Towards Controllable and Interpretable Latent Modeling for Vision and Beyond

JIYANG ZHENG

Doctor of Philosophy

Supervisor: Associate Professor Tongliang Liu
Associate Supervisor: Dr Dadong Wang

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

27 February 2026

Abstract

Deep learning models for vision and multimodal data rely on high-dimensional latent representations to achieve strong empirical performance. However, such representations are often difficult to interpret and control, limiting their reliability and adaptability in tasks requiring structured reasoning. This thesis investigates how latent representations and latent structures can be organised to support both interpretability and controllability across visual and multimodal learning settings. This work propose that effective latent modeling requires explicitly distinguishing task-relevant semantic factors from task-irrelevant variability, and enforcing selective invariance and alignment during learning. Rather than relying on fully entangled latent spaces or opaque internal processes, the thesis adopts the perspective that latent structures should expose meaningful organisation that can be selectively preserved, modified, or aligned according to task requirements. The first part of the thesis examines this principle in ordinal visual learning, where discriminative information is subtle and order-dependent. It shows that commonly used representation learning strategies induce excessive invariance, obscuring ordinal semantics in the latent space. By encouraging minimal and targeted latent variation, the proposed approach retains ordinal meaning while remaining robust to irrelevant changes. The second part studies controllability in visual in-context learning, demonstrating that latent representations in large autoregressive vision models can be made more interpretable through structured intermediate representations that reflect progressive visual reasoning. The third part extends these principles to multimodal generation by learning selectively aligned latent spaces that capture shared semantic factors while excluding modality-specific variability. Finally, the thesis generalises latent modeling beyond internal representations to interpretable procedural structures, showing that agentic workflows can be treated as latent constructs whose organisation is learned under explicit constraints.

Statement of Originality

I certify that, to the best of my knowledge, this thesis contains no material previously published or written by another person, except where due acknowledgement is made. This thesis has not been submitted, either in whole or in part, for the award of any degree or diploma at this or any other institution.

I further certify that the intellectual content of this thesis is the result of my own work. All sources of information and assistance received in the preparation of this thesis have been appropriately acknowledged.

Jiyang Zheng
School of Computer Science
Faculty of Engineering
The University of Sydney

27th February 2026

Statement of Generative AI

During the preparation of this thesis, I used ChatGPT (OpenAI) and Gemini (Google) for the purpose of language enhancement. The use of this generative artificial intelligence tool was limited to minor paraphrasing, sentence restructuring, and correction of spelling and grammatical errors in selected draft chapters. All AI-assisted modifications were carefully reviewed by the author to ensure accuracy, coherence, and the absence of errors or unintended bias. The author takes full responsibility for the content of the submitted thesis and confirms that the work is original and that generative AI was used strictly within permitted and ethical academic guidelines.

Jiyang Zheng
School of Computer Science
Faculty of Engineering
The University of Sydney

27th February 2026

Authorship Attribution Statement

This thesis was undertaken at the University of Sydney between 2023 and 2026, under the supervision of Associate Professor Tongliang Liu. The principal findings presented in this dissertation were originally reported in the following publications:

(1) **Jiyang Zheng**, Yu Yao, Bo Han, Dadong Wang, Tongliang Liu. "Enhancing Contrastive Learning for Ordinal Regression via Ordinal Content Preserved Data Augmentation". In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. Presented in Chapter 2. I conceived the research, carried out the experiments, and prepared the manuscript. Revisions to the paper were undertaken collaboratively with the other co-authors.

(2) **Jiyang Zheng**, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, Tongliang Liu. "Chain-of-Focus Prompting: Leveraging Sequential Visual Cues to Prompt Large Autoregressive Vision Models". *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. Presented in Chapter 3. I conceived the research, carried out the experiments, and prepared the manuscript. Revisions to the paper were undertaken collaboratively with the other co-authors.

(3) **Jiyang Zheng**, Siqi Pan, Yu Yao, Zhaoqing Wang, Dadong Wang, Tongliang Liu. "Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation". In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. Presented in Chapter 4. I conceived the research, carried out the experiments, and prepared the manuscript. Revisions to the paper were undertaken collaboratively with the other co-authors.

(4) **Jiyang Zheng**, Islam Nassar, Thanh Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong, Yuan-Fang Li. "MedDCR: Learning to Design Agentic Workflows for Medical Coding". In *arXiv preprint arXiv:2511.13361*, 2025. Presented in Chapter 5.

I conceived the research, carried out the experiments, and prepared the manuscript.

Revisions to the paper were undertaken collaboratively with the other co-authors.

In addition to the statements above, for publications in which I am not the corresponding author, permission to include the published material has been obtained from the corresponding author.

Jiyang Zheng
School of Computer Science
Faculty of Engineering
The University of Sydney 27th February 2026

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

Tongliang Liu
School of Computer Science
Faculty of Engineering
The University of Sydney 27th February 2026

Sidere mens eadem mutato.

Acknowledgements

Pursuing a PhD is a demanding and challenging journey. I consider myself very fortunate to have received invaluable support from my supervisors, family members, colleagues, and friends. This thesis would not have been possible without the guidance, encouragement, and assistance of each of them.

First and foremost, I am deeply grateful to my principal supervisor, Associate Professor Tongliang Liu, for his invaluable guidance, encouragement, and intellectual support throughout my doctoral studies. Beyond his role as an academic supervisor, he has been a mentor who has profoundly influenced my personal and professional development. His insight, patience, and rigorous standards have been instrumental in shaping my growth as a researcher.

I am also sincerely thankful to my other mentors: Dr. Dadong Wang, Dr. Yu Yao, and Dr. Yuan-Fang Li, as well as my collaborators, including Dr. Bo Han, Jialiang Shen, Zhaoqing Wang, Dr. Islam Nassar, Dr. Thanh Vu, Dr. Xu Zhong, Dr. Yang Lin, Dr. Long Duong, Weijie Tu, Dr. Weijian Deng, Dr. Dylan Campbell, Dr. Tom Gedeon, Dr. Siqi Pan, and Dr. Min Wang, Dr. Yang Yang, for their constructive feedback, insightful discussions, and continued support at various stages of this research.

My appreciation extends to my colleagues in the Trustworthy Machine Learning Lab and the Sydney AI Centre. Being part of this community has been a truly enjoyable and rewarding experience, from which I have gained friendship, support, and life inspiration.

I am especially grateful to the University of Sydney and CSIRO for supporting my PhD research through the Faculty of Engineering Research Scholarship and Next Generation Graduates Scholarship Program.

In closing, I would like to thank my father, my mother, and my wife for their unwavering support throughout my PhD studies. To them, I owe my deepest gratitude.

List of Publications

The majority of the works presented in this thesis have been previously published, including:

(1) **Jiyang Zheng**, Yu Yao, Bo Han, Dadong Wang, Tongliang Liu. "Enhancing Contrastive Learning for Ordinal Regression via Ordinal Content Preserved Data Augmentation". In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. [**Chapter 2**]

(2) **Jiyang Zheng**, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, Tongliang Liu. "Chain-of-Focus Prompting: Leveraging Sequential Visual Cues to Prompt Large Autoregressive Vision Models". *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. [**Chapter 3**]

(3) **Jiyang Zheng**, Siqi Pan, Yu Yao, Zhaoqing Wang, Dadong Wang, Tongliang Liu. "Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation". In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. [**Chapter 4**]

(4) **Jiyang Zheng**, Islam Nassar, Thanh Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong, Yuan-Fang Li. "MedDCR: Learning to Design Agentic Workflows for Medical Coding". In *arXiv preprint arXiv:2511.13361*, 2025. [**Chapter 5**]

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Recent years have witnessed rapid progress in vision and multimodal learning, driven by advances in representation learning (Chen et al., 2020b; Caron et al., 2021; Radford et al., 2021), large-scale datasets (Deng et al., 2009; Lin et al., 2014; Schuhmann et al., 2022), and increasingly expressive neural architectures (He et al., 2016; Vaswani et al., 2017). Across supervised, self-supervised, and generative paradigms, modern models rely heavily on latent representations as the primary medium through which information is encoded, transformed, and synthesized. These latent spaces underpin a wide range of capabilities, including visual recognition (Shafiq and Gu, 2022), reasoning (Zakari et al., 2022; Sun et al., 2025), and cross-modal generation (Żelaszczyk and Mańdziuk, 2024; He et al., 2024), and have become a central abstraction for understanding and designing learning systems.

Despite their empirical success, latent representations learned by contemporary models are often difficult to interpret and manipulate in a principled manner. In practice, latent variables tend to capture complex mixtures of semantic factors and incidental variations, with little explicit alignment to human-interpretable concepts (Gandelsman et al., 2024; Xie et al., 2025). This lack of structure poses a fundamental challenge for tasks that require model transparency, controlled intervention, or reliable generalisation beyond the training distribution. As models grow in scale and are deployed in increasingly complex settings, the limitations of opaque and entangled latent spaces become more pronounced.

A common characteristic of many existing approaches is that latent representations are shaped implicitly by optimisation objectives that prioritise predictive or generative performance (Gui et al., 2024). While this implicit learning often yields strong task performance, it provides limited guarantees about the semantic organisation of the resulting latent space. Task-relevant

factors are frequently entangled with nuisance variations, and individual latent dimensions lack clear functional roles (Higgins et al., 2017b; Wang et al., 2024c). As a consequence, models may rely on spurious correlations, encode brittle decision rules, or exhibit failure modes that are difficult to diagnose or correct.

