffusion Probabilistic Models (DPM). These models offer stable training objectives and have achieved state-of-the-art results in general computer vision, paving the way for high-fidelity medical image synthesis. In particular, diffusion models [27, 55] and transformer-based diffusion backbones such as DiT [50] have become a practical foundation for controllable, high-fidelity generation. More recently, rectified-flow formulations have been actively studied as an alternative that can support fewer-step sampling while preserving image quality, especially with transformer backbones [18]. This trend motivates our choice of a large-scale transformer backbone in Chapter 5 and the emphasis on spatial controllability in Chapters 4,5. Before introducing AR and diffusion in detail, Table 3.2 lists common building blocks across token-based, diffusion-based, and flow-matching pipelines, which will be referenced in Chapters 4–5.

3.2 Autoregressive Generative Models

Autoregressive models treat image generation as a sequence modeling problem. They factorize the joint probability distribution of an image into a product of conditional probabilities.

3.2.1 Sequence Modeling and Next-Token Prediction

Given an image represented as a sequence of discrete tokens $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, the joint probability $p(\mathbf{x})$ is modeled as:

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n \mid x_{<n}; \theta), \tag{3.1}$$

where $x_{<n}$ denotes the sequence of tokens preceding position n, and θ represents the model parameters. Learning is performed by maximizing the log-likelihood of the training data:

$$\mathcal{L}_{AR} = \sum_{n=1}^{N} \log p(x_n \mid x_{<n}; \theta). \tag{3.2}$$

This "next-token prediction" paradigm, popularized by models like GPT in natural language processing, allows for robust density estimation. In computer vision, images are commonly tokenized via Vector Quantized Variational Autoencoders (VQ-VAE) [63], converting continuous pixel values into a sequence of discrete codebook indices.

In practice, image tokens are obtained by rasterizing a 2D latent grid into a 1D sequence with a fixed ordering (e.g., row-major). Positional encodings then provide spatial indices for each token. This design choice is not innocuous: ordering and positional bias may affect where the model allocates causal attention, which motivates the attention analysis and modulation introduced in Chapter 4.

Representative autoregressive or token-based image generation models include iGPT [12], which directly applies GPT-style decoding to visual sequences, as well as VQ-VAE-based pipelines that model discrete latent codes with transformer priors. In addition, non-autoregressive token models such as MaskGIT [10] replace strict raster-scan decoding with masked token prediction and iterative refinement, offering faster generation while retaining transformer-based sequence modeling.

3.2.2 Transformers and Causal Self-Attention

The Transformer architecture [65] has become the de facto backbone for autoregressive modeling due to its ability to capture long-range dependencies. The core component of the Transformer is self-attention.

For a sequence of input embeddings $X \in \mathbb{R}^{N \times d}$, the model computes Queries (Q), Keys (K), and Values (V) via linear projections. The attention output is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right) V, \tag{3.3}$$

where d_k is the dimension of the key vectors, and M is a **causal mask**.

The causal mask M is crucial for autoregressive generation. It ensures that the prediction for token x_t depends only on past tokens $x_1, \ldots, x_{t-1}$ by setting attention scores for future positions to $-\infty$ (pre-softmax):

$$M_{ij} = \begin{cases} 0 & \text{if } i \geq j, \\ -\infty & \text{otherwise.} \end{cases} \tag{3.4}$$

In Chapter 4, we will introduce a modification to this standard attention mechanism, termed *Gaussian-Biased Causal Attention* (GBCA), to explicitly inject spatial lesion priors into the generation process.

3.3 Diffusion Probabilistic Models

Diffusion models [27] are a class of latent variable models inspired by non-equilibrium thermodynamics. They learn to generate data by reversing a gradual noise-adding process.

3.3.1 Forward and Reverse Processes

Denoising Diffusion Probabilistic Models (DDPMs) operate through two distinct Markov chains: a forward process that destroys information and a reverse process that restores it.

The Forward Process (Diffusion)

The **forward process**, denoted as q, is a fixed Markov chain that gradually adds Gaussian noise to the real data $x_0 \sim q(x_0)$ over a pre-defined number of steps K. At each diffusion step τ, the transition is parameterized by a variance schedule $\beta_\tau \in (0, 1)$:

$$q(x_\tau \mid x_{\tau-1}) = \mathcal{N}(x_\tau; \sqrt{1-\beta_\tau}x_{\tau-1}, \beta_\tau \mathbf{I}). \tag{3.5}$$

where x_τ represents the latent variable (noisy image) at time step t. β_τ controls the step size of the noise injection. $\mathbf{I}$ is the identity matrix, indicating that noise is added independently to each dimension.

A key property of this process is the ability to sample x_τ at any diffusion step τ directly from x_0 without iterating through intermediate steps. Let $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$. The marginal distribution $q(x_\tau \mid x_0)$ can be expressed in closed form:

$$q(x_\tau \mid x_0) = \mathcal{N}(x_\tau; \sqrt{\bar{\alpha}_\tau}x_0, (1-\bar{\alpha}_\tau)\mathbf{I}). \tag{3.6}$$

Using the reparameterization trick, we can express a sample x_τ as:

$$x_\tau = \sqrt{\bar{\alpha}_\tau}x_0 + \sqrt{1-\bar{\alpha}_\tau}\epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3.7}$$

As $K \to \infty$, $\bar{\alpha}_K \to 0$, and the distribution of x_K converges to a standard isotropic Gaussian $\mathcal{N}(0, \mathbf{I})$, meaning all original signal is lost.

The Reverse Process (Denoising)

The goal of the generative model is to reverse this process—starting from pure noise $x_K \sim \mathcal{N}(0, \mathbf{I})$ and sequentially denoising it to recover a sample from the data distribution x_0. Since the true posterior $q(x_{\tau-1} \mid x_\tau)$ is intractable, we approximate it using a parameterized probabilistic model p_θ:

$$p_\theta(x_{\tau-1} \mid x_\tau) = \mathcal{N}(x_{\tau-1}; \mu_\theta(x_\tau, \tau), \Sigma_\theta(x_\tau, \tau)). \tag{3.8}$$

where θ represents the learnable parameters of a neural network (typically a U-Net or Transformer). $\mu_\theta(x_\tau, \tau)$ is the predicted mean of the denoised distribution. $\Sigma_\theta(x_\tau, \tau)$ is the variance, often fixed for stability in standard implementations.

Training Objective

Ho et al. [27] demonstrated that it is more stable to parameterize the model to predict the noise ϵ added to x_0. The training objective is derived from the ELBO but simplified to a reweighted mean squared error:

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0, \epsilon, \tau}\left[\|\epsilon - \epsilon_\theta(x_\tau, \tau)\|^2\right], \tag{3.9}$$

where x_0 is a real image sampled from the training dataset. τ is a time step sampled uniformly from $\{1, \ldots, T\}$. $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the injected Gaussian noise. $\epsilon_\theta(x_\tau, \tau)$ is the noise predicted by the neural network given the noisy input x_τ and time embedding τ.

By minimizing this loss, the network learns to iteratively remove noise, guiding the trajectory from the latent noise space back to the data manifold.

3.3.2 Practical advances: fast sampling, guidance, and flow-matching

Fast samplers such as DDIM [60] reduce the number of reverse steps without changing the training objective, making diffusion practical for high-resolution synthesis. Classifier-free guidance [28] further improves conditional controllability by interpolating conditional and unconditional predictions. Beyond classical DDPM formulations, EDM [34] provides a clearer design space and improved preconditioning, yielding better quality–efficiency trade-offs. More recently, rectified-flow/flow-matching formulations have been explored as an alternative that can support fewer-step generation with transformer backbones [18], aligning with our Chapter 5 backbone choice.

3.3.3 Conditioning mechanisms for controllable diffusion

Practical diffusion systems rely on explicit conditioning interfaces to control generation. A common approach is to inject text or other conditions through cross-attention, where

the denoising network attends to condition embeddings (e.g., prompt tokens) while processing visual latents. Image-conditioned diffusion can also be implemented by concatenating condition features with the noisy latents (channel-wise concatenation) or by using FiLM/adaLN-style modulation, where scale-and-shift parameters are predicted from the condition vector and applied to normalized activations. For spatially structured conditions (e.g., edges, masks, bounding boxes), ControlNet [74] introduces an additional conditioning branch to preserve the original backbone while enabling strong spatial control. These conditioning interfaces provide natural insertion points for spatial priors, which we exploit in Chapters 4–5 to steer attention and edits toward lesion-relevant regions.

3.3.4 Latent Diffusion and DiT Architectures

Latent Diffusion Models (LDMs)

Standard diffusion models operating directly in pixel space face significant computational challenges due to the high dimensionality of image data. To mitigate this, **Latent Diffusion Models (LDMs)** [55] propose a two-stage approach that separates the learning of perceptual compression from the generative modeling of semantic content.

LDMs utilize a pre-trained perceptual compression model consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The encoder maps an input image $I \in \mathbb{R}^{H\times W\times 3}$ into a lower-dimensional latent representation $z = \mathcal{E}(I) \in \mathbb{R}^{h\times w\times C}$. The generative diffusion process is then performed in latent space. The training objective is modified to:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon,\tau}\left[\|\epsilon - \epsilon_\theta(z_\tau, \tau, y)\|^2\right], \tag{3.10}$$

where z_τ is the noisy latent at timestep τ, y represents conditioning inputs (e.g., text prompts), and ϵ_θ is the denoising network. Once the latent z_0 is generated, the decoder reconstructs the pixel-space image $\hat{I} = \mathcal{D}(z_0)$. This paradigm shifts computation from pixel space to a compressed latent manifold, enabling high-fidelity synthesis with reduced resource consumption.

Diffusion Transformers (DiT)

While early LDMs relied on U-Net backbones, recent advancements have introduced the **Diffusion Transformer (DiT)** [50], which aligns the diffusion backbone with the scalable Transformer architecture [65]. Unlike U-Nets, which rely on inductive biases favoring spatial locality and translation invariance, DiT treats image generation as a token-sequence processing problem, offering superior scalability and global context modeling.

Patchification and Embedding. Given a latent input $z \in \mathbb{R}^{h\times w\times C}$, DiT divides it into a sequence of patches of size $p \times p$. These patches are flattened and linearly projected into a hidden dimension d, resulting in a token sequence $\mathbf{X} \in \mathbb{R}^{N\times d}$, where $N = \frac{hw}{p^2}$.

Model	Training	Grounding output
CLIP [53]	Contrastive	None (global embedding)
MedCLIP [75]	Medical contrastive	None (global embedding)
MedKLIP [70]	Medical VLP	None (global embedding)
LLaVA-Med [38]	Instruction-tuned MLLM	Text; grounding varies by variant
Qwen-VL [2]	Grounded MLLM	Box/coordinate tokens

Table 3.3: VLMs and whether they provide explicit visual grounding outputs.

Positional embeddings are added to retain spatial information, analogous to Vision Transformers (ViT).

Transformer Blocks with adaptive Layer Norm (adaLN). DiT stacks standard Transformer blocks containing multi-head self-attention and feed-forward networks. A key conditioning mechanism is **adaLN-Zero**, which injects diffusion step τ and condition y by regressing scale (γ) and shift (β) parameters from the conditioning embeddings:

$$\text{adaLN}(X, \tau, y) = \gamma(\tau, y) \cdot \text{Norm}(X) + \beta(\tau, y). \tag{3.11}$$

This design enables strong, stable conditioning across diffusion timesteps.

Scalability and relevance. DiT models exhibit favorable scaling behavior, where increasing model size and token count improves sample quality. In Chapter 5, we adopt **FLUX.1-dev**, a large-scale rectified-flow transformer backbone. Its global dependency modeling via self-attention is particularly suitable for our proposed vital-layer manipulation and spatially guided editing.

3.4 Vision-Language Models (VLMs) as Spatial Priors

Table 3.3 compares representative VLMs by whether they provide explicit grounding outputs, which enables our coordinate-to-Gaussian spatial priors in Chapter 4.

The generation of longitudinal medical images requires not only high visual fidelity but also precise semantic control over disease progression. While generative models (AR and diffusion) excel at texture synthesis, they often lack robust understanding of complex clinical language and spatial reasoning. To bridge this gap, we leverage large Vision-Language Models (VLMs) as a semantic interpreter and spatial guide.

3.4.1 The Evolution from Contrastive to Generative VLMs

The field of vision-language modeling has undergone a paradigm shift from discriminative alignment to generative reasoning.

Contrastive Dual-Encoder Models

Early foundational models, most notably Contrastive Language-Image Pre-training (CLIP) [53], adopt a dual-encoder architecture. They consist of separate image and text encoders that project data into a shared embedding space. The training objective maximizes similarity between matched image-text pairs while minimizing it for unmatched pairs via a contrastive loss (e.g., InfoNCE). While CLIP excels at zero-shot classification and global image-text matching, it aggregates information into a global vector and thus provides limited spatial grounding for dense localization.

Generative Multimodal Large Language Models (MLLMs)

To address the limitations of contrastive models, recent work has converged on Generative VLMs (MLLMs), such as **LLaMA-Vision** [23] and **Qwen-VL** [2]. These models typically comprise: (i) a visual encoder (e.g., ViT), (ii) an adapter/projector aligning visual features to the LLM token space, and (iii) an LLM backbone that performs autoregressive next-token prediction over interleaved visual and text tokens. Instruction tuning further enables them to follow complex prompts and produce structured outputs.

3.4.2 Visual Grounding and Coordinate Prediction

A capability central to this thesis is **visual grounding**, also referred to as Referring Expression Comprehension (REC). This task localizes a specific region in an image described by a natural language query (e.g., "the consolidation in the right lower lobe"). Unlike closed-set object detectors (e.g., Faster R-CNN [21]), generative VLMs can treat grounding as sequence generation by emitting discretized coordinate tokens. For example, Qwen-VL [2] quantizes coordinates into bins and outputs bounding boxes as structured tokens.

3.4.3 From coordinates to Gaussian spatial priors

In this thesis, we convert VLM-predicted coordinates into continuous spatial priors used for attention modulation and editing. Let the VLM output a bounding box

$$\mathbf{b}^{\text{box}} = (x_{\min}, y_{\min}, x_{\max}, y_{\max}), \tag{3.12}$$

and define its center as $\mu = (\frac{x_{\min}+x_{\max}}{2}, \frac{y_{\min}+y_{\max}}{2})$. We then construct a 2D Gaussian prior over pixel coordinates (u, v):

$$G(u, v) = \exp\left(-\frac{\|(u, v) - \mu\|_2^2}{2\sigma^2}\right), \tag{3.13}$$

where σ controls the spatial spread. This soft prior provides a lightweight mechanism to bias attention toward lesion-relevant regions without requiring dense pixel-level supervision.