These issues are particularly acute in settings where the underlying semantic signals are subtle, structured, or only partially observable. For example, in ordinal prediction tasks, the discriminative information may lie in fine-grained and ordered variations rather than categorical differences (Wang et al., 2025a). In visual reasoning, successful inference often depends on selectively attending to relational or causal structure rather than global appearance statistics (Ke et al., 2025). Similarly, in multimodal generation, different modalities frequently exhibit incomplete or asymmetric correspondence, such that forcing full alignment across modalities can obscure modality-specific semantics or introduce semantic noise (Xie et al., 2025). In these scenarios, indiscriminate enforcement of invariance or alignment can degrade both interpretability and performance.

Existing research has explored disentanglement, invariance, and interpretability from a variety of perspectives, including variational modeling, contrastive learning, and self-supervised objectives (Higgins et al., 2017b; Khemakhem et al., 2020; Von Kügelgen et al., 2021). While these efforts have yielded important theoretical and empirical insights, many approaches still treat latent spaces as homogeneous embeddings. As a result, they offer limited mechanisms for selectively preserving, suppressing, or aligning specific semantic factors based on task requirements. This gap motivates a more structured view of latent representation learning.

This thesis is motivated by the observation that controllability and interpretability are closely related properties of latent representations, and that both can be improved by explicitly distinguishing task-relevant semantic factors from task-irrelevant variation during learning (Von Kügelgen et al., 2021). Rather than treating latent spaces as monolithic vectors, the thesis adopts the perspective that latent representations should expose internal structure, with different components serving distinct semantic purposes. From this viewpoint, effective representation learning is not solely about maximizing information content, but about organising information in a manner that supports selective use, intervention, and analysis.

Central to this perspective is the notion of selective modeling (Xiao et al., 2021; Von Kügelgen et al., 2021; Xie et al., 2025). Instead of enforcing uniform invariance or alignment across all latent dimensions, selective modeling aims to preserve information that is causally or semantically relevant to a given task, while suppressing variation that is irrelevant. This principle manifests in several forms throughout the thesis, including selective invariance for ordinal learning, structured latent reasoning for vision models, and selective cross-modal alignment for multimodal generation. In each case, the goal is to guide the learning process toward latents that are both more interpretable and amenable to controlled manipulation.

Beyond internal feature representations, this thesis further argues that latent structure can arise at the level of reasoning processes and computational workflows. In complex decision-making settings, model behaviour is often governed not only by latent embeddings, but by latent procedural structures that determine how intermediate representations are generated, combined, and validated (Luo et al., 2025). These structures, such as reasoning sequences or agentic workflows are typically implicit, manually designed, or fixed a priori, limiting both transparency and adaptability (Wu et al., 2024a). By treating such workflows as latent objects that can be learned, constrained, and analysed (Hu et al., 2025; Zhang et al., 2025b), the same principles of selective modeling can be extended from representation spaces to procedural reasoning structures, enabling interpretable and controllable behaviour at the system level.

The scope of this thesis primarily focuses on vision and closely related multimodal learning problems, with an emphasis on representation learning and generative modeling. The final part of the thesis extends these ideas to agentic reasoning pipelines, illustrating how latent modeling principles generalise beyond perception to structured decision-making tasks. Throughout, the emphasis remains on settings where explicit supervision is limited and architectural modifications are kept minimal. This reflects a practical concern: improving interpretability and controllability should not come at the cost of excessive supervision or fundamentally altering well-established model architectures. Instead, the thesis explores how latent structure can be induced through principled objectives, constraints, and training strategies that integrate naturally with existing frameworks.

Collectively, this thesis aims to provide a coherent framework for understanding and designing latent representations and latent structures that support transparent, interpretable, and controllable model behaviour across diverse learning paradigms. By grounding the analysis in concrete vision, multimodal, and agentic reasoning tasks, the thesis seeks to bridge theoretical insights on latent structure with practical considerations in modern deep learning systems.

1.1 Thesis Outline

The remainder of this thesis is organised as follows.

Chapter 2, Controllable Latent Representations for Ordinal Visual Learning, investigates how latent representations can be structured to preserve task-relevant semantic order while suppressing nuisance variation. Focusing on ordinal prediction tasks, this chapter shows that commonly used representation learning objectives induce excessive invariance, which obscures fine-grained ordinal semantics in the latent space. It then develops principled strategies for encouraging minimal and targeted latent variation, yielding representations that are both more interpretable and more amenable to controlled manipulation. This chapter establishes the thesis's core argument that interpretability and controllability depend on explicitly structuring latent representations according to task semantics.

Chapter 3, Structured Latent Reasoning in Large Vision Models, extends the notion of structured latent modeling from static feature representations to dynamic reasoning processes. It examines how latent representations in large autoregressive vision models can be organised to expose intermediate reasoning structure, enabling interpretable and controllable visual in-context learning. By introducing structured latent transitions that reflect progressive visual reasoning, this chapter demonstrates how control can be exerted over model behaviour through interpretable internal representations rather than opaque end-to-end inference.

Chapter 4, Interpretable Cross-Modal Latent Modeling for Multimodal Generation, generalises the proposed latent modeling principles to multimodal generative settings. Focusing on text-to-audio-video generation, this chapter addresses the challenge of partial and

asymmetric correspondence between modalities. It proposes selectively aligned latent spaces that capture shared semantic factors while excluding modality-specific noise, resulting in improved interpretability, controllability, and cross-modal coherence. This chapter highlights how selective latent alignment supports robust multimodal generation without enforcing unnecessary or harmful invariance.

Chapter 5, Interpretable Latent Workflow Modeling for Agentic Reasoning, further extends the thesis beyond internal representations to latent procedural structures that govern complex reasoning systems. It treats agentic workflows as latent computational graphs whose components and execution order are learned rather than manually designed. By explicitly modelling and optimising these workflows under task and constraint feedback, the chapter demonstrates how the principles of controllable and interpretable latent modeling generalise to system-level reasoning processes, enabling transparent and controllable decision-making in real-world tasks.

Chapter 6, Conclusions, concludes the thesis by summarising the main contributions, discussing limitations of the proposed approaches, and outlining directions for future research on structured, controllable, and interpretable latent modeling across vision, multimodal, and agentic learning systems.

CHAPTER 2

Controllable Latent Representations for Ordinal Visual Learning

---

This chapter examines ordinal visual learning as an initial case study for controllable and interpretable latent modeling, focusing on prediction tasks where ordered labels depend on subtle and localized semantic cues. It highlights a fundamental mismatch between standard contrastive learning, particularly its reliance on strong, predefined augmentations, and the semantic requirements of ordinal regression, where task-relevant information can be easily disrupted. To address this limitation, the chapter introduces a generative, content-preserving augmentation framework based on disentangled latent representations and the principle of minimal change, enabling selective invariance aligned with ordinal semantics. By doing so, the chapter establishes the central thesis argument that effective representation learning in structured prediction tasks requires explicit control over latent factors.

2.1 Introduction

Ordinal visual learning addresses a class of prediction problems in which target labels are discrete but inherently ordered, such as age estimation, disease severity grading, and quality assessment (Niu et al., 2016; Liu et al., 2017; Beckham and Pal, 2017). Unlike standard categorical classification, ordinal regression requires models to capture subtle, monotonic semantic variations that reflect relative ordering rather than absolute class identity. As a result, performance in ordinal tasks depends critically on the representation's ability to preserve fine-grained semantic structure while remaining invariant to task-irrelevant variations.

Recent advances in representation learning have demonstrated that contrastive learning provides a powerful mechanism for learning invariant and transferable visual features by

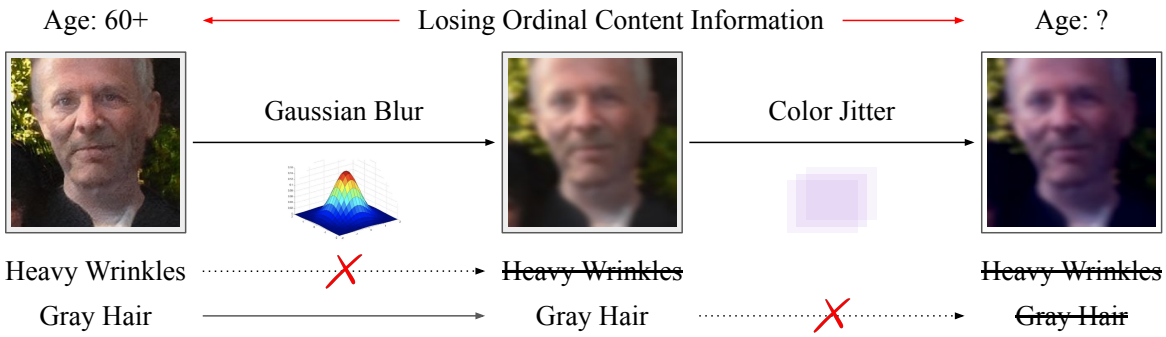

FIGURE 2.1. Ordinal content information in data can be easily distorted by standard augmentations in contrastive learning. As illustrated in the example, the age-related features: wrinkles and hair color are eliminated after Gaussian blurring and color jitter, making the age become unidentifiable.

enforcing consistency between multiple views of the same instance (Oord et al., 2018; He et al., 2019; Khosla et al., 2020). By contrasting strongly and weakly augmented samples, contrastive objectives encourage the extraction of shared semantic information while suppressing nuisance variability (Chen et al., 2020b). This paradigm has been shown to be highly effective across supervised and self-supervised learning settings, contributing substantially to progress in visual representation learning (Jaiswal et al., 2020).