3.4.4 VLMs in Medical Imaging

In the context of our LeGend framework (Chapter 4) and diffusion-based editing (Chapter 5), we utilize generative VLMs as an offline reasoning module. Given a reference CXR and a progression description, we instruct the model to perform **medical visual grounding**: identifying the image region where the described pathological change occurs. The predicted coordinates are subsequently transformed into Gaussian spatial priors (Section 3.4.3), decoupling "understanding medical text" from "generating pixels".

Recent medical-domain VLMs such as MedCLIP [75] and MedKLIP [70] demonstrate that domain-adapted vision-language pretraining improves radiology-centric alignment, while instruction-tuned biomedical MLLMs such as LLaVA-Med [38] enable open-ended reasoning over biomedical images. These developments support the feasibility of leveraging VLM-generated coordinates as a lightweight yet effective spatial prior for guiding longitudinal generation.

3.5 Summary

This chapter reviewed the fundamental technologies underpinning this thesis. Autoregressive models provide a scalable sequence-based approach to generation but require careful handling of tokenization, ordering, and spatial control. Diffusion models, particularly those based on Transformer backbones (DiT) and enhanced by practical techniques such as fast sampling and guidance, offer strong fidelity and stability with flexible conditioning interfaces. Finally, VLMs provide semantic grounding and coordinate-level spatial cues that can be converted into continuous Gaussian priors. In the following chapters, we will demonstrate how integrating these components leads to accurate and clinically meaningful longitudinal chest X-ray generation.

Chapter 4

Longitudinal Chest X-ray Generation via Autoregression Model

In this chapter, we systematically revealed a corner-biased attention pattern in autoregressive longitudinal CXR generation and introduced LeGend with GBCA to correct it. By injecting lightweight, lesion-centered Gaussian spatial priors into mid-depth causal attention layers, GBCA improves lesion attention, yielding follow-up CXRs with clearer diagnostic cues. Using paired studies as objective surrogates of disease trajectories, our evaluations demonstrate that LeGend is able to produce follow-ups that accurately capture the described progression while preserving high visual realism and achieving state-of-the-art performance.

For clarity, we use distinct symbols for different notions of sequence indexing and attention analysis throughout this chapter. Specifically, Δ denotes the follow-up interval, n denotes the autoregressive token position, D denotes the progression description, and X_D denotes its token sequence. We use S for the total input sequence length and $S^{(\ell)}$ for the pre-softmax attention logits at layer ℓ. In addition, ϕ_A denotes the mapping from a reference-image token index to its 2D spatial location, and $\mathcal{H}^{(\ell)}$ denotes the corresponding layer-wise spatial attention heatmap.

4.1 Motivations and Contributions

As discussed in previous chapters, autoregressive models offer strong scalability and controllability for image generation, but suffer from structural limitations when applied to longitudinal medical imaging. Notably, conventional autoregressive decoding tends to allocate excessive attention to image corners and borders—an issue we call corner-edge attention bias. This bias weakens the model's capacity to represent clinically meaningful lesion evolution and leads to generated images that are visually plausible yet semantically inconsistent. In particular, Figure 4.2 provides quantitative evidence of this phenomenon: the attention mass concentrates disproportionately on corner/edge

regions (high CBI/EAR) while exhibiting reduced overlap with clinically relevant lesion areas (lower Attn-IoU) across decoding depth. This layer-wise profiling indicates that the model can produce visually plausible follow-up CXRs yet allocate causal attention to peripheral, low-semantic regions, motivating an explicit lesion-centric spatial prior to correct the bias.

Our analysis reveals that this bias is especially problematic in follow-up CXR synthesis. Medical lesions tend to be fine-grained, highly localized, and variable in size and morphology. Autoregressive models, trained without spatial priors, often fail to properly focus on these regions. To address this, we propose LeGend, a lightweight and modular enhancement that introduces lesion-guided attention into the autoregressive generation pipeline.

The main contributions of our work are:

We identify and quantify the spatial attention bias present in standard autoregressive decoding when applied to longitudinal CXR synthesis.

We propose a novel Gaussian-Biased Causal Attention (GBCA) mechanism that dynamically modulates self-attention using lesion priors.

We develop a practical pipeline to extract lesion coordinates from a vision-language model and inject them into the generation process.

We demonstrate that LeGend significantly improves lesion alignment, attention interpretability, image fidelity, and downstream disease classification performance.

4.2 Methodology

The longitudinal CXR generation task aims to generate the follow-up X-ray image $\hat{I}_B$, given the prior CXR image I_A and the progression description D. We begin with a diagnostic analysis of autoregressive decoding and introduce quantitative metrics to assess causal alignment between attention and pathology. This reveals a previously unrecognized corner/edge bias, an important conceptual contribution of this work (Section 4.2.1). Guided by these findings, we map sparse lesion coordinates from a vision–language model to a 2D Gaussian spatial prior and inject it as an additive bias into selected causal self-attention logits (Section 4.2.2). LeGend integrates this Gaussian-Biased Causal Attention (GBCA) into a general-purpose autoregressive backbone (Section 4.2.3), substantially improving lesion-aware causal alignment. The inference pipeline is shown in Figure 4.1.

To avoid ambiguity, we distinguish two regimes in this chapter. For training and for the attention-bias diagnosis in Section 4.2.1, we use the teacher-forcing setting, where the decoder observes the ground-truth follow-up prefix. For generation at test time, we use autoregressive inference, where the decoder has access only to the reference image, the progression description, and the previously generated follow-up prefix. The same

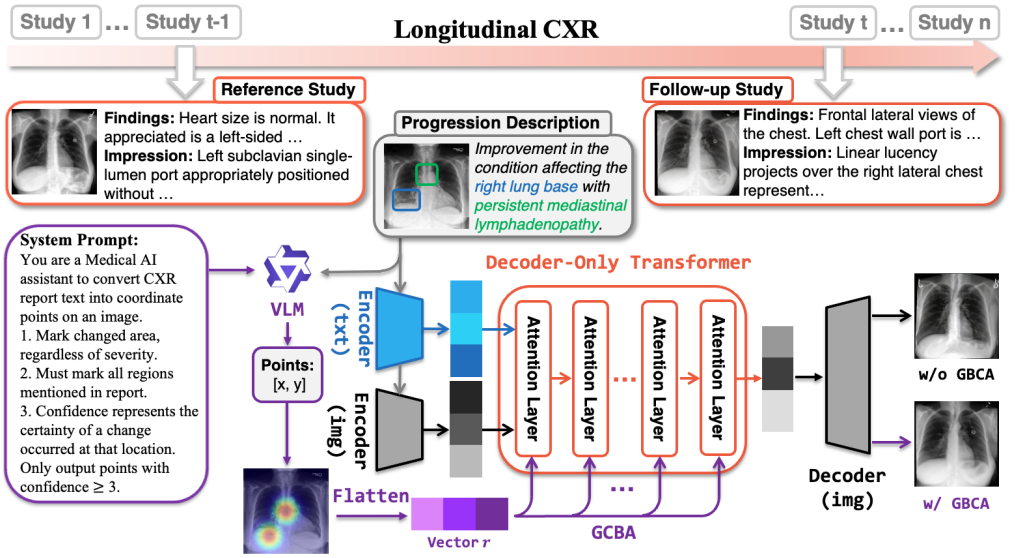


Figure 4.1: Inference-time overview of LeGend. Given a reference CXR I_A and a progression description D, a vision–language model (VLM) predicts sparse lesion coordinates on the reference image. These coordinates are converted into a 2D Gaussian prior, which is injected as an additive bias into the decoder's causal self-attention logits (GBCA), steering attention toward clinically relevant reference regions while preserving causality. The decoder then autoregressively predicts the follow-up token sequence $\hat{X}_B$; after each step, the newly generated token is fed back as part of the visible prefix for subsequent decoding, and the final sequence is detokenized into the generated follow-up image $\hat{I}_B$. Ground-truth follow-up tokens X_B are not used in this inference pipeline.

decoder and GBCA formulation are used in both regimes; the difference lies only in whether the visible follow-up prefix comes from $X_{B,<n}$ or from $\hat{X}_{B,<n}$.

In the current formulation, the follow-up interval Δ is not provided as an explicit conditioning variable to the generator. Instead, the model is conditioned on the reference image and the progression description only, so the temporal aspect is represented implicitly through the semantics of the text rather than through a dedicated interval encoding.

4.2.1 Layer-wise Attention Bias Profiling

Conventional autoregressive fine-tuning struggles to precisely control lesion location, morphology, and severity in longitudinal follow-up CXR generation due to the absence of medical prior knowledge. Although fine-tuning improves perceptual realism, preliminary qualitative inspection suggests a skew of attention toward non-lesional corners and borders of the reference image, a "corner-focus" pattern that suggests sampling alignment rather than genuine lesion-semantic understanding. To examine this discrepancy between appearance and pathology, we analyze how the decoder allocates causal self-attention while synthesizing $\hat{I}_B$ from I_A.

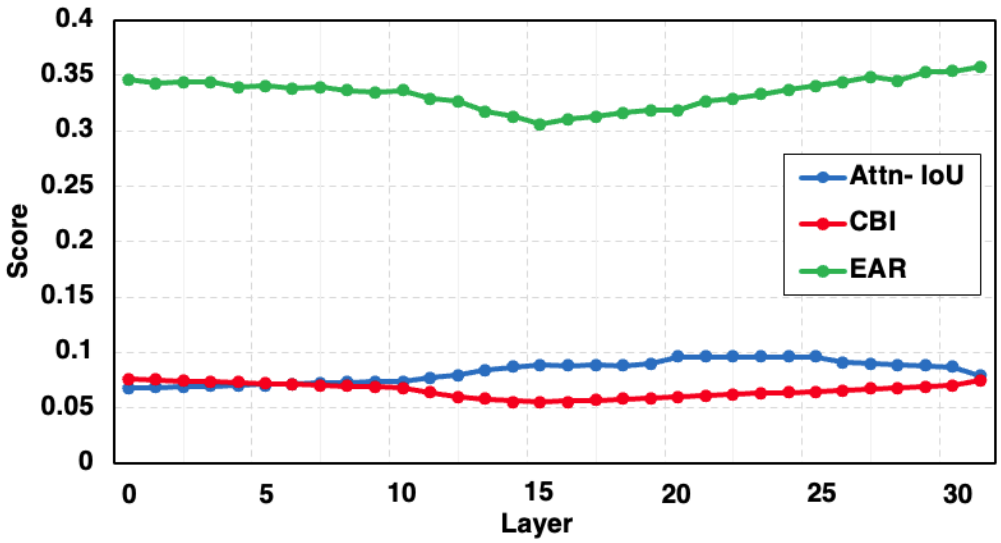


Figure 4.2: The layer-wise analysis of causal attention distribution with Attn-IoU, CBI, and EAR for the vanilla autoregressive model.

Teacher-forcing attention analysis. For the bias diagnosis in this subsection, we explicitly analyze the autoregressive generation model under the *teacher-forcing* setting rather than free-running inference. Concretely, we tokenize the reference image I_A, the progression description D, and the ground-truth follow-up image I_B into reference-image tokens X_A, text tokens X_D, and target follow-up tokens X_B, respectively, and form the input sequence $\{X_A, X_D, X_B\}$ of length S. Under teacher forcing, when predicting the n-th target token, the decoder is conditioned on the reference tokens, the text tokens, and the previous ground-truth follow-up prefix $X_{B,<n}$ under the causal mask.

By contrast, at inference time the ground-truth follow-up tokens are unavailable, and the same decoder predicts the next token conditioned on X_A, X_D, and the previously generated prefix $\hat{X}_{B,<n}$.

When predicting the n-th visual token in X_B, the query position $q \in Q_B$ attends over the visible keys $k \in \mathcal{K}_{\text{vis}}(q)$, where

$$\mathcal{K}_{\text{vis}}(q) = X_A \cup X_D \cup X_{B,<n},$$

and all other positions are masked out by causality. For each decoder layer ℓ, we read out the *pre-softmax* attention score (logit) matrix $S^{(\ell)} \in \mathbb{R}^{S\times S}$ with elements

$$S^{(\ell)}_{q,k} = \frac{Q^{(\ell)}_q \cdot K^{(\ell)}_k}{\sqrt{d_k}} + M_{q,k}, \tag{4.1}$$

where rows index query positions, columns index key positions, M enforces causality and visibility (invalid entries set to $-\infty$ so that they vanish after the Softmax). Let Q_B denote the index set of target follow-up token positions in X_B, and let K_A denote the index set of reference-image token positions in X_A. The sub-matrix $S^{(\ell)}_{B\to A}$ captures the

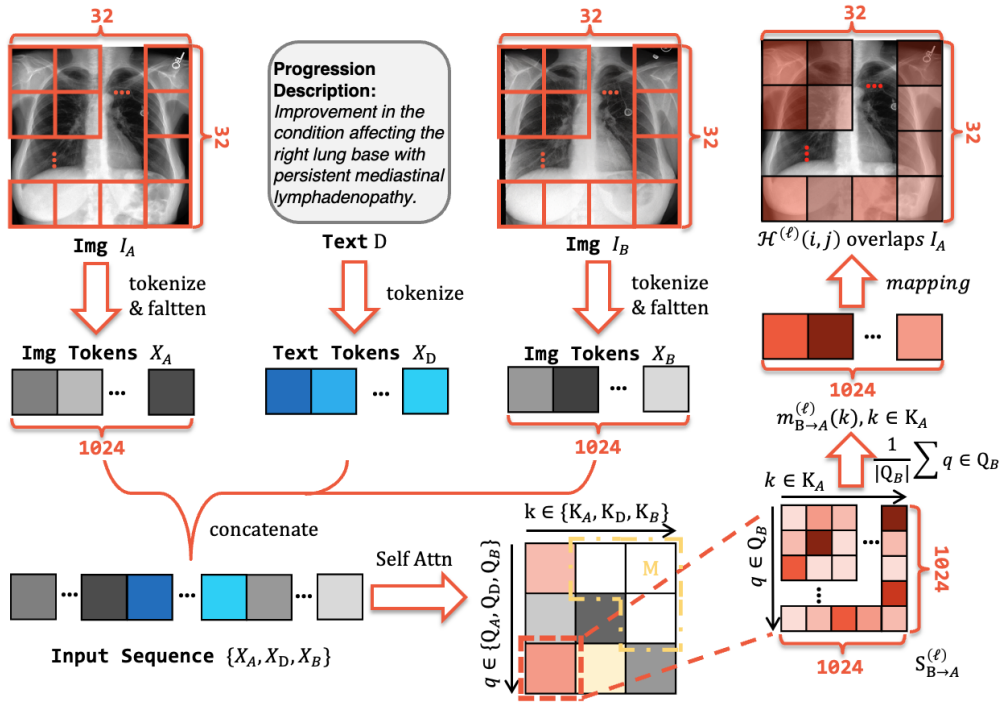


Figure 4.3: Teacher-forcing setup used for attention-bias profiling. The decoder input is the concatenated sequence $\{X_A, X_D, X_B\}$, where X_A denotes reference-image tokens, X_D denotes progression-text tokens, and X_B denotes the ground-truth follow-up tokens. When analyzing the query corresponding to the n-th target token in X_B, the causal mask allows attention only to X_A, X_D, and the preceding ground-truth prefix $X_{B,<n}$, while future follow-up tokens remain invisible. The heatmap is obtained by aggregating attention from follow-up queries to reference-image keys.

logits from follow-up queries to reference-image keys at layer ℓ:

$$S^{(\ell)}_{B\to A} = S^{(\ell)}[Q_B, K_A] \in \mathbb{R}^{|Q_B|\times|K_A|}. \tag{4.2}$$

We then average over the query dimension to obtain a single score for each reference-image token:

$$m^{(\ell)}_{B\to A}(k) = \frac{1}{|Q_B|}\sum_{q\in Q_B}\left[S^{(\ell)}_{B\to A}\right]_{q,k}, \qquad k \in K_A. \tag{4.3}$$

Intuitively, $m^{(\ell)}_{\hat{B}\to A}(k)$ measures how strongly the follow-up image, as a whole, attends to the k-th reference token at layer ℓ in terms of raw attention logits.

Finally, we project these token-wise scores back to the spatial grid of the reference image. Let $\phi_A : k \mapsto (i,j)$ map a reference-image token index k to its spatial coordinates (i,j) on the $H \times W$ grid (here $H = W = 32$). The layer-wise spatial attention heatmap for $(i,j) \in \{0,\dots,H-1\} \times \{0,\dots,W-1\}$ is then defined as

$$\mathcal{H}^{(\ell)}(i,j) = m^{(\ell)}_{B\to A}(\phi_A^{-1}(i,j)), \tag{4.4}$$

where ϕ_A is a bijection between the flattened reference-token indices and the $H \times W$ spatial lattice, and $\phi_A^{-1}(i,j)$ returns the token index corresponding to spatial location (i,j).

Using these layer-wise heatmaps, Figure 1.2 shows that fine-tuned autoregressive transformers frequently allocate disproportionate mass to image borders and corners, even when lesions lie far from these regions. This *corner–edge over-focus* suggests reliance on sampling alignment rather than lesion semantics, motivating a quantitative analysis.

Spatial Bias Analysis. To systematically quantify this spatial imbalance and characterize how attention is distributed during decoding, we introduce two metrics: the **Corner Bias Index (CBI)** and the **Edge Attention Ratio (EAR)** based on $\mathcal{H}^{(\ell)}(i,j)$. CBI is defined as the proportion of attention mass that falls within the four corner regions $\mathcal{C}$. The corner mask $\mathcal{C}$ consists of four rectangular regions, each with dimensions $(\alpha \cdot W) \times (\alpha \cdot H)$. EAR is defined as the proportion of attention mass concentrated within an annular ring (or border band) $\mathcal{E}$. The edge mask $\mathcal{E}$ is a band extending inwards from the image boundary with a width of $\beta \cdot \min(H, W)$. In our case, $\alpha = 0.12, \beta = 0.08$. The metrics are calculated as follows:

$$\text{CBI} = \frac{\sum_{(i,j)\in\mathcal{C}} \mathcal{H}^{(\ell)}(i,j)}{\sum_{(i,j)} \mathcal{H}^{(\ell)}(i,j)}, \tag{4.5}$$

$$\text{EAR} = \frac{\sum_{(i,j)\in\mathcal{E}} \mathcal{H}^{(\ell)}(i,j)}{\sum_{(i,j)} \mathcal{H}^{(\ell)}(i,j)}. \tag{4.6}$$

Moreover, to complement spatial bias measurement, we assess whether attention aligns with clinically relevant regions by introducing the **Attention-Lesion Overlap (Attn-IoU)** metric. It compares the attention heatmap $\mathcal{H}^{(\ell)}(i,j)$ with the binary lesion mask Ω. Attn-IoU measures the fraction of total attention mass that falls inside Ω:

$$\text{Attn-IoU} = \frac{\sum_{(i,j)\in\Omega} \mathcal{H}^{(\ell)}(i,j)}{\sum_{(i,j)} \mathcal{H}^{(\ell)}(i,j)}. \tag{4.7}$$

The lesion mask Ω is constructed from report-based lesion points: we first convert each point into a 2D Gaussian bump on the same $H \times W$ grid, max-pool these bumps into a continuous saliency map, and then binarize it by thresholding at a fixed fraction η of its maximum value (we use $\eta = 0.5$ in all experiments). The exact construction of the underlying Gaussian spatial prior is detailed in Section 4.2.2.

Using these metrics, we conduct a layer-wise analysis of the native autoregressive model's attention distribution, as shown in Figure 4.2. Early layers exhibit high CBI and EAR, indicating strong corner- and edge-focused attention driven by easy, high-contrast structures rather than pathology. As depth increases, this peripheral bias temporarily weakens as CBI/EAR decreases and Attn-IoU rises, suggesting partial integration of semantic cues that redirect attention toward lesion regions. In deeper layers, however, CBI and EAR increase again while Attn-IoU falls, because the decoder prioritizes global

structure completion and border continuity during final reconstruction, pulling attention back to edges at the expense of lesion focus. This pattern motivates us to import prior knowledge to mitigate the bias issue.

4.2.2 2D Gaussian Spatial Prior from VLM

To obtain proper prior knowledge, we leverage the strong multimodal understanding and alignment capabilities of a general-purpose Vision-Language Model (VLM). In our method, we use the offline model Qwen2.5-VL-7B-Instruct [2], which avoids leaking patients' privacy information. Given the prior image I_A and the progression description D, the VLM predicts N_p lesion-relevant coordinates in the original pixel space, denoted by $\mathbf{p}_k = (x_k, y_k)$, where $x_k \in [0, W_{\text{img}} - 1)$ and $y_k \in [0, H_{\text{img}} - 1)$.

Since the autoregressive decoder operates on a discrete visual-token grid rather than the full image lattice, we project each pixel-space coordinate onto the reference token grid of size $H \times W$ (here $H = W = 32$). The projected token-grid coordinates are defined as

$$\tilde{x}_k = \left\lfloor \frac{x_k}{W_{\text{img}}} W \right\rfloor, \qquad \tilde{y}_k = \left\lfloor \frac{y_k}{H_{\text{img}}} H \right\rfloor. \tag{4.8}$$

Here, the Gaussian prior is constructed on the discrete visual-token lattice associated with the reference image in the autoregressive decoder, obtained by projecting the VLM-predicted pixel coordinates onto that lattice.

Each projected point is then assigned a 2D Gaussian with $\sigma = \text{ratio} \cdot \min(H, W)$ (we use ratio $= 0.10$). For each token-grid cell (i, j), we compute its association with all projected points. Since a cell may be close to multiple points, we use point-wise max aggregation:

$$R_{\text{tok}}(i, j) = \max_{k \in \{1, \ldots, N_p\}} \exp\left(-\frac{(i - \tilde{y}_k)^2 + (j - \tilde{x}_k)^2}{2\sigma^2} \right), \tag{4.9}$$

where σ^2 controls the spatial spread of the Gaussian kernel on the token grid. This max-aggregation ensures that the value of $R_{\text{tok}}(i, j)$ is determined by the nearest projected lesion point, thereby creating a token-grid spatial prior that guides the subsequent autoregressive generation process.

Finally, R_{tok} is flattened into a 1D vector $\mathbf{r}$ of length $H \times W$. This vector $\mathbf{r}$ serves as a positional bias over the reference visual keys within decoder causal self-attention. It provides a score representing the strength of association between each visual token's spatial region and the key information derived from the text.

4.2.3 Gaussian-Biased Causal Attention

Recall that in standard decoder self-attention, the attention scores (logits) are computed as:

$$S = \frac{QK^T}{\sqrt{d_k}}. \tag{4.10}$$

Our Gaussian-Biased Causal Attention introduces B_{gaussian} into the standard decoder self-attention. B_{gaussian} is constructed from the prior vector $\mathbf{r}$ (aligned with I_A), extended (via zero-padding) to the input sequence length. This yields a per-position bias vector $\mathbf{a}$ over keys. We then broadcast $\mathbf{a}$ across all heads and query positions to form a bias tensor B_{gaussian} with the same shape as S. We apply the Gaussian bias additively to the pre-softmax attention logits and then enforce causality with the mask M, yielding the final attention logits S':

$$S' = S + s \cdot B_{\text{gaussian}} + M, \tag{4.11}$$

where causality and visibility are enforced via a mask matrix M, s is a learnable parameter which is obtained via a 2-layer MLP. The learnable strength s is initialized to a very small value, and an upper bound is set to prevent excessive bias. We choose to apply the bias at the logits level, rather than performing linear mixing after Softmax, to avoid the numerical instability and gradient control difficulties that can arise from non-linear compression. This Gaussian bias can be flexibly added to specific decoder layers of the model. Our model architecture is shown in Figure 4.1.

Equation 4.11 defines the GBCA-modified attention logits for the decoder in both regimes. The distinction between teacher forcing and inference does not lie in the attention rule itself, but in the source of the visible follow-up prefix: $X_{B,<n}$ during training/analysis versus $\hat{X}_{B,<n}$ during autoregressive inference.

4.2.4 Training and Inference Pipelines

Training (teacher forcing). During training, the decoder receives the token sequence $\{X_A, X_D, X_B\}$, where X_B is the ground-truth follow-up token sequence. The next-token loss is computed only on the follow-up positions:

$$L_{\text{CE}} = -\sum_{n=1}^{N_B} \log p(X_B[n] \mid X_A, X_D, X_{B,<n}) \,. \tag{4.12}$$

Thus, at the prediction of the n-th target token, the visible follow-up prefix is the ground-truth prefix $X_{B,<n}$. This teacher-forcing regime is used both for model optimization and for the attention-bias analysis in Section 4.2.1.

Inference (autoregressive decoding). At inference time, the ground-truth follow-up sequence is unavailable. The decoder is initialized with the reference-image tokens X_A and text tokens X_D only. A VLM first processes I_A and D to predict lesion coordinates, from which the token-grid Gaussian prior R_{tok} and the bias tensor B_{gaussian} are constructed. The decoder then autoregressively predicts

$$\hat{X}_B[n] \sim p\big(\cdot \mid X_A, X_D, \hat{X}_{B,<n}\big) \,, \tag{4.13}$$

Table 4.1: Performance comparison of image generation quality and downstream classification.

Method	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
BioMedJourney [24]	47.5307	35.01	0.4479	12.09	0.6166	0.7423
PIE [40]	49.0129	35.13	0.4982	12.92	0.6479	0.7764
CXR-IRGen [57]	44.1238	32.38	0.4760	12.26	0.6230	0.7544
ProgEmu [43]	35.0192	35.45	0.4884	13.22	0.7153	0.8132
LeGend (Ours)	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**

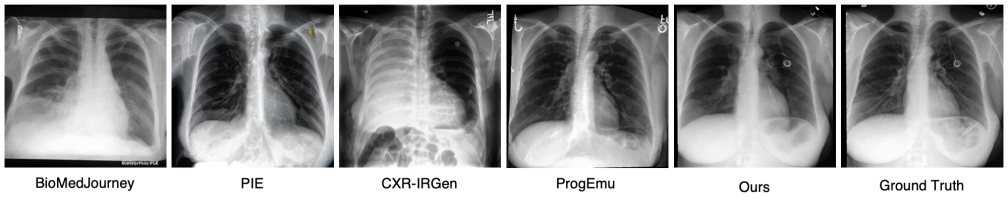


Figure 4.4: Visual comparison of follow-up CXR generation. Our LeGend shows higher similarity with regard to the ground truth image.

and each newly generated token is appended back to the visible prefix for the next step. After all follow-up tokens are generated, $\hat{X}_B$ is detokenized into the generated follow-up image $\hat{I}_B$. During this process, the Gaussian bias is applied only to the reference-image visual keys derived from I_A, while causality over the follow-up prefix is preserved by the same causal mask as in training.

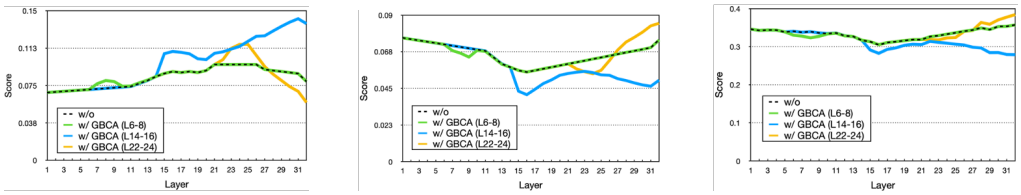
4.3 Experiments

4.3.1 Dataset

All experiments were conducted on ICG-CXR, the longitudinal counterfactual CXR dataset introduced in ProgEmu [43]. It is derived from MIMIC-CXR [32] and CheXpert-Plus [9] by selecting PA-view exam pairs with a reference/follow-up interval under 100 days, yielding 11,439 pairs from 7,388 patients. Each sample includes a reference and follow-up CXR, the corresponding reports, and an LLM-generated progression description. All image pairs are rigidly registered to reduce pose variation. As noted in [43], ICG-CXR is currently the only public dataset designed for longitudinal, report-conditioned counterfactual CXR generation, making it the standard and only feasible benchmark for this task.

4.3.2 Evaluation Metrics

To comprehensively evaluate GBCA, we assess both **attention behavior**, **follow-up image quality** and **classifier-based downstream task evaluation**, covering spatial alignment, visual realism, semantic consistency, structural fidelity, and downstream clinical



(a) Attn-IoU ↑ **(b)** CBI ↓ **(c)** EAR ↓

Figure 4.5: Attention profiling across layers highlights the mid-layer region where GBCA most effectively improves lesion focus.

utility.

Attention alignment metrics. To evaluate whether GBCA improves lesion-relevant attention, we report **Attn-IoU**, together with **CBI** and **EAR** (defined in Section 4.2.1). **Attn-IoU** measures the overlap between the model's reference-side attention map and the lesion region, where higher values indicate better lesion-focused attention. In contrast, **CBI** and **EAR** quantify undesired corner-biased and edge-biased attention, respectively, so lower values are preferred.

Generation quality metrics. For follow-up image generation quality, we use four metrics. **FID** [26] measures the distributional similarity between generated and real follow-up images, with lower values indicating more realistic generation. **CLIP-T** evaluates the semantic consistency between the generated image and the progression text using a pretrained BiomedCLIP [75] encoder, where higher values indicate better multimodal alignment. **MS-SSIM** [67] and **PSNR** measure structural and pixel-level similarity between the generated follow-up image and the paired ground-truth follow-up image. These metrics are particularly important in longitudinal medical image generation, where the background anatomy and patient-specific structure should remain largely consistent while lesions evolve locally.