However, when directly applied to ordinal regression, contrastive learning exhibits noticeably diminished effectiveness. This limitation arises not from the contrastive objective itself, but from a mismatch between the assumptions underlying standard contrastive augmentation strategies and the intrinsic structure of ordinal data. In conventional contrastive learning pipelines, strong predefined augmentations, such as aggressive color jittering, color dropping, or blurring are introduced to eliminate superficial correlations and enforce semantic invariance (Von Kügelgen et al., 2021; Xiao et al., 2021). While suitable for category-level discrimination, these transformations can be detrimental in ordinal settings, where the discriminative information is often subtle, localized, and semantically fragile (Wang et al., 2025a).

In ordinal visual tasks, the information that determines relative ordering frequently manifests in small-scale or low-contrast visual cues. For example, in age estimation, ordinal distinctions may depend on localized facial features such as wrinkles, skin texture, or hair coloration,

which occupy only a small fraction of the image (See Figure 2.1). Similarly, in diabetic retinopathy grading (Liu et al., 2022), disease severity is indicated by fine-grained retinal lesions, such as microaneurysms or hemorrhages, that are both spatially localized and visually subtle. Strong augmentations commonly used in contrastive learning can distort or entirely remove these critical cues, thereby suppressing the very information required for accurate ordinal prediction. Consequently, enforcing invariance across such augmented views risks collapsing ordinal content into nuisance variation, undermining both representation quality and downstream performance.

This observation highlights a broader issue that recurs throughout this thesis: effective controllable latent modeling requires explicitly distinguishing task-relevant semantic factors from task-irrelevant variability (Xiao et al., 2021). In ordinal learning, this distinction is particularly acute, as ordinal content information must be preserved rather than abstracted away. From this perspective, the failure of standard contrastive learning in ordinal regression reflects a lack of control over which latent factors are encouraged to remain invariant and which are permitted to vary.

Motivated by this insight, this chapter proposes an alternative approach to contrastive augmentation for ordinal visual learning, one that explicitly preserves ordinal content while enabling controlled variation in non-ordinal factors. Rather than relying on handcrafted strong augmentations (Chen et al., 2020b), we introduce a generative, content-preserving augmentation mechanism grounded in latent factor disentanglement (Kong et al., 2022). The core idea is to construct augmentations that differ in visual style while remaining semantically consistent with respect to ordinal labels.

To achieve this, we employ a generative model trained under the principle of minimal change (Xie et al., 2022), which constrains variations in latent space to have limited impact on ordinal content. Generative models, such as Generative Adversarial Networks (Goodfellow et al., 2014), provide a natural framework for synthesizing diverse visual instances. However, without additional constraints, their latent spaces typically entangle ordinal and non-ordinal factors, resulting in limited controllability over semantic attributes. Such entanglement hinders

the generation of instances with fixed ordinal content and undermines their use as reliable contrastive views.

The minimal change principle addresses this challenge by explicitly regulating the influence of latent factors on the generated output. In this framework, the latent representation is partitioned into two complementary components: one dedicated to ordinal content and another capturing non-ordinal variations such as style, illumination, or background (Von Kügelgen et al., 2021). The ordinal latent component is constrained to undergo minimal change during generation, ensuring that it encodes only the essential information required to determine ordinal labels. Once reconstruction fidelity is achieved under this constraint, ordinal and non-ordinal information become effectively disentangled, yielding a controllable latent structure aligned with the requirements of ordinal learning.

Building on this disentangled representation (Wang et al., 2022a), the proposed method generates ordinal-preserving augmented instances by fixing the ordinal latent factors and resampling the non-ordinal ones. The resulting synthetic images exhibit substantial stylistic diversity while maintaining consistent ordinal semantics. These instances can be used as strong contrastive views, enabling contrastive learning objectives to operate without corrupting ordinal content. Importantly, this strategy allows contrastive learning to be integrated seamlessly into existing ordinal regression frameworks (Li et al., 2019; Li et al., 2021; Shin et al., 2022), while restoring its effectiveness through principled control over latent variation.

In the context of the broader thesis, this chapter establishes a foundational case study demonstrating how controllable and interpretable latent representations can address task-specific failures of generic representation learning techniques (Khosla et al., 2020; Zha et al., 2022). By explicitly aligning latent structure with ordinal semantics, the proposed approach illustrates how selective invariance and controlled generation can improve learning outcomes in settings where semantic signals are weak, localized, or easily disrupted. The ideas developed here foreshadow subsequent chapters, which extend the same principles to structured reasoning in large vision models and interpretable cross-modal alignment in multimodal generation.

The remainder of this chapter is organized as follows. Section 2 reviews related work in ordinal regression and contrastive learning, with particular emphasis on their limitations in handling subtle semantic structure. Section 3 presents the proposed generative augmentation framework and details the application of the minimal change principle. Section 4 evaluates the approach across multiple ordinal visual learning benchmarks. Section 5 concludes the chapter and discusses implications for controllable latent modeling more broadly.

2.2 Related Works

Method for Ordinal Regression. Recent advancements often frame ordinal regression as a classification task (Niu et al., 2016; Liu et al., 2017; Beckham and Pal, 2017). Liu et al., 2018b introduced a constrained optimization formulation for ordinal regression. This approach minimizes the negative log-likelihood across multiple classes while simultaneously preserving the inherent order relationship between instances. Diaz and Marathe, 2019 leveraged the natural ordinal relationships between targets, imparting them as prior knowledge to the model through soft labels. Liu et al., 2019 addressed the task from a probabilistic modeling perspective, where a Gaussian Process model is attached to the output layer of the deep neural network to model uncertainty. Li et al., 2021 proposed a framework that employs probabilistic embeddings to model data uncertainty. Their method enforces a constraint between the learned embedding distributions and pre-defined ordinal distributions, ensuring that the learned embedding space remains ordered. Shin et al., 2022 proposed a regression-based rank estimation algorithm that learns to model the order relationship between instances. Cheng et al., 2023b propose a data fusion approach to address the class-imbalance issue in ordinal regression datasets. While most of the previous studies primarily focused on aligning the model's final predictions with the target, our approach emphasizes the importance of preserving ordinal content information when augmenting ordinal regression data. Our method is orthogonal to end-to-end trainable ordinal regression model, which can serve as a plug-and-play solution to improve the performance of existing state-of-the-art ordinal regression frameworks.

Contrastive Learning and Data Augmentations. Contrastive learning (CL) extracts discriminative information from data by organizing samples into similar and dissimilar pairs. It amplifies the similarity of similar pairs and increases the difference between dissimilar pairs in the feature space. In a self-supervised setting (He et al., 2019; Chen et al., 2020b), similar pairs are generated through a data augmentation module. Given a reference image, this module introduces modifications such as random scaling, cropping, color jittering, blurring, and flipping to generate new perspectives of the image. The original image and its augmented versions constitute a similar pair. In contrast, other images in the batch are treated as dissimilar samples, forming dissimilar pairs. Khosla et al., 2020 expanded CL to a supervised setting. In addition to the augmented versions of the image, samples from the same class also become part of a similar pair. Zha et al., 2022 introduced a supervised CL framework for regression tasks, ensuring that the order of representations in the feature space corresponds to their target values. The results demonstrate that existing regression methods consistently benefit from a CL module for extracting discriminative features from data.

Furthermore, intensive data augmentations have been found crucial for the success of the contrastive learning framework across all settings (Chen et al., 2020b; Chen et al., 2020c; Khosla et al., 2020; Li et al., 2023c; Li et al., 2022a; Huang et al., 2021; Zheng et al., 2022). However, such aggressive augmentations can compromise an image's content. Xiao et al., 2021 addressed this breach of the invariance assumption (i.e., data augmentations altering the data's semantic information) by decomposing a compound series of augmentations into individual operations and creating distinct heads for each single augmentation. This method is effective when data is sensitive to a few specific augmentations within the full augmentation sequence but maintains its invariance assumption with others. Given that content information in ordinal data is particularly susceptible, many augmentation techniques can potentially distort an image's semantics, thereby limiting the method's efficacy. Compared to their approach, our method does not depend on any predetermined augmentation method. Instead, we guide the model to discern which aspects of the augmented data should be preserved as content variables and which can be modified to introduce diversity as style variables.

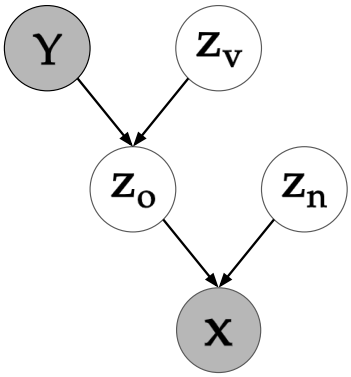

FIGURE 2.2. The data generative process employed by our method. The shaded variables are observable and the unshaded variables are latent.

2.3 Ordinal Content Preserving Contrastive Learning (OCP-CL)

In this section, we introduce our Ordinal Content Preserving Contrastive Learning (OCP-CL) framework. Specifically, to improve the utility of contrastive learning for ordinal regression, we propose a novel ordinal content-preserving augmentation method that replaces the predefined strong augmentations in a contrastive learning framework. First, we explain the generative process of ordinal regression data, which serves as the foundational understanding of our proposed generative model. Then, we introduce our approach for disentangling ordinal content and non-ordinal factors via minimal change, and detail the implementation of the generative model. Next, we describe the process of generating content-preserving data augmentations through interventions on non-ordinal latent factors. Finally, we present the contrastive learning formulation with our generated augmentations from the original instances. The contrastive learning objective can be integrated into any existing end-to-end trainable deep ordinal regression methods to form a joint objective.

Data Generative Process. We first explain the causal data generative process (Glymour and Zhang, 2019; Yao et al., 2023) as illustrated in Figure 2.2. The graph outlines the generative process for ordinal regression data, segregating latent factors into different functional groups based on their relationships to the observed variables.

Specifically, $\boldsymbol{z}_{\mathrm{v}}$ denotes a set of invariant ordinal factors that capture all ordinal content relevant features across different ordinal categories. For example, in age estimation, $\boldsymbol{z}_{\mathrm{v}}$ encompasses a comprehensive collection of age-specific attributes such as the severity of wrinkles or variations in skin texture across age groups. The ordinal label y serves as a constraining variable, ensuring that $\boldsymbol{z}_{\mathrm{o}}$ selectively retains only those features from $\boldsymbol{z}_{\mathrm{v}}$ that pertain to its corresponding ordinal category. Similarly, $\boldsymbol{z}_{\mathrm{n}}$ denotes the set of non-ordinal, style-related factors. Together, $\boldsymbol{z}_{\mathrm{o}}$ and $\boldsymbol{z}_{\mathrm{n}}$ collaboratively generate x, the observed data instance. Our primary objective for the generative model is to disentangle the learned latent factors $\hat{\boldsymbol{z}}$ into $\hat{\boldsymbol{z}}_{\mathrm{o}}$ and $\hat{\boldsymbol{z}}_{\mathrm{n}}$, in alignment with the proposed SCM. In this setup, $\hat{\boldsymbol{z}}_{\mathrm{o}}$ is approximated to only contain ordinal content information that determines the ordinal category, whereas $\hat{\boldsymbol{z}}_{\mathrm{n}}$ holds styling information. Generating content-preserving samples hinges on accurately recovering the true joint distribution of image and class, denoted as $P(\boldsymbol{X}, \boldsymbol{Y})$. Based on our proposed generative process, the causal factorization of the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{v}}, \boldsymbol{Z}_{\mathrm{n}})$ is:

$$P(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{v}}, \boldsymbol{Z}_{\mathrm{n}}) = P(\boldsymbol{Y})P(\boldsymbol{Z}_{\mathrm{v}})P(\boldsymbol{Z}_{\mathrm{n}})P(\boldsymbol{Z}_{\mathrm{o}}|\boldsymbol{Y}, \boldsymbol{Z}_{\mathrm{v}})P(\boldsymbol{X}|\boldsymbol{Z}_{\mathrm{o}}, \boldsymbol{Z}_{\mathrm{n}}). \tag{2.1}$$

Our method is designed to fulfill the this generative process by modelling each probability in Eq. 2.1.

2.3.1 Philosophy of Disentangling Ordinal Content Factors.

The principle of minimal change is pivotal in our generative process, aiding in the effective disentanglement of non-ordinal factors from ordinal content factors. To underscore the importance of this principle, we first discuss the limitations of relying solely on ordinal labels y. Subsequently, we explore scenarios that incorporate the *minimal change* principle. We consider a generative model (Zhou et al., 2023; Yao et al., 2021) that aligns with the generative process depicted in Figure 2.2 and infers the factors $\hat{\boldsymbol{z}}_{\mathrm{n}}$, $\hat{\boldsymbol{z}}_{\mathrm{o}}$, and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. After the learning phase, in which the reconstruction error is minimized, our goal is to align $\hat{\boldsymbol{z}}_{\mathrm{n}}$ with non-ordinal factors $\boldsymbol{z}_{\mathrm{n}}$; align $\hat{\boldsymbol{z}}_{\mathrm{o}}$ with ordinal content factors $\boldsymbol{z}_{\mathrm{o}}$; and align $\hat{\boldsymbol{z}}_{\mathrm{v}}$ with the set of invariant ordinal factors $\boldsymbol{z}_{\mathrm{v}}$. Note that $\boldsymbol{z}_{\mathrm{n}}$, $\boldsymbol{z}_{\mathrm{o}}$ and $\boldsymbol{z}_{\mathrm{v}}$ are the true latent factors in the data generative process.

If we rely solely on ordinal labels y, the generation of $\hat{\boldsymbol{z}}_{\mathrm{o}}$ is influenced by both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. This relationship can be mathematically expressed as: $\hat{\boldsymbol{z}}_{\mathrm{o}} = g(y, \hat{\boldsymbol{z}}_{\mathrm{v}}) + \epsilon$. In essence, this generation mechanism allows $\hat{\boldsymbol{z}}_{\mathrm{o}}$ to assimilate information from both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$. Given the generative model's typical assumption that both $\hat{\boldsymbol{z}}_{\mathrm{n}}$ and $\hat{\boldsymbol{z}}_{\mathrm{v}}$ follow a standard Gaussian distribution, non-ordinal information can feasibly reside in either $\hat{\boldsymbol{z}}_{\mathrm{v}}$ or $\hat{\boldsymbol{z}}_{\mathrm{n}}$. These factors are fundamentally similar, and their differences are purely notational. Hence, the choice of where to store non-ordinal information does not influence the reconstruction error. As a result, $\hat{\boldsymbol{z}}_{\mathrm{o}}$, being influenced by both y and $\hat{\boldsymbol{z}}_{\mathrm{v}}$, inevitably contains non-ordinal information. Thus, the disentanglement can not be achieved solely with y.

To address this, we apply the minimal change principle in the disentanglement process. The minimal change principle serves as a constraint on image generation, ensuring that the influence of certain factors during instance generation remains minimal. Specifically, when applied to $\hat{\boldsymbol{z}}_{\mathrm{o}}$, this principle limits the factor's influence. Consequently, the information within $\hat{\boldsymbol{z}}_{\mathrm{o}}$ remains minimal and focused. By introducing an additional constraint on the generative function g, we ensure that $\hat{\boldsymbol{z}}_{\mathrm{v}}$ inherently carries information related to Y. Assuming the reconstruction error is minimized, this suggests that ordinal content information is primarily included within the latent factors $\hat{\boldsymbol{z}}_{\mathrm{o}}$. As $\hat{\boldsymbol{z}}_{\mathrm{n}}$ is generated independently of the ordinal label Y, it will not contain ordinal content information. Therefore, all ordinal content information becomes localized within $\hat{\boldsymbol{z}}_{\mathrm{o}}$. Furthermore, by minimizing the influence of $\hat{\boldsymbol{z}}_{\mathrm{o}}$, it becomes exclusively representative of ordinal content information.

2.3.2 Ordinal Content and Non-Ordinal Information Disentanglement via Minimal Change

Our primary objective is to achieve *minimal change* in the generation process, specifically by constraining the influence of ordinal content factors $\hat{\boldsymbol{z}}_{\mathrm{o}}$. This is realized by limiting the number of these factors. To this end, we introduce a mask operation, consistent with our data generative process. The essence of this mask is to regulate the quantity of ordinal content factors. A sparser mask translates to fewer ordinal content factors. To promote this sparsity, we impose an L1 loss on the mask, represented as $\mathcal{L}_{\mathrm{sp}} = \|M\|_1$. Let's define our latent factors

as $\hat{\boldsymbol{z}} := [\hat{z}_o, \hat{z}_n]$ and $\hat{\boldsymbol{z}}_{on} := [\hat{z}_v, \hat{z}_n]$. The mask operation is then given by:

$$\hat{\boldsymbol{z}}_{\boldsymbol{on}} = \hat{\boldsymbol{z}} + \boldsymbol{M} \odot f_{\boldsymbol{y}}(\hat{\boldsymbol{z}}), \;\; \hat{z} \sim \mathcal{N}(0, \mathbf{I}), \;\; y \sim P(Y). \tag{2.2}$$

In the above equation, $\hat{\boldsymbol{z}}$ encapsulates both the invariant ordinal content factors $\hat{\boldsymbol{z}}_{\text{v}}$ and the non-ordinal factors $\hat{\boldsymbol{z}}_{\text{n}}$. The label distribution $P(Y)$ is derived empirically by counting the occurrences of each label within the dataset and then normalizing these counts to form a probability distribution. We utilize the Deep Sigmoidal Flow (Huang et al., 2018) for the function $f_{\boldsymbol{y}}$. It is a type of normalization flow characterized by the use of small neural networks with sigmoid units. These units introduce inflection points in the transformation function, enabling the modeling of complex probability distributions. Within our context, it serves as a component-wise transformation function that transforms $\hat{z}_{on}$ in a component-wise manner. This function, when applied, acts as a label influence. Although it impacts all elements in $\hat{\boldsymbol{z}}$, each element undergoes an independent transformation. The mask operation on $\hat{\boldsymbol{z}}$ rejects certain transformed elements. By adding $\hat{\boldsymbol{z}}$ back, we ensure that certain elements remain uninfluenced by the label Y, effectively distinguishing them as non-ordinal factors.