Classifier-based clinical utility metrics. Visual realism alone does not guarantee clinical correctness. Therefore, we further evaluate whether generated lesions remain clinically identifiable using a pretrained downstream disease classifier. Specifically, we report **AUC** [5] and **F1 score** computed from classifier predictions on generated follow-up images against the pathology labels of the ground-truth follow-up studies. Higher values indicate that the generated images preserve diagnostically meaningful pathology cues that can be recognized by a standard CXR analysis model.

Details of the downstream disease classifier. For downstream evaluation, we use the public TorchXrayVision model "densenet121-res224-all" [14] without any fine-tuning on our generated images or icg-cxr dataset. This model is based on a DenseNet-121 backbone adapted for single-channel radiographs and operates at a fixed input resolution of 224×224. It outputs logits over the default 18-pathology TorchXrayVision label

space, including pathologies such as atelectasis, consolidation, effusion, pneumonia, cardiomegaly, and lung opacity. In our experiments, AUC and F1 are computed on the ICG-CXR pathology labels that overlap with the TorchXrayVision output categories. The classifier itself was originally trained in a multi-label setting on a large mixture of public CXR datasets, including NIH ChestX-ray14, PadChest, CheXpert, MIMIC-CXR, OpenI, the Google relabeled NIH set, and the RSNA Pneumonia dataset.

Summary. Together, these metrics provide a complementary evaluation protocol: attention metrics assess whether GBCA improves lesion-relevant reference attention; generation metrics evaluate realism, semantic consistency, and structural fidelity; and classifier-based metrics assess whether the synthesized follow-up images remain clinically meaningful. All models are evaluated on the same test set under identical preprocessing, registration, and resolution settings.

Table 4.2: Last-layer attention statistics under different GBCA injection locations.

Method	Attn-IoU↑	CBI↓	EAR↓
w/o GBCA	0.0790	0.0750	0.3574
w/ GBCA(L6-8)	0.0789	0.0746	0.3573
w/ GBCA(L14-16)	**0.1372**	**0.0511**	**0.2795**
w/ GBCA(L22-24)	0.0579	0.0854	0.3843

4.3.3 Implementation Details

We fine-tune the Emu3 backbone using LoRA [30] adapters inserted into all self-attention projection layers, namely q_proj, k_proj, v_proj, and o_proj. Each LoRA module uses a rank of $rank = 32$, scaling factor $lora\_alpha = 64$, dropout = 0.05, and no additional bias terms. Training is launched on 2× NVIDIA RTX A6000 GPUs for 6k optimization steps, together with gradient checkpointing for memory efficiency. Batch size is 16 with gradient accumulation. The optimizer is Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-6}$, weight decay = 0.1, and gradient clipping at $\|g\|_2 \leq 5.0$. The base learning rate is 1×10^{-5} with a cosine schedule down to a minimum of 1×10^{-6}, preceded by 30 warm-up steps.

4.3.4 Results

Comparison with existing methods. We compare our method against representative diffusion-based (BioMedJourney [24], PIE [40], CXR-IRGen [57]) and autoregressive (ProgEmu [43]) models on the ICG-CXR dataset. The performance of all baseline models in table 4.1 come from re-running the authors' released code (and public weights when available) on ICG-CXR test set under a unified protocol (same resolution and preprocessing).

Table 4.1 reports the quantitative results of both the CXR generation quality (left) and the downstream classification performance (right). As seen, our LeGend achieves

Table 4.3: Performance across different GBCA injection locations.

Method	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
Ours (L6-8)	34.7979	35.50	0.4910	13.30	0.7160	0.8140
Ours (L14-16)	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**
Ours (L22-24)	40.2545	33.20	0.4710	12.60	0.6127	0.7250
Ours (L14-24)	41.8016	32.49	0.4592	12.88	0.5721	0.6817

the best overall generation quality across all metrics, showing the lowest FID and the highest PSNR, MS-SSIM, and CLIP-T scores. Diffusion models such as BioMedJourney [75], PIE [40] and CXR-IRGen [57] produce visually realistic outputs but exhibit lower PSNR/MS-SSIM values, indicating smoother textures and weaker lesion contrast due to their stochastic noise-sampling process. Autoregressive methods like ProgEmu [43] preserve better structural continuity, yet our GBCA further enhances both global fidelity and local precision by steering causal attention toward lesion regions through the proposed Gaussian bias.

More importantly, to assess whether the generated follow-up CXRs encode clinically relevant information, we further conduct a downstream disease-classification task using a pretrained ResNet-based thoracic disease classifier [14] to predict 14 pathologies. GBCA again achieves the highest F1 and AUC, confirming that its lesion-aware attention improves not only perceptual quality but also the diagnostic consistency of generated images.

In addition to the quantitative results, Figure 4.4 provides visual evidence of method behavior. BioMedJourney produces smooth but overly blurred regions, weakening lesion depiction and suppressing subtle progression cues; anatomical boundaries (e.g., lungs, diaphragm) appear indistinct. PIE preserves global structure but often shows oversaturation, spatial shifts, and sharpened edges, leaving noticeable editing artifacts. CXR-IRGen generates plausible CXRs but departs from the true follow-up, consistent with its lower PSNR and MS-SSIM. ProgEmu offers the sharpest anatomy among baselines and introduces localized pathological changes, though lesion positions may drift. In contrast, GBCA yields follow-up CXRs with fine anatomical detail, accurately localized lesion progression, and high radiological plausibility aligned with the textual description. These observations confirm that Gaussian-biased causal attention guides the autoregressive decoder toward clinically relevant regions while maintaining global realism.

4.3.5 Ablation Study

Prior injection location. To determine where the Gaussian spatial prior most effectively influences the decoder, we compare GBCA injected at different depths of the 32-layer transformer decoder. Because the injected bias propagates forward through the causal

stack, its final effect is best reflected in the last-layer attention statistics reported in Table 4.2. As shown, injecting the prior in shallow layers (6–8) produces almost no change compared with the baseline, whereas mid-layer injection (14–16) yields the largest improvement—Attn-IoU increases from 0.0790 to 0.1372, and both CBI and EAR drop substantially. In contrast, deep-layer injection (22–24) worsens all three metrics, indicating disrupted semantic focus and stronger corner/edge bias.

Figure 4.5 provides the layer-wise explanation for this behavior. Recall that for the vanilla model, shallow layers show high CBI/EAR, mid-layers achieve maximal lesion alignment (highest Attn-IoU), and deep layers gradually shift attention back toward global structure (Figure 4.2). Consequently, the prior injected in shallow layers is overwritten by later layers, whereas injecting it at mid-depth (14–16) aligns with the model's natural semantic grounding stage and allows the Gaussian spatial prior to propagate cleanly forward. Injecting it too late (22–24) interferes with global refinement, leading to degraded attention quality.

Beyond attention statistics, Table 4.3 evaluates how different injection locations affect image quality and downstream classification. The trends closely follow those in Table 4.2 and Figure 4.5: injecting the prior in shallow layers (L6–8) yields only marginal gains, while mid-depth injection (14–16) delivers the best overall results—lowest FID (26.12), highest PSNR (15.87), MS-SSIM (0.6617), and CLIP-T (38.27), together with the strongest classifier performance (AUC = 0.7559, F1 = 0.8402). This mid-layer region offers the best balance between semantic guidance and reconstruction fidelity. In contrast, deep injection (L22–24) over-constrains the decoder, leading to blurred structures and reduced AUC/F1, consistent with its degraded attention patterns. Excessive injection across many layers (e.g., L14–24) further degrades performance, as same-direction bias accumulates and induces inter-layer interference, weakening late-stage refinement, reducing global consistency, and ultimately lowering sharpness and classifier accuracy.

Prior injection layer number. Given that mid-layer injection (14–16) yields the best overall location, we further perform a fine-grained ablation to identify the most effective layer(s) within this region. As shown in Table 4.4, injecting the prior into a single mid-layer (L15) already improves both visual and diagnostic metrics over the baseline. Notably, the layer L15 emerges as the most effective single-layer injection point, achieving the best FID(28.1581), CLIP-T(37.69), PSNR(0.6509), MS-SSIM(14.78), AUC(0.7543), and F1(0.8385). When the prior is applied to adjacent layers jointly (L14-15 or L15-16), performance consistently improves across both generative and classification metrics, indicating that a slightly broader mid-layer window allows the spatial prior to propagate more stably through the network. Extending it across three consecutive layers (L14–16) leads to optimal configuration, achieving the best performance.

Effect of Gaussian σ's ratio. The Gaussian standard deviation σ controls the spatial spread of the injected prior, defined as $\sigma = \text{ratio} \cdot \min(H, W)$. table 4.5 shows how different ratios affect performance. A very small σ (ratio = 0.05) produces an overly

Table 4.4: Effect of injecting GBCA into different mid-depth layer combinations.

Method	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
Ours (L14)	28.6369	36.86	0.6131	14.24	0.7436	0.8270
Ours (L15)	28.1581	37.69	0.6509	14.78	0.7543	0.8385
Ours (L16)	29.1533	36.43	0.5928	13.92	0.7452	0.8287
Ours (L14-15)	27.0684	38.18	0.6581	15.77	0.7554	0.8396
Ours (L15-16)	27.3550	38.12	0.6557	15.31	0.7550	0.8393
Ours (L14-16)	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**

Table 4.5: The ablation of Gaussian σ's ratio.

Ratio	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
0.05	31.5546	38.36	0.6527	15.71	0.7381	0.8333
0.1	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**
0.2	41.1478	37.90	0.6603	15.68	0.7466	0.8206

concentrated prior, limiting spatial generalization and demanding more accurate coordinates from the VLM. Conversely, a very large σ (ratio = 0.2) diffuses the prior too broadly, weakening ROI localization and introducing redundant spatial bias. The moderate setting (ratio = 0.1) achieves the best balance between focus and coverage, yielding the best generation quality and capturing diseases. Notably, even suboptimal ratios (0.05 and 0.2) outperform competing methods in disease classification, showing that GBCA reliably enhances disease-relevant feature capture.

Different VLMs for Gaussian spatial prior generation. Table 4.6 compares GBCA when lesion coordinates are obtained from two vision–language models, Qwen2.5-VL and Llama-3.2-Vision. Both variants yield consistent improvements over the baseline, indicating that GBCA is effective across different VLM providers. Qwen2.5-VL achieves slightly better image-generation quality (e.g., FID and CLIP-T), while Llama-3.2-Vision remains competitive on downstream classification metrics. Overall, the relatively small gap across metrics suggests that GBCA is reasonably stable with respect to the choice of VLM-derived spatial prior. We note that this comparison evaluates downstream robustness across alternative VLM priors, rather than the absolute localization accuracy of the priors themselves, since our dataset does not provide ground-truth lesion masks

Table 4.6: Ablation of different VLMs for spatial-prior generation.

VLM	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
Qwen2.5-VL	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**
Llama-3.2-vision	29.0563	37.68	0.6396	15.54	0.7742	0.8370

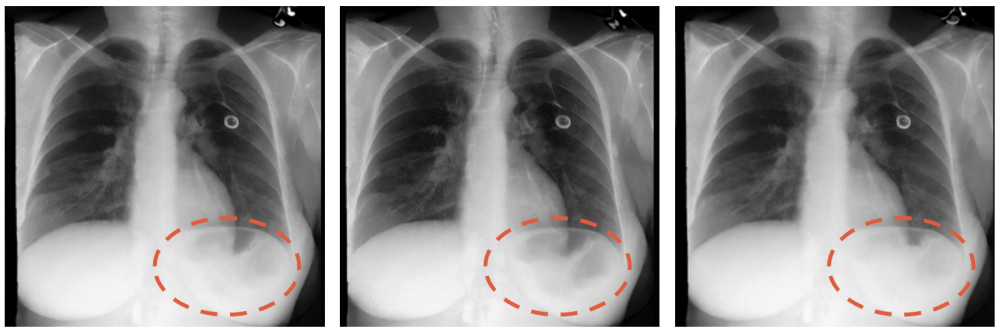


Figure 4.6: Ablation of different VLMs for spatial-prior generation. This comparison evaluates the robustness of GBCA across alternative VLM-derived spatial priors.

or independent lesion-location annotations for direct IoU evaluation.

4.4 Generality of GBCA Across Autoregressive Backbones

To evaluate whether GBCA generalizes beyond our primary autoregressive backbone, we apply GBCA to EditAR [45], a unified AR generator that performs next-token prediction to synthesize edited images from conditioning inputs and text prompts. After fine-tuning EditAR on the ICG-CXR dataset, we freeze all EditAR parameters and train only the GBCA module, injected into its 18th (middle) transformer layer. As reported in Table 4.7, GBCA improves both image quality and downstream classification metrics after only 500 training steps, demonstrating its effectiveness across different autoregressive architectures.

Layer-wise attention diagnosis (EditAR). Before injecting GBCA, we conduct a layer-wise diagnosis on the fine-tuned EditAR backbone (without Gaussian prior; denoted as *no_gaussian*) by measuring three attention-related metrics: Attn-IoU (lesion-region attention overlap), Corner Bias Index (CBI), and Edge Attention Ratio (EAR). As shown in Figure 4.7, EAR stays consistently high across layers and exhibits a pronounced peak around the 18th (middle) transformer layer, indicating the strongest tendency to allocate attention to image borders/edges at this stage. In contrast, Attn-IoU remains comparatively low and changes mildly across layers, suggesting limited lesion-aligned attention under the vanilla backbone. This layer-wise pattern motivates our design choice to inject GBCA into the 18th layer, where the edge-dominant attention is most evident and thus the lesion-centric Gaussian bias is expected to yield the largest corrective effect.

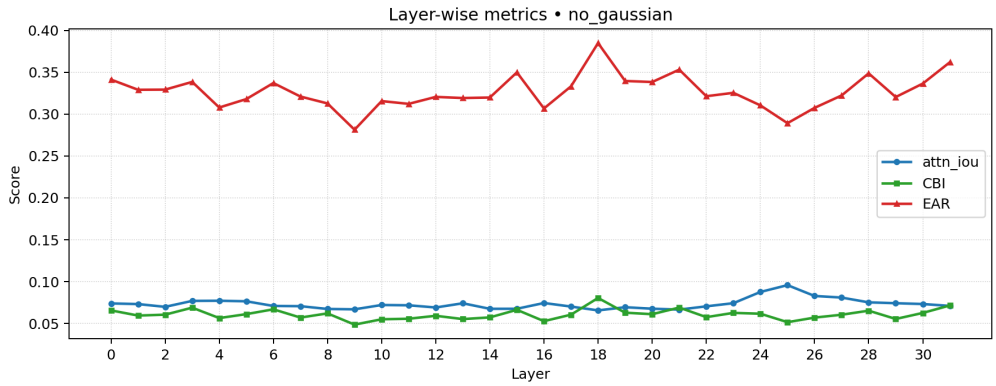


Figure 4.7: Layer-wise attention metrics of EditAR under the *no_gaussian* setting. We report Attn-IoU, CBI, and EAR across transformer layers. The marked EAR peak around the 18th layer indicates a strong edge-attention tendency, motivating GBCA injection at this middle layer.