It's worth noting that for elements unaffected by the mask but influenced by y, the addition of elements from $\hat{\boldsymbol{z}}$ is inconsequential. Given that $\hat{\boldsymbol{z}}$ is sampled from a high-dimensional Gaussian distribution, adding it to these elements is akin to introducing random Gaussian noise, ensuring our method remains consistent with the proposed data generative process.

To make latent factors, we employ a Generative Adversarial Network (GAN) model (Mirza and Osindero, 2014). The architecture of this model is illustrated in Figure 2.3. The GAN comprises two main components: a generator G_θ and a discriminator D_ϕ, each parameterized by their respective learnable parameters θ and ϕ. The generator's role is to craft realistic instances, while the discriminator endeavors to differentiate between genuine and generated instances. The GAN loss, vital for image reconstruction, is formally articulated as:

$$\mathcal{L}_{\text{gan}} = \mathbb{E}[\log(D_\phi(\boldsymbol{x}))] + \mathbb{E}[\log(1 - D_\phi(G_\theta(\hat{\boldsymbol{z}}_{\boldsymbol{on}})))]. \tag{2.3}$$

In the given formulation, $D_\phi(\boldsymbol{x})$ represents the discriminator's estimated probability that the instance $\boldsymbol{x}$ is sampled from the real data distribution. The generator, denoted by G_θ, aims to

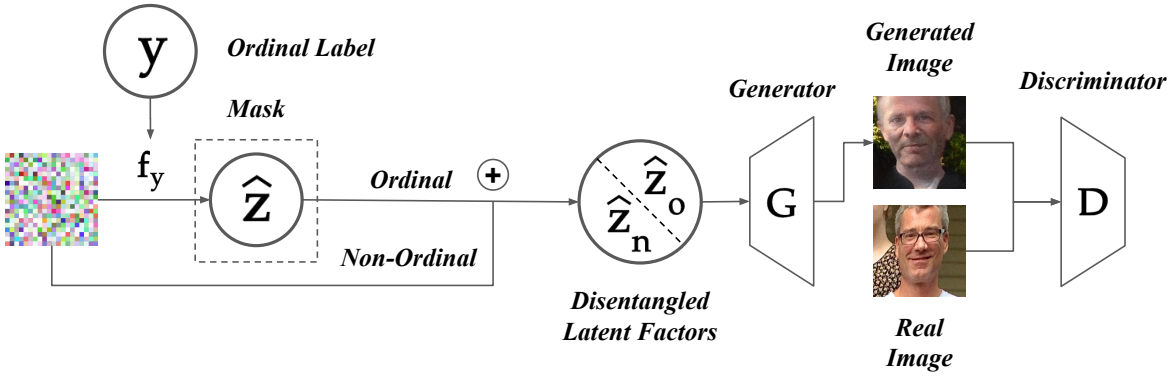

FIGURE 2.3. Architecture of Our Generative Model. The class label $\boldsymbol{y}$ is leveraged for both disentangling latent factors and enforcing minimal change. An ordinal head is appended to the discriminator to preserve the ordinal distribution of generated samples in relation to their class.

produce instances that the discriminator D_ϕ perceives as real, maximizing the likelihood of them being classified as genuine. Conversely, the discriminator D_ϕ endeavors to distinguish real instances from those generated by G_θ, classifying them accurately as either real or fake.

The objective function for disentangling ordinal content and non-ordinal information is combined with the GAN loss and the sparsity loss, i.e.,

$$\arg \min_{\{\phi,\theta,f_y,\boldsymbol{M}\}} \mathcal{L}_{\text{Aug}} = \arg \min_{\{\phi,\theta,f_y,\boldsymbol{M}\}} \mathcal{L}_{\text{gan}} + \lambda \cdot \mathcal{L}_{\text{sp}}. \tag{2.4}$$

where λ is the coefficient to control the contribution of the sparsity loss to the overall objective.

2.3.3 Content-Preserving Augmentation for Ordinal Regression

Our method is crafted to complement existing ordinal regression techniques, leveraging the strengths of contrastive learning. After training, we could have a generator $G_{\hat{\theta}}$. To generate an instance x' corresponding to a specific ordinal label, we employ Eq. 2.2. For instance, generating an example for the ordinal label $Y = 1$ can be achieved by:

$$\boldsymbol{x}'_{\boldsymbol{i}} = G_{\hat{\theta}}(\hat{z}_{on}), \;\; \hat{z}_{on} = \hat{\boldsymbol{z}} + \boldsymbol{M} \odot f_{\mathbf{1}}(\hat{\boldsymbol{z}}), \;\; \hat{z} \sim \mathcal{N}(0, \mathbf{I}), \;\; Y = 1. \tag{2.5}$$

Specifically, we start by sampling $\hat{z}_{on}$. By setting the ordinal label $Y = 1$, we utilize $f_{\mathbf{1}}(\hat{\boldsymbol{z}})$ to generate $\hat{z}_{on}$ specific to the ordinal label $Y = 1$. By sampling different $\hat{z}_{on}$ values and

maintaining the ordinal label $Y = 1$ constant, we can generate diverse instances that, while exhibiting stylistic variations, consistently belong to the label Y.

To integrate with existing ordinal regression methods, x' can be employed as the strongly augmented data. We then incorporate the supervised contrastive loss (Khosla et al., 2020) as a regularization term for the prevailing method. This loss emphasizes intra-class similarities while concurrently maximizing inter-class disparities. Let's define $\mathbf{X}$ as the feature space of x. We introduce h_ψ as a model with learnable parameters ψ, which is used by the ordinal regression method. For a given instance x, $h_\psi(x) = z$ outputs the latent representation z used for ordinal regression. The contrastive loss on h_ψ is defined as:

$$\mathcal{L}^{\text{con}} = \sum_{i \in I} \mathcal{L}_i^{\text{con}} = -\sum_{i \in I} \frac{1}{|\mathcal{S}(h_\psi(\boldsymbol{x_i}))|} \sum_{h_\psi(\boldsymbol{x_s}) \in \mathcal{S}(h_\psi(\boldsymbol{x_i}))} \log \frac{\exp\left(h_\psi(\boldsymbol{x_i}) \cdot h_\psi(\boldsymbol{x_i'})/\tau\right)}{\sum_{b \in B} \exp\left(h_\psi(\boldsymbol{x_i}) \cdot h_\psi(\boldsymbol{x_b})/\tau\right)}. \tag{2.6}$$

In the equation above, I denotes the set of sample indices in the batch. For each instance $\boldsymbol{x_i}$ belonging to y, its strongly augmented counterpart x_i' is generated using our method by setting the ordinal label $Y = y$ during the generation process.

2.4 Experiments

In this section, we evaluate our method across three real-world applications within the domain of ordinal regression: age estimation, diabetic retinopathy rating, and weather condition prediction. Due to space constraints, we include qualitative analyses of our generative model in Appendix A4.

Baselines. We employ five state-of-the-art deep learning-based ordinal regression methods as our baselines. OR-CNN (Niu et al., 2016) utilizes a series of binary classifiers and optimizes the model through the one-hot encoding of labels. CNNPOR (Liu et al., 2018b) reduces the multi-class negative log-likelihood while concurrently maintaining the intrinsic ordinal relationship among instances. SORD (Diaz and Marathe, 2019) employs a soft labeling strategy during training. POE (Li et al., 2021) captures data uncertainty via probabilistic embeddings. MWR (Shin et al., 2022) leverages an auxiliary set of reference images to model

ordinal relationships. All these models are end-to-end trainable. We seamlessly integrate our contrastive learning objective into their original loss formulations, augmented with ordinal content-preserving data transformations.

Experimental Settings. For the generative model, we use StyleGAN2 (Karras et al., 2020) as the base model, λ_1 is set to 1e-4 across all settings. For all ordinal regression methods, we use VGG16 (Simonyan and Zisserman, 2014) the base deep neural network architecture, with ImageNet (Deng et al., 2009) pre-trained weight for initialization. We employ an embedding layer before the final output layer in the model to extract feature embeddings. The dimension of feature embedding is set to 128. The ratio of contrastive loss is consistently set to 1e-4 for OR-CNN, CNNPOR and POE, and 1e-5 for SORE and MWR. For the three datasets, the input images are resized into 256×256 and center cropped into a sub-region of 224×224. Adam (Kingma and Ba, 2014) optimizer is used for all baseline methods, with a base learning rate of 1e-4. We uniformly train all baseline models for 200 epochs with a batch size of 256 for all baselines except MWR. For MWR, the batch size is set to 128 due to memory constraints. We report the results via the accuracy and mean absolute error (MAE) metrics. For the other parameters in the baselines, we adhere to the original settings designed in the papers unless specified in our experimental settings. While we employ our techniques for strong augmentations, weak augmentations are achieved solely through resizing, center cropping, and normalizing the original instances. No other augmentation methods are applied to the data. All experiments are conducted in on two 48GB NVIDIA RTX A6000 GPUs.

2.4.1 Age Estimation

Dataset. Age estimation is the task of predicting age groups based on facial images. The Adience dataset (Eidinger et al., 2014) comprises 26,580 photos from Flickr, featuring 2,284 subjects. These photos are annotated across eight age groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60 years. The dataset adheres to a standard five-fold, subject-exclusive cross-validation protocol, as widely utilized in previous studies (Rothe et al., 2018; Shen et al., 2018; Li et al., 2019; Li et al., 2021). The generative model is trained in accordance with the training fold of this protocol. For each instance in the dataset, we generate 3 augmented

Adience	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	54.6 ± 5.5	0.60 ± 0.09	57.1 ± 5.1 *(+2.5)*	0.56 ± 0.06 *(+0.04)*
CNNPOR (Liu et al., 2018b)	55.1 ± 6.0	0.60 ± 0.08	57.7 ± 4.2 *(+2.6)*	0.55 ± 0.07 *(+0.05)*
SORD (Diaz and Marathe, 2019)	57.8 ± 4.9	0.53 ± 0.06	59.9 ± 5.0 *(+2.1)*	0.49 ± 0.06 *(+0.04)*
POE (Li et al., 2021)	60.5 ± 4.8	0.47 ± 0.08	63.7 ± 4.6 *(+3.2)*	0.43 ± 0.07 *(+0.04)*
MWR (Shin et al., 2022)	62.6 ± 5.0	0.45 ± 0.08	63.6 ± 4.7 *(+1.0)*	0.43 ± 0.07 *(+0.02)*

Diabetic Retinopathy	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	71.9 ± 1.3	0.42 ± 0.01	72.8 ± 0.7 *(+0.9)*	0.41 ± 0.00 *(+0.01)*
CNNPOR (Liu et al., 2018b)	71.3 ± 1.1	0.42 ± 0.02	72.6 ± 1.0 *(+1.3)*	0.41 ± 0.01 *(+0.01)*
SORD (Diaz and Marathe, 2019)	69.1 ± 1.0	0.45 ± 0.01	69.9 ± 1.1 *(+0.8)*	0.44 ± 0.01 *(+0.01)*
POE (Li et al., 2021)	73.6 ± 1.0	0.40 ± 0.01	74.8 ± 0.8 *(+1.2)*	0.38 ± 0.00 *(+0.02)*
MWR (Shin et al., 2022)	74.5 ± 1.1	0.38 ± 0.02	75.1 ± 1.1 *(+0.6)*	0.37 ± 0.01 *(+0.01)*

SkyFinder	w/o OCP-CL		w/ OCP-CL	
	Accuracy (↑)	MAE (↓)	Accuracy (↑)	MAE (↓)
OR-CNN (Niu et al., 2016)	60.3 ± 2.1	0.48 ± 0.03	62.1 ± 2.3 *(+1.8)*	0.46 ± 0.04 *(+0.02)*
CNNPOR (Liu et al., 2018b)	57.6 ± 1.6	0.52 ± 0.03	59.7 ± 1.5 *(+2.1)*	0.49 ± 0.03 *(+0.03)*
SORD (Diaz and Marathe, 2019)	58.2 ± 1.9	0.51 ± 0.06	60.5 ± 2.0 *(+2.3)*	0.48 ± 0.04 *(+0.03)*
POE (Li et al., 2021)	61.9 ± 1.7	0.46 ± 0.05	64.1 ± 1.6 *(+2.2)*	0.42 ± 0.05 *(+0.04)*
MWR (Shin et al., 2022)	62.4 ± 1.8	0.45 ± 0.05	63.2 ± 1.9 *(+0.8)*	0.44 ± 0.06 *(+0.01)*

TABLE 2.1. Accuracy (%) and MAE comparison on Adience dataset (Eidinger et al., 2014), Diabetic Retinopathy dataset (Liu et al., 2018a) and SkyFinder dataset (Mihail et al., 2016).

views using the generative model, the augmented views and the original instances are jointly trained by the models.

Results. We present the experimental results in Table 2.1 (Top). Employing our proposed OCP-CL method, OR-CNN experiences a 4.58% boost in accuracy and a 6.67% reduction in MAE. CNNPOR benefits from a 4.72% increase in accuracy and an 8.33% improvement in MAE. SORD's performance is uplifted by 3.63% in accuracy and 7.55% in MAE. POE sees the largest accuracy improvement of 5.29% and an MAE reduction of 8.51%. Lastly, MWR has a modest 1.6% increase in accuracy and a 4.44% decrease in MAE. This consistent

improvement across multiple ordinal regression methods validate the efficacy of our OCP-CL approach for the task of age estimation. Additionally, we visualise the generative augmentations in Figure 2.1. We observe that the augmentations have preserved the ordinal content information for their respective age groups, capturing details such as the sparse eyebrows of children, silky skin texture of young adults, and the pronounced wrinkles of seniors, thereby allowing the ordinal regression methods and the contrastive learning framework to effectively learn the critical ordinal content information.

2.4.2 Diabetic Retinopathy Rating

Dataset. The Diabetic Retinopathy dataset[1] is utilized for predicting the severity stages of Diabetic Retinopathy based on high-resolution RGB retina images. The dataset consists of 35,126 individual instances, each annotated into one of five ordinal categories representing increasing levels of severity (*i.e.*, *No DR*, *Mild*, *Moderate*, *Severe*, and *Proliferative DR*). The dataset is partitioned into training, validation, and testing sets, which constitute 80%, 5%, and 15% of the total dataset, respectively. The dataset contains 25,810, 2443, 5292, 873 and 708 images for each category, respectively. Account for the imbalances between adjacent categories, we generate augmented views dynamically depending on the ground truth label, which mitigates the class imbalance issue. Specifically, the number of augmentations for instances from each increasing level of severity is set as [1, 3, 2, 5, 5].

Results. In Table 2.1 (Middle), we evaluate the performance of various ordinal regression methods on the Diabetic Retinopathy dataset, with and without the incorporation of our proposed OCP-CL (Ordinal Content-Preserving Contrastive Learning) module. Remarkably, all the compared methods exhibit improvement in both accuracy and MAE upon integration with the OCP-CL module. Practicularly, the incorporation of the OCP-CL module results in accuracy improvements of 1.25%, 1.82%, 1.16%, 1.63%, and 0.81% for OR-CNN, CNNPOR, SORD, POE, and MWR, respectively. Concurrently, the MAE reduces by 2.38%, 2.38%, 2.22%, 5.00%, and 2.63%, respectively. These results collectively indicate that the introduction of the OCP-CL module consistently enhances the performance across a diverse set of

[1]Accessible from https://www.kaggle.com/competitions/diabetic-retinopathy-detection

Age Group: 4-6 *Age Group: 25-32* *Age Group: 60+*

FIGURE 2.4. Generated augmentations for the age estimate task, the collections corresponds to the age group of (4-6), (25-32), and 60+ respectively.

ordinal regression models. This validates the generalizability and efficacy of our proposed OCP-CL approach in boosting performance for ordinal regression tasks.

2.4.3 Weather Condition Prediction

Dataset. The SkyFinder Dataset (Mihail et al., 2016) comprises 94,804 labeled outdoor images, sourced from 53 static webcams affiliated with the Archive of Many Outdoor Scenes (AMOS). These images encapsulate a broad spectrum of weather and lighting conditions. A specialized subset of 62,988 images, specifically featuring the weather conditions of *Clear*, *Partly Cloudy*, and *Mostly Cloudy*, has been curated to create a weather prediction dataset. This subset is further partitioned into training, validation, and testing sets, constituting 80%, 5%, and 15% of the dataset, respectively. For each instance in the dataset, we generate 3 augmented views using the generative model, the augmented views and the original instances are jointly trained by the models.

Results. Table 2.1 (Bottom) presents the results of our experiments, highlighting the performance improvements achieved by all baseline models upon the incorporation of the contrastive module. Specifically, the accuracy improvements for the baselines are 2.9%, 3.5%, 3.8%, 3.4%, and 1.1% for OR-CNN, CNNPOR, SORD, POE, and MWR, respectively. Similarly, the improvements in MAE for the baselines are 4.2%, 5.8%, 5.9%, 8.7%, and 2.2%, respectively. With an average improvement of 2.94% in accuracy and 5.42% in MAE, these results demonstrate the efficacy of our method in enhancing the performance of deep-learning-based ordinal regression models on the weather condition estimation task.

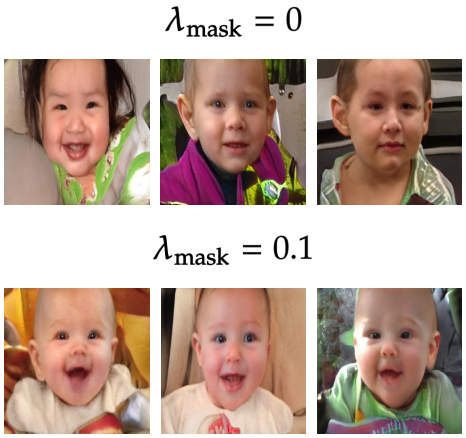

FIGURE 2.5. Influence of *minimal change* in image generation.