Table 4.7: Performance comparison of EditAR with/without GBCA.

EditAR	**Generation Quality**				**Classifier Quality**	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
w/o GBCA	39.5823	0.3557	0.5973	14.08	0.7521	0.7875
w/ GBCA	**33.5722**	**0.3621**	**0.6313**	**14.74**	**0.7872**	**0.8040**

4.5 Visualization of Diverse VLM-Predicted Spatial Priors

Figure 4.8 compares the lesion-location points predicted by two different vision–language models, Qwen2.5-VL and Llama-3.2-Vision. The two models produce noticeably different coordinate sets, reflecting natural variability in VLM-based spatial localization. Despite this diversity in predicted priors, our method achieves stable performance across both VLMs (see Table 4.6). This qualitative comparison illustrates the variability of the input priors, while the quantitative results in Table 4.6 show that GBCA remains effective across these alternative VLM-derived spatial priors. Since our dataset does not provide ground-truth lesion masks, this comparison should be interpreted as a robustness analysis with respect to different prior sources, rather than a direct measurement of prior accuracy.

4.6 Prompts Design for VLM Point Annotation

Prompt Design for Point Annotation We use a deterministic instruction set for a VLM (Qwen2.5-VL-7B-Instruct in our case), which returns point coordinates marking the anatomical locations where changes occur (worsening, persistence, improvement, or resolution). The model does not diagnose; it only produces standardized positional annotations.

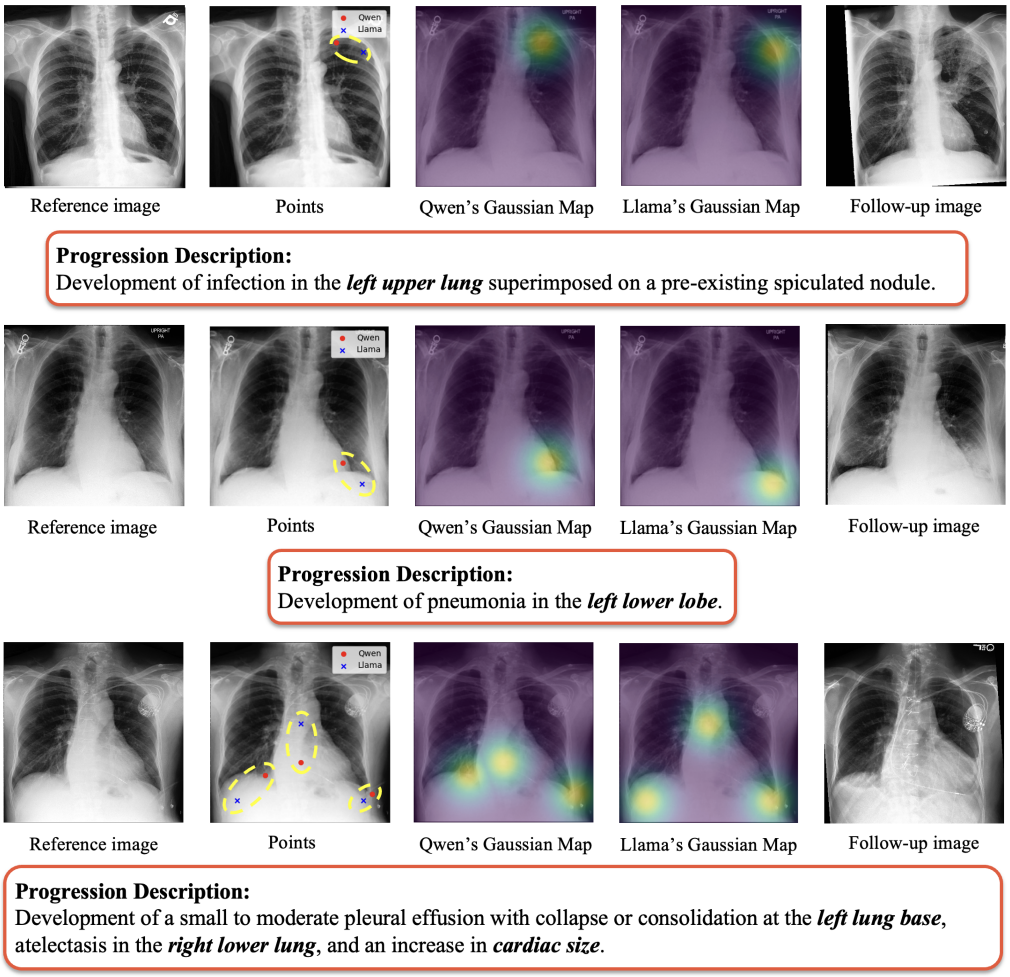


Figure 4.8: Visualization of spatial priors predicted by two VLMs (Qwen2.5-VL and Llama-3.2-Vision). The two models produce different lesion-coordinate sets, illustrating the natural variability in VLM-based localization. The corresponding quantitative results in Table 4.6 evaluate the robustness of GBCA across these alternative VLM-derived priors.

The design principles are as follow: (1). Centered on "change". Confidence reflects certainty where a change occurred, not disease severity. Improvements/resolution count as positive evidence when localized. Confidence ranges 1-5. The point with confidence lower than 3 is discarded. (2). Clinical laterality convention. Textual "left/right" refers to patient laterality; on standard PA/AP CXR, patient-left appears on image-right, and patient-right on image-left. (3). Coverage over minimality. The VLM model is required to generate as many points as needed to cover distinct described region (bilateral/diffuse patterns must annotate both lung). Based on the distance between the points, redundant points are merged to keep outputs compact and interpretable.

The **SYSTEM PROMPT** we used is as follow:

You are a medical AI assistant that converts chest X-ray (CXR) progression descriptions into pixel coordinates marking disease-relevant locations. You do not diagnose; you only produce standardized positional annotations.

Inputs

- A CXR image (pixel-based; infer width W and height H).
- Disease progression description.

Core principles

- Annotate *where changes occur* (worsening/persistence/improvement); severity is irrelevant.
- Output as many points as needed for distinct regions; avoid redundant points.
- Bilateral/diffuse findings $\rightarrow$ annotate both lungs.

Laterality (patient vs. image)

- "Left/right" refers to the *patient's* side.
- Patient-left = image-right; patient-right = image-left.

Localization rules

- Visually estimate lung fields and derive `BBOX_R` (patient-right/image-left) and `BBOX_L` (patient-left/image-right).
- Split each lung box into three equal-height zones (Upper/Middle/Lower); use zone centers as candidate points.
- If lung boxes are unreliable, use coarse priors: image-left centers at (0.30W,0.20H), (0.30W,0.50H), (0.30W,0.80H); image-right centers at (0.70W,0.20H), (0.70W,0.50H), (0.70W,0.80H).

- Extra-pulmonary targets: pleural effusion → costophrenic angle; pneumothorax → apicolateral pleura; cardiomegaly/mediastinum → cardiac/mediastinal center; devices/bone/soft-tissue findings → described tip/structure.

Confidence (output only $\{3, 4, 5\}$)

- 5: explicit side/zone (or lobe) *and* explicit change at same site.
- 4: clear side with partially specified zone, change consistent with text.
- 3: bilateral/diffuse or weakly localized change.
- Points with confidence 1 or 2 are discarded.

Point placement and deduplication

- Unilateral finding → at least one point on that side (zone per text).
- Bilateral/diffuse → at least one representative point per lung.
- Multifocal → one point per distinct focus/structure.
- Merge points closer than $\delta = 0.03 \cdot \min(W, H)$ into their average location, keeping higher confidence.

Output (strict JSON; no extra text)

```
{
  "points": [
    {"x": <int>, "y": <int>,
     "confidence": <int>}
  ]
}
```

Use `"points": []` if healthy or no location meets confidence ≥ 3.

Chapter 5

Longitudinal Chest X-ray Generation via Diffusion-based Model

5.1 Motivations and Contributions

Longitudinal chest X-ray (CXR) image generation—predicting future radiographs based on prior scans and textual disease progression descriptions—holds significant clinical value in monitoring disease development, evaluating treatment response, and supporting follow-up decisions. For example, generating a hypothetical follow-up image conditioned on a progression description could help clinicians visualize lesion evolution and preemptively assess outcomes. However, this task presents unique challenges. Lesion changes tend to be local and subtle, requiring generative models to faithfully preserve global thoracic structures while precisely depicting small-scale, clinically meaningful variations. Moreover, the generated image must be semantically aligned with the progression description, accurately reflecting changes in lesion size, shape, and location. Achieving both visual realism and clinical faithfulness demands strong spatial awareness and precise semantic control—capabilities current models often lack.

5.2 Methodology

We propose a novel framework for longitudinal chest X-ray (CXR) generation that effectively disentangles anatomical structural preservation from pathological evolution. The core objective is to synthesize a high-fidelity follow-up image I_B, conditioned on a reference image I_A and a textual progression description D_P. As illustrated in Figure 5.1, our pipeline is built upon a Diffusion Transformer (DiT) backbone and integrates the structural consistency of Stable Flow [1] with the precise lesion localization of our Gaussian-Biased Causal Attention (GBCA).

The overall framework operates in three distinct phases: (1) Offline Structural Calibration to identify layers responsible for anatomical integrity; (2) Semantic-to-Spatial

Mapping to convert textual descriptions into geometric guidance; and (3) Dual-Path Guided Generation to synthesize the final output via feature injection.

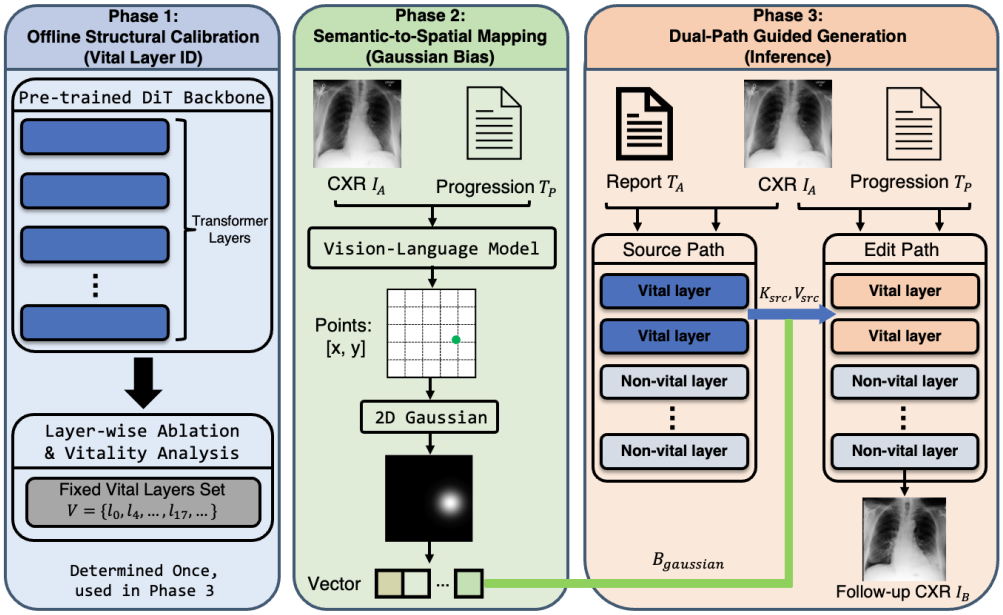


Figure 5.1: Overview of the proposed LeGend-Diffusion framework. The pipeline consists of three phases: **Phase 1** identifies Vital Layers ($\mathcal{V}$) offline via ablation analysis to lock anatomical structures. **Phase 2** leverages a Vision-Language Model (VLM) to map the progression description D_P into a Gaussian spatial prior B_{gaussian}. **Phase 3** performs dual-path inference, where the Edit Path generates the follow-up image I_B by retrieving structural keys/values (K_{src}, V_{src}) from the Source Path and integrating the Gaussian bias (B_{gaussian}) specifically at the identified Vital Layers.

5.2.1 Phase 1: Offline Structural Calibration (Vital Layer ID)

The first phase addresses the challenge of structural preservation in isotropic architectures. Unlike U-Net architectures, which possess explicit hierarchical scales (coarse-to-fine) due to pooling layers, Diffusion Transformers (DiTs) process images as flat sequences of tokens. This isotropic nature makes structural information diffuse across all layers, complicating the task of preserving patient-specific anatomy (e.g., rib cage, cardiac silhouette) during editing.

To pinpoint the layers encoding high-level geometric structure, we adopt the *Vitality Analysis* approach proposed in Stable Flow [1]. As illustrated in Figure 5.2, we define a layer's "vitality" by the perceptual impact of bypassing it.

Let $\mathcal{F}$ denote the complete pre-trained DiT model with L layers. We define a modified model, $\mathcal{F}_{\text{skip-}l}$, where the l-th layer is physically skipped via a residual bypass connection. To estimate the expected perceptual deviation, we generate N pairs of images. For the i-th sample, let $z_i \sim \mathcal{N}(0, \mathbf{I})$ be the initial Gaussian noise and p_i be a

randomly selected text prompt. Let $I_i = \mathcal{F}(z_i, p_i)$ be the image generated by the full model, and $I_i^{(l)} = \mathcal{F}_{\text{skip-}l}(z_i, p_i)$ be the image generated when layer l is skipped.

To quantify the perceptual deviation between these outcomes, we employ a robust semantic similarity metric, denoted as $\text{Sim}(\cdot, \cdot)$. We utilize the cosine similarity of deep features extracted by a frozen DINOv2 [47] encoder, denoted as $\Phi(\cdot)$. For any two images U and V, the metric is explicitly defined as:

$$\text{Sim}(U, V) = \frac{\Phi(U) \cdot \Phi(V)}{\|\Phi(U)\|_2 \|\Phi(V)\|_2}. \tag{5.1}$$

Pixel-wise MSE is unsuitable here as it is overly sensitive to high-frequency noise rather than structural semantics.

The vitality score $v(l)$ for layer l is then computed as the complement of the average similarity across all N samples:

$$v(l) = 1 - \underbrace{\frac{1}{N} \sum_{i=1}^{N} \text{Sim}\left(I_i, I_i^{(l)}\right)}_{\text{Average Similarity}}. \tag{5.2}$$

Intuitively, if skipping layer l results in a generated image $I_i^{(l)}$ that is very similar to the original I_i, the average similarity will be close to 1, resulting in a low vitality score $v(l) \approx 0$. Conversely, a low similarity implies a high structural deviation, indicating that layer l is "vital."

Finally, we define the fixed set of **Vital Layers** $\mathcal{V}$ using a predefined threshold τ_{vit}:

$$\mathcal{V} = \{l \in \{1, \ldots, L\} \mid v(l) \geq \tau_{vit}\}. \tag{5.3}$$

Empirical analysis on the FLUX.1 backbone reveals that vital layers are typically distributed at the input (early processing) and middle (semantic formation) blocks. These layers serve as "structural anchors," which we explicitly manipulate in the subsequent generation phase.

5.2.2 Phase 2: Semantic-to-Spatial Mapping (Gaussian Bias)

Standard attention mechanisms in DiTs often suffer from "corner-focus" bias or fail to localize small pathological changes described in text. This issue stems from the *modality gap*: the textual description T_P provides high-level semantic instructions (e.g., "right lower lobe"), while the visual generation operates on low-level latent tokens without explicit spatial grounding. To bridge this gap, we construct a lesion-specific spatial prior, transforming discrete semantic coordinates into a continuous, differentiable attention bias.

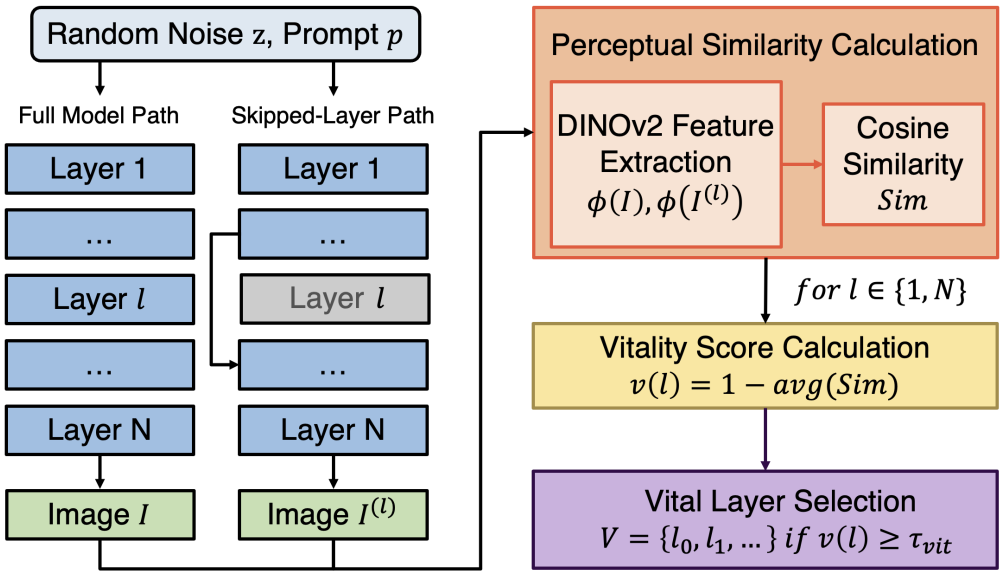


Figure 5.2: Workflow for Offline Structural Calibration (Phase 1).

Coordinate Prediction via VLM

We utilize a Generative Vision-Language Model (VLM), specifically Qwen2.5-VL [2], to act as a semantic interpreter. Given the reference image I_A and the progression description D_P, the VLM is prompted to perform visual grounding. It predicts a set of discrete coordinates $P = \{(x_k, y_k)\}_{k=1}^{N_p}$ defined in the original image pixel space $\mathbb{R}^{H \times W}$, where each point corresponds to the centroid of a region of interest (ROI) mentioned in the progression text (e.g., "enlarging opacity").

Latent Space Transformation and Gaussian Modeling

Since the Diffusion Transformer operates in a compressed latent space rather than the original image lattice, the pixel-space coordinates must be projected onto the latent grid. Let the image resolution be $(H_{\text{img}}, W_{\text{img}})$ and the latent resolution be (h, w). The projected latent coordinates are defined as

$$x_k^{\text{lat}} = \left\lfloor \frac{x_k}{W_{\text{img}}} w \right\rfloor, \qquad y_k^{\text{lat}} = \left\lfloor \frac{y_k}{H_{\text{img}}} h \right\rfloor. \tag{5.4}$$

Unlike Chapter 4, where the Gaussian prior is constructed on the discrete visual-token lattice used by the autoregressive decoder, here the prior is first defined on the 2D latent spatial grid of the DiT backbone and only then serialized into transformer tokens through flattening or patchification.

To integrate these sparse points into the dense attention mechanism, we model the spatial prior as a continuous 2D latent-space heatmap $R_{\text{lat}} \in \mathbb{R}^{h \times w}$. To avoid artificial hotspots caused by overlapping summations, we employ a max-aggregation of Gaussian

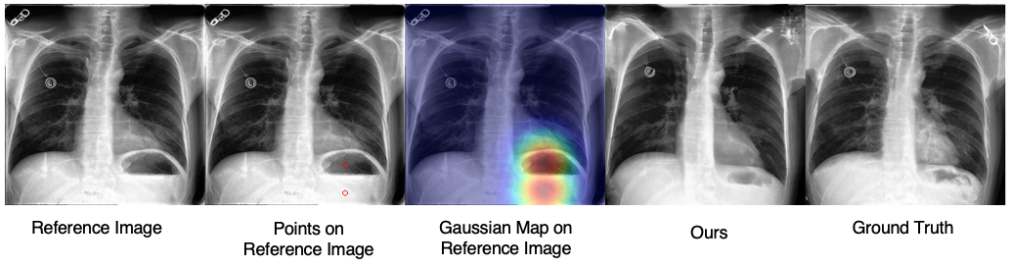


Figure 5.3: Visualization of Semantic-to-Spatial Mapping and Generation Results. This figure illustrates the intermediate spatial guidance signals and the final output. **(1) Reference Image:** The baseline CXR input (I_A). **(2) Points on Reference Image:** Discrete lesion coordinates (red circles) predicted by the VLM based on the progression description, identifying the anatomical ROI (e.g., left lower lobe). **(3) Gaussian Map:** The continuous Gaussian spatial prior (B_{gaussian}) derived from the discrete points, which serves as the attention bias B_{gaussian} injected into the DiT. **(4) Ours:** The final follow-up image generated by our full framework (LeGend-Diffusion), showing precise lesion synthesis. **(5) Ground Truth:** The real follow-up exam. Comparing the Gaussian Map with the generated outcome confirms that our GBCA mechanism effectively translates semantic coordinates into accurate visual pathology.

kernels. For each latent-grid cell (i, j), the response is defined as

$$R_{\text{lat}}(i,j) = \max_{k \in \{1,\dots,N_p\}} \exp\left(-\frac{(i - y_k^{\text{lat}})^2 + (j - x_k^{\text{lat}})^2}{2\sigma^2}\right), \tag{5.5}$$

where $\sigma = 0.1 \cdot \min(h, w)$ controls the spatial spread of the latent-space Gaussian prior.

Sequence Alignment and Bias Injection

The input to the DiT backbone is a concatenated sequence of text tokens and flattened image tokens. Let N_{txt} be the number of text tokens and $N_{img} = h \times w$ be the number of image tokens. The total sequence length is $L_{seq} = N_{txt} + N_{img}$.

To align the 2D spatial prior R with this 1D sequence structure, we perform the following operations:

1. **Flattening:** The latent-space heatmap R_{lat} is flattened into a visual bias vector $\mathbf{r}_{img} \in \mathbb{R}^{N_{fu}}$.
2. **Padding:** Since the spatial prior should not bias the text tokens (which lack spatial geometry), we pad the vector with zeros (neutral bias) for the text positions:

$$\mathbf{r}_{\text{seq}} = [\underbrace{0, \dots, 0}_{N_{\text{ref}}}, \underbrace{0, \dots, 0}_{N_{\text{txt}}}, \underbrace{\mathbf{r}_{\text{img}}^{(1)}, \dots, \mathbf{r}_{\text{img}}^{(N_{\text{fu}})}}_{N_{\text{fu}}}]^\top \in \mathbb{R}^{L_{\text{seq}}}. \tag{5.6}$$

3. **Broadcasting:** We construct the final bias matrix $B_{\text{gaussian}} \in \mathbb{R}^{L_{seq} \times L_{seq}}$. In our design, the bias acts on the *Keys* (K) of the attention mechanism. Therefore, we

broadcast $\mathbf{r}_{seq}$ across the query dimension:

$$B_{\text{gaussian}}[m, n] = \mathbf{r}_{seq}[n]. \tag{5.7}$$

This matrix ensures that when any token (whether text or image) attends to a visual token located at a lesion site (index n), the attention score is boosted by the Gaussian value.

Finally, to allow the model to adaptively control the influence of this prior, we introduce a learnable scalar s. This scalar is predicted by a lightweight 2-layer MLP, and the final term added to the attention logits is $s \cdot B_{\text{gaussian}}$.

5.2.3 Phase 3: Dual-Path Guided Generation (Inference)

The final phase executes the image generation via a synchronized dual-path inference strategy. This process is designed to navigate the trade-off between semantic fidelity (adhering to the progression text) and structural fidelity (preserving the patient's anatomy). The inference relies on the interaction between two denoising trajectories: the **Source Path** ($\mathcal{T}_{src}$) and the **Edit Path** ($\mathcal{T}_{edit}$).

Latent Initialization via DDIM Inversion

Before generation commences, we must ensure that the generative process starts from a noise distribution that structurally encodes the reference image I_A. Instead of sampling random Gaussian noise, we employ **DDIM Inversion** [60]. Given the reference image latent $z_0 = \mathcal{E}(I_A)$, we run the deterministic DDIM reverse process in the forward direction (from $\tau = 0$ to $\tau = K$) conditioned on the source text D_A. This yields a structured noise latent z_K, which serves as the shared starting point for both trajectories:

$$z_K^{(\text{src})} = z_K^{(\text{edit})} = \text{DDIM}_{\text{invert}}(z_0, D_A, \epsilon_\theta). \tag{5.8}$$

Sharing z_K ensures that both paths originate from the same global structural layout.

Synchronized Denoising Trajectories

During the generative denoising steps ($\tau = K \to 0$), the two paths evolve in parallel. Let $h_\tau^{(\text{src},l)}$ and $h_\tau^{(\text{edit},l)}$ denote the intermediate hidden states at layer l and diffusion step τ for the source and edit paths, respectively.

- **Source Path ($\mathcal{T}_{\text{src}}$):** Conditioned on D_A. Its primary role is to reconstruct the reference anatomy. It acts as a "memory bank," providing structural features (Keys and Values) to the edit path.
- **Edit Path ($\mathcal{T}_{\text{edit}}$):** Conditioned on the progression description D_P. Its goal is to generate pathology-consistent follow-up changes defined by D_P while retrieving structural constraints from $\mathcal{T}_{\text{src}}$.

Feature Injection and Attention Modulation

The core innovation lies in how these two paths interact within the DiT backbone. This interaction is strictly confined to the set of **Vital Layers** ($\mathcal{V}$) identified in Phase 1. For a given layer l, the Self-Attention (SA) mechanism in the Edit Path is modified as follows:

Case 1: Independent Generation (Non-Vital Layers, $l \notin \mathcal{V}$). In layers identified as non-structural (typically responsible for fine textures or background noise), the two paths remain decoupled. The Edit Path relies solely on its own projections to refine local details consistent with the progression text:

$$\text{Attn}_{edit}^{(l)} = \text{Softmax}\left(\frac{Q_{edit}^{(l)}(K_{edit}^{(l)})^\top}{\sqrt{d_k}} + M\right) V_{edit}^{(l)}, \tag{5.9}$$

where Q, K, V are projected from $h_\tau^{(\text{edit},l)}$. This independence allows the model to synthesize new textures (e.g., consolidation patterns) that do not exist in the source image.

Case 2: Structure-Locked Lesion Generation (Vital Layers, $l \in \mathcal{V}$). In Vital Layers, which control global geometry (e.g., rib alignment, organ shape), we impose strict constraints. We propose the **Gaussian-Biased Vital-Layer Attention (GBCA)** mechanism, which performs two simultaneous operations:

1. **Structural Feature Injection (SFI):** We replace the Keys (K) and Values (V) of the Edit Path with those computed from the Source Path. Since K and V in self-attention determine the content and layout of the retrieved features, this operation "locks" the generated anatomy to match the reference image I_A. 2. **Spatial Saliency Modulation:** To prevent the Source Keys from completely suppressing the new lesion (since the lesion is absent in I_A), we inject the Gaussian bias B_{gaussian} (from Phase 2) into the attention logits.

The unified formulation for the Edit Path attention at Vital Layers is:

$$\text{Attn}_{\text{edit}}^{(l)} = \text{Softmax}\left(\underbrace{\frac{Q_{\text{edit}}^{(l)}(K_{\text{src}}^{(l)})^\top}{\sqrt{d_k}}}_{\text{Structural Context}} + \underbrace{s(\tau)\cdot B_{\text{gaussian}}}_{\text{Lesion Guidance}} + M\right) V_{\text{src}}^{(l)}. \tag{5.10}$$

where $Q_{\text{edit}}^{(l)}$ encodes the semantic request for the new pathology (driven by D_P). $K_{src}^{(l)}, V_{src}^{(l)}$ provide the anatomical blueprint from the reference scan. M is the standard attention mask (handling text/image token separation). $s(\tau)$ is a time-dependent learnable scalar, parameterized by a small MLP taking the timestep τ as input.

Mechanism Interpretation: The term $s(\tau) \cdot B_{\text{gaussian}}$ acts as a "soft spotlight." It biases the attention logits such that the query tokens in the Edit Path are encouraged to attend strongly to the specific spatial regions defined by the VLM, even though the structural Keys (K_{src}) come from the healthy reference. The scalar $s(\tau)$ allows the model

to dynamically adjust this guidance strength—typically learning to apply stronger guidance during the early semantic formation stages (large τ) and relaxing it during texture refinement (small τ). This synergy enables the precise synthesis of pathological changes within an invariant structural context.

5.3 Experiments

5.3.1 Dataset and Implementation Details

Dataset and Preprocessing. Consistent with the autoregressive experiments in Chapter 4, we conduct all evaluations on the ICG-CXR dataset [43], a derived subset of MIMIC-CXR specifically curated for longitudinal analysis. This dataset comprises 11,439 paired chest X-ray exams (reference and follow-up) from 7,388 unique patients, along with temporal progression descriptions generated by LLMs. To adapt the data for the FLUX.1-dev architecture, all radiographs are resized to 256×256 resolution using bicubic interpolation and normalized to the range $[-1, 1]$.

Implementation Details. Our framework is built upon FLUX.1-dev, a state-of-the-art Diffusion Transformer (DiT) with 12 billion parameters. Given the substantial domain shift between natural images (on which FLUX.1 is pre-trained) and medical radiographs, we adopt a **Two-Stage Training Strategy** to balance domain adaptability with structural preservation.