2.4.4 Analysis

Transfer Learning. In this section, we evaluate the transfer learning performance of our contrastive learning approach. Initially, we pre-train the encoder using a contrastive learning objective for feature extraction. Following this, we freeze the trained encoder and employ the extracted features as input to a single-layer MLP predictor, which is then fine-tuned on the training data. We assess the efficacy of our approach against recent state-of-the-art supervised contrastive learning frameworks across three different tasks. To mitigate performance degradation due to parameter settings, we dynamically adopt the recommended configurations from the original papers. However, for SupCon (Khosla et al., 2020), a batch size of 1024 is unfeasible for image instances of size 224 by 224. To ensure convergence, we reduced the image size to 64 by 64 and the batch size to 512. The results are presented in Table 2.2. Notably, our method significantly outperforms all existing approaches in the task of ordinal regression. The benefits of ordinal content-preserving data augmentation become evident when benchmarked against SupMoCo (He et al., 2019). Specifically, we adopt the SupMoCo framework as the baseline contrastive learning framework and integrate our augmentation strategy by replacing the original data augmentation modules. This aids in evaluating transfer learning performance. The contrastive loss formulation in SupMoCo aligns with our Eq. 2.6, and an additional momentum encoder is incorporated to ensure training convergence. The MoCo strategy is not employed in other experiments.

Dataset	SupMoCo (He et al., 2019)		SupCon (Khosla et al., 2020)		S-LooC (Xiao et al., 2021)		SupCReg (Zha et al., 2022)		OCP-CL (Ours)	
	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy (↑)	MAE (↓)
DR	63.6	0.52	60.5	0.55	63.0	0.52	62.7	0.53	**65.2**	**0.50**
Adience	51.9	0.65	51.7	0.67	52.4	0.64	51.2	0.63	**53.1**	**0.61**
SkyFinder	56.1	0.55	53.9	0.59	55.6	0.56	55.1	0.55	**57.5**	**0.52**

TABLE 2.2. Linear evaluation on supervised contrastive learning frameworks. Accuracy (%) and MAE are reported for various ordinal datasets including Diabetic Retinopathy dataset, Adience (Levi and Hassner, 2015) and SkyFinder dataset (Mihail et al., 2016).

Minimal Change in Image Generation. We study the effect of minimal change on ordinal data generation by manipulating the mask hyperparameter. When this hyperparameter is set to zero, sparsity is not enforced, effectively removing the minimal change constraint from the generative process. As illustrated in Figure 2.5 ($\lambda_{\text{mask}} = 0 \rightarrow$ no minimal change), the absence of minimal change leads to the inclusion of non-ordinal features, such as hair, which do not contribute to identifying infant age groups and should be considered non-ordinal factors. However, in the case where minimal change is not applied, these features are learned as ordinal content factors and appear in all generated images, thereby demonstrating poor disentanglement performance. By enforcing minimal change, these non-ordinal elements are suppressed, enhancing the overall quality of the generated instances.

2.5 Conclusion

In this paper, we address the open challenge of applying contrastive learning to ordinal regression tasks. We find that the strong data augmentations in the contrastive learning frameworks often diminish the intrinsic discriminative semantic information associated with ordinal labels. Consequently, when contrastive learning is used to identify invariant features between weakly and strongly augmented views, the extracted features frequently lack the essential ordinal content information. To mitigate this issue, we introduce a novel augmentation method grounded in the principle of *minimal change*. This generative approach ensures that the images retain the essential ordinal content information during the data augmentation process. As a result, our method enhances the applicability of contrastive learning to ordinal regression tasks.

Extensive experiments validate the efficacy of this approach in improving the performance of existing ordinal regression models. This work not only broadens the scope of contrastive learning in ordinal regression but also provides valuable insights for future research aimed at preserving crucial task-specific information during data augmentation.

CHAPTER 3

Structured Latent Reasoning in Large Vision Models

This chapter investigates structured latent reasoning in large autoregressive vision models, focusing on how intermediate, interpretable structure can be introduced into visual in-context learning. Positioned after the ordinal learning chapter, it shifts from task-specific representation control to reasoning-time control within general-purpose vision models. The chapter identifies a key limitation of existing Large Autoregressive Vision Models: while capable of few-shot visual prediction, their prompting mechanisms lack explicit structure for progressive reasoning and informative context selection. To address this, the chapter proposes Chain-of-Focus prompting, a visual analogue of chain-of-thought prompting that decomposes visual context into ordered, salient intermediate steps and selects prompts based on relevance and annotation richness. Through this formulation, the chapter demonstrates how latent visual representations can be organized to support structured, interpretable reasoning without modifying model parameters, advancing the thesis argument that controllability in large vision models can be achieved through principled structuring of latent and contextual information.

3.1 Introduction

Utilizing a pre-trained, general-purpose vision model to perform multiple downstream visual tasks with only a few illustrative examples represents a significant advancement toward artificial general intelligence. Recently, the emergence of Large Autoregressive Vision Models (LAVMs) (Bai et al., 2024; Guo et al., 2024) has presented a promising approach for achieving this unification of tasks. The principle behind this integration involves building an autoregressive model (Touvron et al., 2023b) that enables visual in-context learning (Bar

et al., 2022; Zhang et al., 2023c; Wang et al., 2023a; Li et al., 2024a), where given a test input and a pair of prompts containing an input image and its visualized target annotation, the vision models endeavor to recognize the visual patterns between the prompt image and its target, thereby making analogous predictions on the test image.

In the realm of large language models (LLMs), in-context learning (ICL) has been extensively studied (Dong et al., 2022b). Among these approaches, Chain-of-Thought (CoT) prompting (Wei et al., 2022; Wang et al., 2022b; Zhang et al., 2022b) is one of the most influential methods, significantly enhancing the predictive abilities of LLMs by introducing intermediate reasoning steps within the contextual language prompts. Given that LLMs and LAVMs share similar autoregressive architectures, we are inspired to explore whether injecting intermediate steps into visual contextual prompts can similarly unlock the capabilities of LAVMs. Building upon the principles of CoT prompting, we propose Chain-of-Focus (CoF) prompting, a novel prompting method tailored for LAVMs.

Nevertheless, implementing contextual and sequential prompts in the vision domain presents two significant challenges. First, unlike text, which follows syntactic and semantic rules, visual data inherently lacks the clear logical structure, making it difficult to decompose and sequence for step-by-step interpretation. Second, in the language domain, hand-crafted prompts can be tailored specifically to the test input by providing analogous examples that closely relate to the problem at hand. For instance, if the test input for LLMs is a geometry problem, the language prompt can include a similar geometry problem with its solution, making the answer more informative to the model for analogy-based predictions. This level of customization is challenging in the visual domain, as images cannot be easily modified or restructured to fit new test inputs.

CoF prompting addresses the first challenge by adapting a cognitive strategy that is fundamental to human visual understanding: visual salience, which enables individuals to sequentially process visual information and draw intermediate conclusions based on the prominence of salient objects in a scene (Wertheimer and Riezler, 1944). For example, when viewing an image of a kitchen containing numerous objects, an observer's attention will initially focus on larger and closer items, such as the benchtop and chairs placed in front of it,

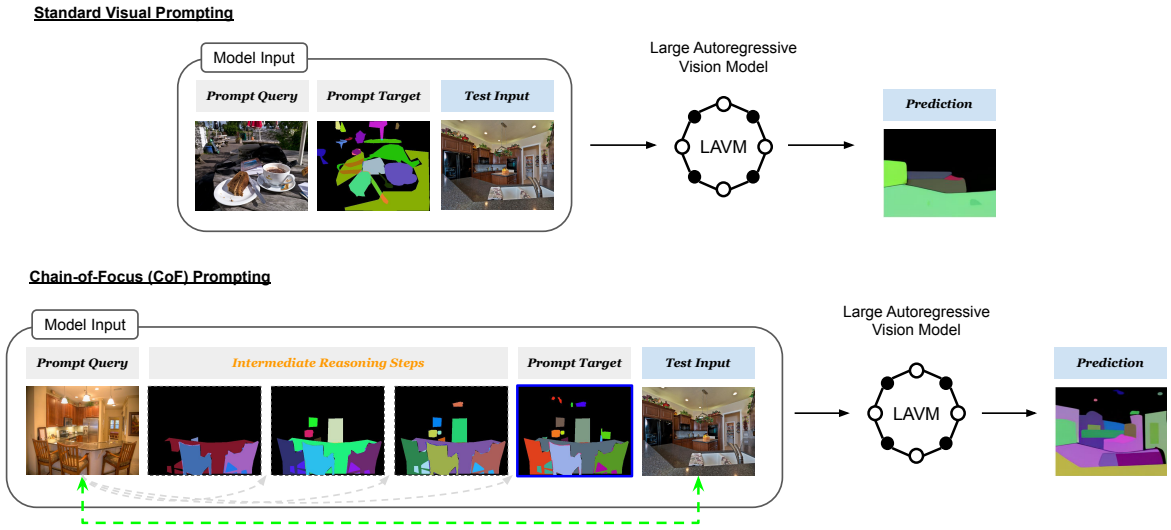

FIGURE 3.1. Illustration of Chain-of-Focus (CoF) prompting. The top section illustrates the current strategy for prompting LAVMs, where the prompt query (image) is randomly selected for the test input, and the task-specific prompt targets are visualized to form a prompt pair, enabling LAVMs to make in-context, analogy-based predictions. CoF prompting (bottom section) generates intermediate steps leading to the prompt target while selecting informative prompt pairs based on prompt query similarities to the test input and the richness of usable information contained in the prompt target.

before shifting to smaller appliances. As illustrated in Figure 3.1, CoF prompting replicates this cognitive process through generating intermediate reasoning steps within the prompt targets by ranking the salient regions of the prompt image in descending order. Specifically, we generate a saliency probability map using a pre-trained saliency detection model (Qin et al., 2020) to obtain the order of salient regions in the prompt image. Incrementally annotating different parts of the image based on saliency scores to create intermediate steps, allowing the models to build context progressively and enhance their predictive capabilities.