- **Stage 1: Domain Adaptation via LoRA.** Directly applying the frozen FLUX.1 model to CXR generation yields suboptimal results due to the domain gap. To address this, we first fine-tune the model on the ICG-CXR training set using **Low-Rank Adaptation (LoRA)** [30]. We inject LoRA adapters (rank $r = 16$, $\alpha = 32$) into the query, key, value, and output projection layers of the attention blocks. This stage adapts the model to the grayscale distribution and anatomical textures of X-rays while keeping the massive pre-trained backbone weights frozen.
- **Stage 2: Optimization for Longitudinal Editing.** In the second stage—which constitutes our proposed framework—we freeze the domain-adapted backbone (including the LoRA weights learned in Stage 1) to preserve its learned generative prior. We then introduce the Stable Flow mechanism and optimize *only* the newly introduced components:
  1. The **GBCA injection parameters**, specifically the projection matrices within the Vital Layers ($\mathcal{V}$) and the MLP governing the scalar $s(\tau)$.
  2. The learnable scalar s is initialized to 0.01 to ensure a gradual introduction of the spatial bias, preventing sudden disruptions to the attention maps.
- **Vital Layers Configuration:** Based on the offline vitality analysis (Section 5.2.1), we explicitly target layers $\mathcal{V} = \{0, 1, 2, 17, 18, 25, 28, 53, 54, 56\}$ for feature injection.

- **Training Hyperparameters:** The model is trained using the **AdamW** optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.01. We employ a constant learning rate of $1e-5$. The training runs for 10 epochs on 2 $\times$ NVIDIA A6000 GPUs with a global batch size of 32 (using gradient accumulation). During inference, we use the DDIM sampler with 50 steps.

5.3.2 Evaluation Metrics

Following the protocol defined in Section 4.3.2, we evaluate generated follow-up images from two complementary perspectives: (1) **generation quality** using FID and CLIP-T, which reflect realism and text-image consistency. And using MS-SSIM and PSNR, which measure similarity to the paired ground-truth follow-up image; and (2) **clinical utility** using downstream classifier-based AUC and F1, which assess whether the generated pathology patterns remain diagnostically recognizable.

This combination is particularly important for longitudinal CXR generation. A model may produce visually plausible images yet fail to preserve patient-specific anatomy or synthesize clinically meaningful lesion progression. Therefore, improvements across these metrics together provide stronger evidence that the generated follow-up images are not only realistic, but also structurally faithful and clinically relevant.

5.3.3 Quantitative Results

Comparison with State-of-the-Art. Table 5.1 presents a comprehensive quantitative comparison against both Autoregressive and Diffusion-based baselines.

In terms of **Generation Quality**, our proposed framework demonstrates superior performance, consistent with the visual comparisons shown in Figure 5.4. Compared to baseline diffusion models (BioMedJourney [24], PIE [40], CXR-IRGen [57]), our method achieves a drastic reduction in FID, indicating a distribution much closer to real medical images. More importantly, we observe a substantial leap in structural similarity metrics (MS-SSIM and PSNR). While standard diffusion models often degrade the anatomical integrity of the reference image, our approach maintains structural fidelity at a level comparable to the ground truth, effectively mitigating the "identity loss" problem.

Regarding **Clinical Utility**, our method outperforms all competing approaches in downstream pathology classification. The improvements in AUC and F1 scores suggest that the generated lesions are not only visually realistic but also contain the necessary diagnostic features to be correctly recognized. This confirms that the dual-path mechanism successfully balances visual fidelity with semantic correctness.

Advantage over Autoregressive Approaches. When compared to the autoregressive LeGend model (Chapter 4), our diffusion-based approach yields comparable semantic alignment (CLIP-T) but offers a distinct advantage in pixel-level structural preservation. This aligns with the theoretical expectation that diffusion models, when properly

constrained, can model fine-grained textures and boundaries better than token-based autoregressive models.

Furthermore, our approach demonstrates a significant advantage in inference efficiency. While both architectures leverage a Transformer backbone, their underlying generative paradigms differ fundamentally: AR models follow a sequential token-by-token generation process, resulting in a computational complexity that scales linearly with the sequence length (typically $O(N)$ steps for N patches). In contrast, our Diffusion Transformer (DiT) employs a parallel denoising mechanism, enabling high-quality image synthesis within a fixed and significantly smaller number of sampling iterations. In our empirical evaluation on a test dataset of 790 samples for 256×256 image generation, the AR model required 9 hours to complete the inference, whereas our DiT-based framework finished the same task in only 45 minutes. Consequently, this framework achieves a $12\times$ speedup in inference latency compared to AR models, which are often bottlenecked by the cumulative latency inherent in serial decoding.

Table 5.1: Performance comparison of image generation quality and downstream classification. Best results are highlighted in **bold**.

Method	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
BioMedJourney [24]	47.5307	35.01	0.4479	12.09	0.6166	0.7423
PIE [40]	49.0129	35.13	0.4982	12.92	0.6479	0.7764
CXR-IRGen [57]	44.1238	32.38	0.4760	12.26	0.6230	0.7544
ProgEmu [43]	35.0192	35.45	0.4884	13.22	0.7153	0.8132
LeGend (Chapter 4)	**26.1247**	38.27	0.6617	15.87	0.7559	0.8402
LeGend-Diffusion	26.1931	**38.55**	**0.6963**	**16.07**	**0.8490**	**0.8676**

5.3.4 Ablation Study

To rigorously decouple the contributions of the Vital Layer Strategy and the GBCA Module, we conducted a controlled ablation study. We analyze the impact of each component through both quantitative metrics (Table 5.2) and detailed visual inspection (Figure 5.5).

Quantitative Component Analysis

Baseline (Standard DiT Editing). The baseline model fine-tunes all layers of the DiT backbone without any structural constraints. As shown in the first row of Table 5.2, this unconstrained approach leads to a collapse in structural similarity metrics. This indicates that without explicit guidance, the diffusion process tends to over-edit the image. As shown in 5.5, Base DiT introduces unnecessary modifications in the upper chest outside the target pathological region. Consequently, the FID score is high, reflecting a significant divergence from the realistic medical image manifold.

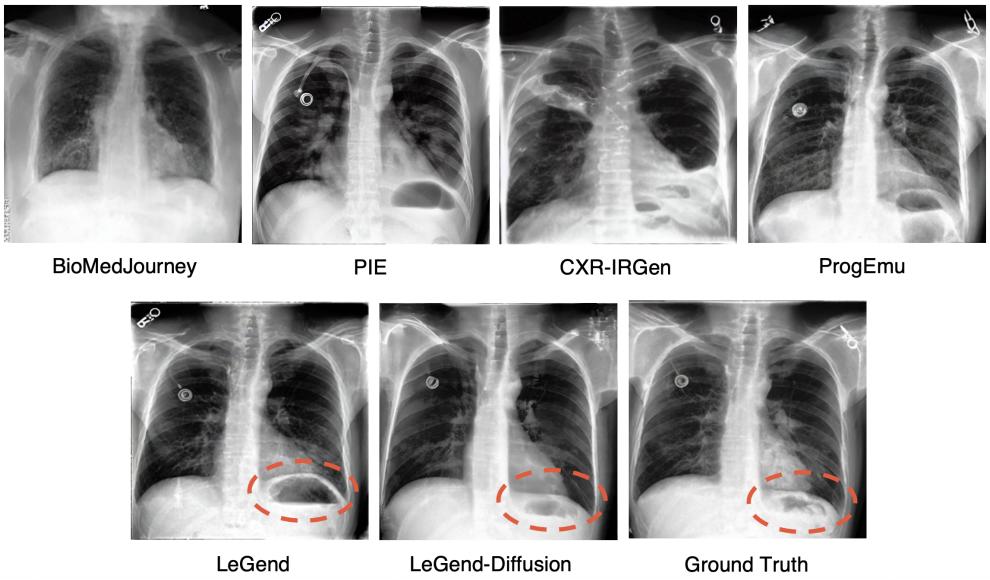


Figure 5.4: Qualitative comparison of longitudinal CXR generation. The top row displays results from baseline methods, which exhibit various degrees of structural distortion or blurring. The bottom row compares our previous AR model (LeGend), our proposed method, and the Ground Truth. Our Diffusion method (LeGend-Diffusion) achieves the best trade-off, preserving the precise anatomical structure of the reference image while faithfully generating the progression lesion.

Effect of Vital Layer Injection (Structure Lock). By identifying and restricting edits in the Vital Layers ($\mathcal{V}$), we observe a recovery in structural fidelity. The MS-SSIM and PSNR both increase. This confirms that transferring features from the source path effectively "anchors" the global geometry. However, relying solely on Vital Layers introduces a trade-off: the strong features from the reference image tend to suppress the generation of new lesions. This results in a moderate improvement in clinical metrics, as the model struggles to synthesize prominent pathologies against the structural constraints.

Effect of GBCA (Pathology Guidance). The integration of Gaussian-Biased Causal Attention (Full Method) resolves the "Structure-vs-Editing" conflict. By explicitly biasing the attention mechanism within the locked Vital Layers, GBCA acts as a spatial override. This configuration achieves the optimal balance: maintaining high structural scores while maximizing clinical utility.

Visual Inspection and Artifact Analysis

To provide a comprehensive assessment, we analyze the model's behavior by cross-referencing the mechanism visualization in Figure 5.3 and the ablation comparison in Figure 5.5. The target progression involves a *"decreased consolidation in the left lower lobe"* alongside an update to the pleural effusion.

Texture Smearing in Base DiT. We first identify the pathology region of interest (ROI) using the "Points on Reference Image" column in Figure 5.3, where the VLM

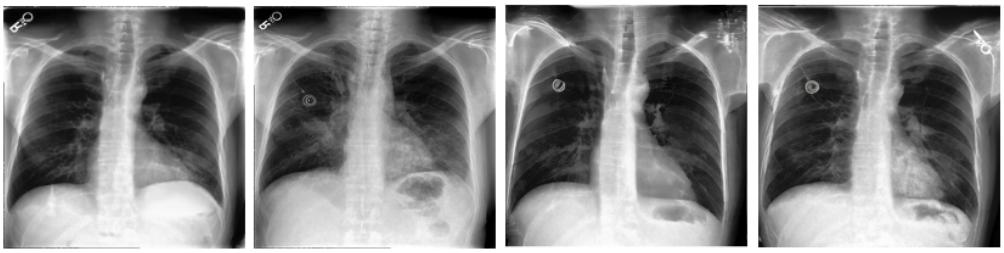


Figure 5.5: Visual ablation of framework components. The text prompt describes progression in the left lung. (a) Base DiT introduces unnecessary changes in non-target upper-chest content while also failing to synthesize the requested pathology faithfully. (b) + Vital Layer improves preservation of the overall image context, but suppresses the requested pathological change. (c) + Vital Layer + GBCA (LeGend-Diffusion) successfully generates the lesion progression while keeping non-target regions largely stable, closely matching (d) the Ground Truth.

Table 5.2: Ablation study on the two core components. **Trend Analysis:** The Baseline suffers from structural collapse. Vital Layer injection recovers structure but limits lesion synthesis due to reference suppression. The addition of GBCA achieves the best trade-off, maximizing both structural fidelity and clinical accuracy.

Method	**Generation Quality**				**Classifier Quality**	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
Base DiT	61.7959	37.55	0.4215	11.34	0.6012	0.6845
+ Vital Layer	45.3210	38.10	0.6963	16.07	0.7640	0.8120
+ Vital + GBCA	**26.1931**	**38.55**	**0.6563**	**16.07**	**0.8490**	**0.8676**

explicitly marks the lower lung field. Comparing this ROI with the "Base DiT" column in Figure 5.5, we observe severe generative artifacts. The texture in the targeted area exhibits an unnatural "smearing" effect. Unlike the "Reference Image" in Figure 5.3, which displays clear, high-frequency granular noise characteristic of authentic X-rays, the Base DiT output appears blurry and "painted-on." This degradation explains the poor FID score reported in Table 5.2, as the generated texture distribution shifts significantly away from the real medical data manifold.

Unnecessary Edits in Non-Target Image Content. Beyond texture, a critical failure mode of the Base DiT is that it alters image regions unrelated to the target pathology. In the example shown in Figure 5.5, visible upper-chest content present in the paired ground-truth follow-up is partially removed and replaced by generic lung-like texture. We do not consider exact preservation of non-anatomical support devices (e.g., monitoring leads) to be a clinical requirement, since such devices may be absent or repositioned across follow-up examinations. Rather, this example illustrates a lack of edit locality: the baseline modifies non-target regions in addition to the pathology-relevant area. This behavior is consistent with the broader identity-loss problem, where an unconstrained diffusion process fails to separate intended pathological changes from irrelevant image

content.

Restoration and Precision via Our Framework. The "+ Vital Layer" column (Figure 5.5) demonstrates that our structural injection strategy successfully restores the support devices and rib alignment, matching the Ground Truth. However, without spatial guidance, the pathological change remains faint due to the reference suppression effect.

Finally, the "LeGend-Diffusion" column (or "+ Vital Layer + GBCA") showcases the efficacy of the full framework: non-target regions remain substantially more stable, and the "smearing" artifact is replaced by a realistic, textured opacity that accurately reflects the disease progression described in the text.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis studied longitudinal chest X-ray (CXR) generation—synthesizing a clinically plausible follow-up radiograph conditioned on a reference study and a textual progression description. A central observation throughout this work is that, while modern transformer-based generators can achieve strong global realism, they may systematically under-attend to clinically meaningful local pathology. This mismatch between where the model allocates attention and where clinical change occurs can lead to poor lesion depiction, peripheral artifacts, and reduced downstream utility.

To address this problem, we contributed a unified, lesion-centric attention prior that is applicable across two major generative paradigms. First, we diagnosed a recurrent peripheral-attention failure mode in transformer-based medical image generation and proposed two lightweight diagnostics—the Corner Bias Index (CBI) and Edge Attention Ratio (EAR)—to quantify corner/edge over-attention and its deviation from lesion-relevant regions. Second, we introduced Gaussian-Biased Causal Attention (GBCA), which injects a lesion-conditioned 2D Gaussian spatial prior into attention logits, explicitly steering attention toward pathology regions derived from progression descriptions while preserving the underlying generation process (e.g., causality in autoregressive decoding).

We validated the effectiveness and generality of the proposed mechanism through two instantiations. (1) In Chapter 4, we implemented an autoregressive follow-up generator (LeGend) and showed that applying GBCA at the semantic-formation stage of the transformer can improve lesion alignment and reduce peripheral artifacts without compromising global anatomical consistency. (2) In Chapter 5, we extended GBCA to diffusion models with transformer backbones by integrating it into a structure-preserving inference pipeline. Specifically, we combined (i) offline structural calibration to identify Vital Layers that are most responsible for anatomical formation and (ii) a dual-path guided generation strategy that preserves patient-specific structure while enabling localized pathology synthesis. Empirically, this design alleviated typical failure modes of transformer-based diffusion in medical imaging, such as identity loss and texture

smearing in the region of interest, and improved both generation fidelity and clinically relevant downstream performance.

Overall, this thesis demonstrates that a lesion-conditioned spatial prior, when injected at the right computational locus, can meaningfully improve the clinical credibility of longitudinal CXR synthesis. Beyond the specific task studied here, the proposed analysis tools (CBI/EAR) and the GBCA mechanism provide a practical template for diagnosing and correcting attention allocation failures in controllable medical image generation.

A key limitation of the present work is that, despite using longitudinal reference / follow-up pairs, the proposed models do not explicitly condition on the elapsed follow-up interval Δ. Consequently, the thesis improves the faithfulness of progression-conditioned follow-up synthesis, but does not yet model time-specific disease trajectories in a way that would support questions such as how a finding may evolve after a prescribed time horizon. Addressing this limitation is an important next step toward greater clinical usability.

6.2 Future Work

While the proposed framework improves lesion-grounded synthesis, several important directions remain open.

(1) Uncertainty-aware and richer spatial priors. GBCA currently uses a compact Gaussian prior parameterized by sparse lesion coordinates. Future work can represent uncertainty in localization (e.g., probabilistic heatmaps), incorporate multiple regions and shapes (mixtures, anisotropic components, learned layouts), and fuse complementary signals such as weak segmentation masks or radiology report grounding to better capture complex or multi-focal disease patterns.

(2) Explicit interval-aware longitudinal forecasting. The current framework uses longitudinal reference/follow-up pairs but does not explicitly encode the elapsed time interval Δ between the two studies. As a result, it cannot directly answer clinically specific questions such as whether a lesion would evolve differently after 6 weeks versus 2 months. A natural next step is therefore to model the conditional distribution

$$x_{\text{fu}} \sim p_\theta(x \mid x_{\text{ref}}, \Delta, y),$$

where Δ is incorporated jointly with the progression cue y. This could be implemented using continuous-time embeddings, discretized interval tokens, or temporal latent-dynamics modules designed for irregularly sampled follow-up studies. Extending the present two-timepoint setting in this way would improve both clinical usability and the realism of patient-specific disease forecasting.

(3) Extension to 3D volumetric modalities and cross-modality longitudinal settings. A key limitation of CXR is projection overlap, whereas CT/MRI provide volumetric evidence of lesion evolution. Applying GBCA-like lesion-centric priors to 3D diffusion/transformer backbones raises new challenges (memory, 3D spatial calibration, slice consistency) but could substantially broaden clinical impact. Similarly, cross-modality longitudinal generation (e.g., CXR-guided CT completion or CT-to-CXR projection-aware synthesis) is an attractive direction.

A key limitation of CXR is projection overlap, whereas CT and MRI provide volumetric evidence of lesion evolution. Extending GBCA-like lesion-centric priors to 3D diffusion/transformer backbones raises new challenges, including memory efficiency, 3D spatial calibration, and inter-slice consistency, but could substantially broaden the clinical applicability of the framework. In addition, cross-modality longitudinal generation (e.g., CXR-guided CT completion or CT-to-CXR projection-aware synthesis) is a promising direction for future research.

(4) Stronger clinical validation and task-aware evaluation. Future work should incorporate radiologist-driven assessments of progression faithfulness and plausibility, and adopt task-aware benchmarks that reflect clinical decision points (e.g., change detection, severity scoring, device preservation). Evaluations can be expanded to include robustness across subgroups, acquisition settings (portable vs. standard), and distribution shifts.

(5) Safety, privacy, and deployment considerations. Since synthetic medical images do not automatically guarantee privacy, future work should include privacy risk assessment under membership inference or data extraction settings, and explore privacy-preserving training or controlled release protocols. For deployment, integrating longitudinal generation into decision-support workflows requires careful human-in-the-loop design, provenance tracking, and calibrated uncertainty reporting to avoid over-reliance on synthetic evidence.

Bibliography

[1] O. Avrahami, O. Patashnik, O. Fried, E. Nemchinov, K. Aberman, D. Lischinski, and D. Cohen-Or. "Stable flow: Vital layers for training-free image editing". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 7877–7888.

[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. "Qwen2. 5-vl technical report". In: *arXiv preprint arXiv:2502.13923* (2025).

[3] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu. "All are worth words: A vit backbone for diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 22669–22679.

[4] C. Bluethgen, P. Chambon, J.-B. Delbrouck, R. Van Der Sluijs, M. Połacin, J. M. Zambrano Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. S. Chaudhari. "A vision–language foundation model for the generation of realistic chest x-ray images". In: *Nature Biomedical Engineering* 9.4 (2025), pp. 494–506.

[5] A. P. Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[6] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vaya. "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical image analysis* 66 (2020), p. 101797.

[7] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. "Extracting training data from diffusion models". In: *32nd USENIX security symposium (USENIX Security 23)*. 2023, pp. 5253–5270.

[8] P. Chambon, C. Bluethgen, J.-B. Delbrouck, R. Van der Sluijs, M. Połacin, J. M. Z. Chaves, T. M. Abraham, S. Purohit, C. P. Langlotz, and A. Chaudhari. "Roentgen: vision-language foundation model for chest x-ray generation". In: *arXiv preprint arXiv:2211.12737* (2022).

[9] P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. Truong, C. T. Chuong, and C. P. Langlotz. "Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats". In: *arXiv preprint arXiv:2405.19538* (2024).

[10] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. "Maskgit: Masked generative image transformer". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11315–11325.

[11] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models". In: *ACM transactions on Graphics (TOG)* 42.4 (2023), pp. 1–10.

[12] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. "Generative pretraining from pixels". In: *International conference on machine learning*. PMLR. 2020, pp. 1691–1703.

[13] H. Chu, X. Qi, H. Wang, and Y. Liang. "Multi-label pathology editing of chest X-rays with a Controlled Diffusion Model". In: *Medical Image Analysis* (2025), p. 103584.

[14] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand. "On the limits of cross-domain generalization in automated X-ray prediction". In: *Medical Imaging with Deep Learning*. PMLR. 2020, pp. 136–155.

[15] Z.-X. Cui, C. Cao, Y. Wang, S. Jia, J. Cheng, X. Liu, H. Zheng, D. Liang, and Y. Zhu. "Spirit-diffusion: Self-consistency driven diffusion model for accelerated mri". In: *IEEE Transactions on Medical Imaging* (2024).

[16] S. U. H. Dar, M. Seyfarth, I. Ayx, T. Papavassiliu, S. O. Schoenberg, R. M. Siepmann, F. C. Laqua, J. Kahmann, N. Frey, B. Baeßler, et al. "Unconditional latent diffusion models memorize patient imaging data". In: *Nature Biomedical Engineering* (2025), pp. 1–15.

[17] S. Dayarathna, K. T. Islam, S. Uribe, G. Yang, M. Hayat, and Z. Chen. "Deep learning based synthesis of MRI, CT and PET: Review and analysis". In: *Medical image analysis* 92 (2024), p. 103046.

[18] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. "Scaling rectified flow transformers for high-resolution image synthesis". In: *Forty-first international conference on machine learning*. 2024.

[19] Y. Fan, J. Xie, Y. Luo, Y. Meng, S. Madhusudhan, G. Y. Lip, L. Cheng, Y. Zheng, and H. Zhao. "t HPM-LDM: Integrating Individual Historical Record with Population Memory in Latent Diffusion-Based Glaucoma Forecasting". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2025, pp. 619–629.

[20] A. Fontanella, G. Mair, J. Wardlaw, E. Trucco, and A. Storkey. "Diffusion models for counterfactual generation and anomaly detection in brain images". In: *IEEE Transactions on Medical Imaging* (2024).

[21] R. Girshick. "Fast r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[22] 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.

[23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. "The llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (2024).

[24] Y. Gu, J. Yang, N. Usuyama, C. Li, S. Zhang, M. P. Lungren, J. Gao, and H. Poon. "Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys". In: *arXiv preprint arXiv:2310.10765* (2023).

[25] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. "Prompt-to-prompt image editing with cross attention control". In: *arXiv preprint arXiv:2208.01626* (2022).

[26] 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).

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

[28] J. Ho and T. Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint arXiv:2207.12598* (2022).

[29] S. Hong, G. Lee, W. Jang, and S. Kim. "Improving sample quality of diffusion models using self-attention guidance". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 7462–7471.

[30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. "Lora: Low-rank adaptation of large language models." In: *ICLR* 1.2 (2022), p. 3.

[31] P. Huang, X. Gao, L. Huang, J. Jiao, X. Li, Y. Wang, and Y. Guo. "Chest-diffusion: a light-weight text-to-image model for report-to-cxr generation". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2024, pp. 1–5.

[32] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1 (2019), p. 317.

[33] C. Jokerst, J. H. Chung, J. B. Ackman, B. Carter, P. M. Colletti, T. D. Crabtree, P. M. de Groot, M. D. Iannettoni, F. Maldonado, B. L. McComb, et al. "ACR Appropriateness Criteria® acute respiratory illness in immunocompetent patients". In: *Journal of the American College of Radiology* 15.11 (2018), S240–S251.

[34] T. Karras, M. Aittala, T. Aila, and S. Laine. "Elucidating the design space of diffusion-based generative models". In: *Advances in neural information processing systems* 35 (2022), pp. 26565–26577.

[35] T. Karras, S. Laine, and T. Aila. "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[36] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. "Diffusion models for medical image analysis: A comprehensive survey". In: *arXiv preprint arXiv:2211.07804* (2022).

[37] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. "Diffusion models in medical imaging: A comprehensive survey". In: *Medical image analysis* 88 (2023), p. 102846.

[38] 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: *Advances in Neural Information Processing Systems* 36 (2023), pp. 28541–28564.

[39] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. "Gligen: Open-set grounded text-to-image generation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 22511–22521.

[40] K. Liang, X. Cao, K.-D. Liao, T. Gao, W. Ye, Z. Chen, J. Cao, T. Nama, and J. Sun. *PIE: Simulating Disease Progression via Progressive Image Editing*. 2023. arXiv: 2309.11745 [eess.IV].

[41] W. S. Lim, S. Baudouin, R. George, A. Hill, C. Jamieson, I. Le Jeune, J. Macfarlane, R. Read, H. Roberts, M. Levy, et al. "BTS guidelines for the management of community acquired pneumonia in adults: update 2009". In: *Thorax* 64.Suppl 3 (2009), pp. iii1–iii55.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[43] C. Ma, Y. Ji, J. Ye, L. Zhang, Y. Chen, T. Li, M. Li, J. He, and H. Shan. "Towards interpretable counterfactual generation via multimodal autoregression". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2025, pp. 611–620.

[44] N. Mouadden, O. Laousy, R. Marini, V. Ong, M.-P. Revel, G. Chassagnon, S. Christodoulidis, and M. Vakalopoulou. "Conditional Latent Diffusion Models for Irregularly Spaced Longitudinal Radiological Data". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2025, pp. 106–115.

[45] J. Mu, N. Vasconcelos, and X. Wang. "Editar: Unified conditional generation with autoregressive models". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 7899–7909.

[46] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. "Medical image synthesis with deep convolutional adversarial networks". In: *IEEE Transactions on Biomedical Engineering* 65.12 (2018), pp. 2720–2730.

[47] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (2023).

[48] W. H. Organization et al. *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. Tech. rep. World Health Organization, 2016.

[49] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Cukur. "Unsupervised medical image translation with adversarial diffusion models". In: *IEEE Transactions on Medical Imaging* 42.12 (2023), pp. 3524–3539.

[50] W. Peebles and S. Xie. "Scalable diffusion models with transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4195–4205.

[51] F. Pérez-García, S. Bond-Taylor, P. P. Sanchez, B. van Breugel, D. C. Castro, H. Sharma, V. Salvatelli, M. T. Wetscherek, H. Richardson, M. P. Lungren, et al.

"Radedit: stress-testing biomedical vision models via diffusion image editing". In: *European Conference on Computer Vision*. Springer. 2024, pp. 358–376.

[52] O. Press, N. A. Smith, and M. Lewis. "Train short, test long: Attention with linear biases enables input length extrapolation". In: *arXiv preprint arXiv:2108.12409* (2021).

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[55] 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.

[56] P. Safety. "Radiation Dose in X-ray and CT Exams". In: *American College of Radiology and Radiological Society of North America (April 2012)* (2012).

[57] J. Shentu and N. Al Moubayed. "Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2024, pp. 5212–5221.

[58] Y. Skandarani, P.-M. Jodoin, and A. Lalande. "Gans for medical image synthesis: An empirical study". In: *Journal of Imaging* 9.3 (2023), p. 69.

[59] S. Skinner. "Guide to thoracic imaging". In: *Australian Family Physician* 44.8 (2015), pp. 558–563.

[60] J. Song, C. Meng, and S. Ermon. "Denoising diffusion implicit models". In: *arXiv preprint arXiv:2010.02502* (2020).

[61] B. Theodorou, A. Dadu, M. Nalls, F. Faghri, and J. Sun. "SECONDGRAM: Self-conditioned diffusion with gradient manipulation for longitudinal MRI imputation". In: *Patterns* 6.5 (2025).

[62] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. "Conditional image generation with pixelcnn decoders". In: *Advances in neural information processing systems* 29 (2016).

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

[64] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. "Pixel recurrent neural networks". In: *International conference on machine learning*. PMLR. 2016, pp. 1747–1756.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[66] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, et al. "Emu3: Next-token prediction is all you need". In: *arXiv preprint arXiv:2409.18869* (2024).

[67] Z. Wang, E. P. Simoncelli, and A. C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The thrity-seventh asilomar conference on signals, systems & computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[68] G. Webber and A. J. Reader. "Diffusion models for medical image reconstruction". In: *BJR| Artificial Intelligence* 1.1 (2024), ubae013.

[69] N. Weng, P. Pegios, E. Petersen, A. Feragen, and S. Bigdeli. "Fast diffusion-based counterfactuals for shortcut removal and generation". In: *European Conference on Computer Vision*. Springer. 2024, pp. 338–357.

[70] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 21372–21383.

[71] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, and M. Z. Shou. "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 7452–7461.

[72] Y. Yeganeh, A. Farshad, I. Charisiadis, M. Hasny, M. Hartenberger, B. Ommer, N. Navab, and E. Adeli. "Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 7685–7695.

[73] X. Yi, E. Walia, and P. Babyn. "Generative adversarial network in medical imaging: A review". In: *Medical image analysis* 58 (2019), p. 101552.

[74] 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.

[75] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs". In: *arXiv preprint arXiv:2303.00915* (2023).

[76] 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.

[77] Z. Zhu, T. Tao, Y. Tao, H. Deng, X. Cai, G. Wu, K. Wang, H. Tang, L. Zhu, Z. Gu, et al. "Loci-diffcom: Longitudinal consistency-informed diffusion model for 3d infant brain image completion". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 249–258.