On the other hand, in the language domain, it has been shown in CoT prompting that finding informative prompt queries is crucial for enhancing LLM's predictive accuracy. Inspired by this, in CoF prompting for visual inputs, we utilize two selection criteria to search for the most informative prompts relative to the test input. First, we consider image relevance, which measures how semantically related the prompt image is to the test input image. Prior research (Zhang et al., 2023c) has demonstrated that images sharing similar semantic meanings

with the test input serve as better illustrations, enabling the model to draw more accurate analogies. However, we find that for certain downstream tasks, these semantically similar images may have sparse annotations, meaning they cannot provide sufficient knowledge to the model. Therefore, we introduce the second criterion, annotation richness, to ensure that the selected prompt images contain comprehensive annotations useful for the test case. By integrating both image relevance and annotation richness, our approach addresses the challenge of creating tailored visual prompts, enhancing the model's ability to generalize from a few examples to unseen inputs.

We build our method upon the framework of Large Autoregressive Vision Models (LAVMs) (Bai et al., 2024; Hao et al., 2024), leveraging their ability to perform simultaneous predictions across multiple downstream tasks within one single pre-trained model. To quantify the similarity between the prompt image and the test image, we employ the encoder from the pre-trained LAVMs and evaluate the distance between their encoded representations. This encoder transforms raw images into discrete indices within a codebook via vector quantization(Esser et al., 2021; Van Den Oord, Vinyals et al., 2017). By treating these codebooks as sets and calculating the intersection over union between them, we effectively capture semantic equivalence while disregarding the specific order of indices. After identifying prompts similar to the test input, we assess the richness of prompt annotations by examining the diversity of entries in the prompt targets' codebooks. This approach ensures that the selected visual prompts are not only highly relevant but also possess rich annotations, thereby enhancing the in-context performance of the LAVMs.

To summarize our contributions, we propose a new visual prompting paradigm called Chain-of-Focus (CoF) prompting. Our approach mimics progressive thinking by incorporating intermediate steps into visual prompts and addresses the challenge of prompt customization by directly selecting the most informative prompts relative to test inputs. Our method can be seamlessly integrated with the recently proposed Large Autoregressive Vision Models (LAVMs) (Bai et al., 2024; Hao et al., 2024) through visual in-context learning, significantly improving their performance on downstream visual tasks.

3.2 Related Works

In-Context Learning and CoT Prompting In-context learning (ICL) (Huang et al., 2024b; Wang et al., 2024a) is a paradigm where models learn to perform tasks by conditioning on examples provided in the input context during inference. Rather than relying on traditional training processes with gradient updates, the models leverage the contextual information from query-target pairs presented at inference time to make predictions on new test inputs. In the language domain, recent advancements have highlighted the effectiveness of hierarchical reasoning techniques, known as Chain-of-Thought prompting, in enhancing the performance of large language models (LLMs) (Kojima et al., 2022; Wang et al., 2022b; Wei et al., 2022; Zhang et al., 2023a; Luo et al., 2024; Zhang et al., 2024b; Lin et al., 2025; Tu et al., 2024). These methods leverage sequential reasoning steps to improve inference. Inspired by these developments, researchers have extended hierarchical reasoning frameworks to the vision-language domain (Lu et al., 2022; Zhang et al., 2023e). Among these, the most related stream of works to ours has attempted to explore the rationale within or across images and express them in textual descriptions (Ge et al., 2023; Mitra et al., 2023; Rose et al., 2023; Zheng et al., 2023a). This integration of visual information into its language counterpart has yielded significant improvements for large vision-language models (LVLMs) (Liu et al., 2023b; Zhang et al., 2022a; Zhang et al., 2024c; Zhou et al., 2024), yet it also reveals the challenges of applying CoT-based methods directly to the pure vision domain (i.e., expressing reasoning without the use of language). Unlike language, images lack explicit symbolic structures, making it challenging to express reasoning steps as in LLMs or LVLMs. In purely visual contexts, Zhang et al. (Zhang et al., 2023c) develop a prompt retrieval framework for selecting in-context examples that maximize models' performance. Chain-of-Spot (Liu et al., 2024b) develops a multimodal promtping method for LVLMs. It leverages language prompts to use only regions of interest (ROIs) for visual understanding. Chain-of-Sight (Huang et al., 2024c) introduces a purely visual framework that employs a sequence of visual resamplers to capture visual details at different spatial levels, generating tokens across multiple scales.

Large Autoregressive Vision Models The inspiration behind autoregressive vision models stems from the advancements of large language models (LLMs) (Brown et al., 2020; Touvron

et al., 2023b; Touvron et al., 2023a). Using contextual information, LLMs are able to capture long-range dependencies and make coherent predictions with sequential modelling techniques. Building on this concept, Bai et al. (Bai et al., 2024) propose Large Autoregressive Vision Models (LAVMs), which adapt this modelling strategy to the visual domain by constructing 'visual sentences' that enable sequential prediction. This approach involves representing visual inputs as sequences of tokens, analogous to the text tokens used in LLMs. By processing visual data sequentially, the model employs self-attention mechanisms to understand dependencies and relationships within the visual context, thereby enabling effective in-context learning from purely visual inputs. By including query-target pairs from different downstream tasks in these visual sentences, the model can accomplish various visual downstream tasks within a single framework. Hao et al. (Hao et al., 2024) extend the work and introduce a data-efficient LAVM, which is designed to operate effectively on limited datasets by making use of data augmentation and knowledge distillation. The primary purpose of LAVMs is to unify all vision tasks within a single model, making the adaptation to downstream tasks highly efficient.

3.3 Chain-of-Focus (CoF) Visual Reasoning

3.3.1 Preliminaries

The Large Autoregressive Vision Model (LAVM) (Bai et al., 2024) is a foundational vision model that synthesizes visual predictions through sequential modeling, inspired by the successes of Large Language Models (LLMs). In LLMs, an autoregressive model predicts the next word in a sentence based on previous words. Similarly, LAVM aims to predict the next visual token in a visual sequence given the previous tokens. This is achieved using a tokenization network $E : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{n\times d}$ that transforms raw images $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{h\times w\times c}$ into visual tokens $Z = \{z_1, z_2, \ldots, z_n\} \in \mathbb{R}^{n\times d}$, followed by a sequential model $f : \mathbb{R}^p \to \mathbb{R}$ that predicts outputs in an autoregressive manner $z_t = f(z_{t-1}, z_{t-2}, \ldots, z_{t-p}) + \varepsilon_t$, where p is the total number of previous time steps in the sequence, t is the current step, and ε is the noise. The predictions are then detokenized back to pixel space by a decoder network $D : \mathbb{R}^{n\times d} \to \mathbb{R}^{h\times w\times c}$.

In implementations of LAVMs (Bai et al., 2024; Hao et al., 2024), a pre-trained VQ-GAN (Esser et al., 2021) model is employed as the tokenizer. The VQ-GAN model encodes the image into a discrete codebook, with the indices in the codebook serving as the tokens for the autoregressive model. The pre-trained VQ-GAN decoder then decodes the codebook/tokens back into pixel space for generating images. At its core, the autoregressive model in LAVM utilizes a causal transformer (Touvron et al., 2023b) which employs causal masking to compute each token's representation based solely on itself and the preceding tokens, thereby preserving the sequence's temporal order. This allows the model to capture dependencies and patterns within the data effectively, enhancing its ability to generate coherent sequences during inference.

The visual sentences used to train LAVM are either derived from natural visual sequences, such as videos or multi-views of a 3D object (Zhan et al., 2022), or handcrafted by connecting raw images with their target annotation pairs from various visual downstream tasks. This allows the model to adapt to any downstream task given images (a.k.a. prompt queries x_{pq}) and annotations (a.k.a. prompt targets x_{pt}). At the inference stage, LAVM employs prompted inference. Given several examples of image and target annotation pairs, the tokenizer first transforms each input into tokens and constructs a visual sentence using the paired image and annotation data. The test input is appended at the end of the visual sentence as the last token. This sentence is then passed into the autoregressive network for the prediction of the next token in the sequence. The predicted tokens are subsequently constructed into a codebook and decoded into pixel space.

3.3.2 Saliency-based Intermediate Reasoning Steps

Sequential Prompt Construction In our approach, we construct prompts that not only present visual queries and targets but also sequentially introduce reasoning steps. Each prompt consists of a visual query x_{pq}, and a series of m intermediate reasoning steps $\{x_{pt}^1, x_{pt}^2, \ldots, x_{pt}^m\}$ leading up to the final target x_{pt}. This setup mimics human reasoning processes, where intermediate conclusions are drawn before reaching a final decision. In practice, when constructing the model input with the intermediate steps, we find that the best order is denoted as: $[x_{pq}, x_{pt}^1, x_{pq}, x_{pt}^2, \ldots, x_{pq}, x_{pt}^m, x_{tq}]$, where the prompt query and intermediate targets are

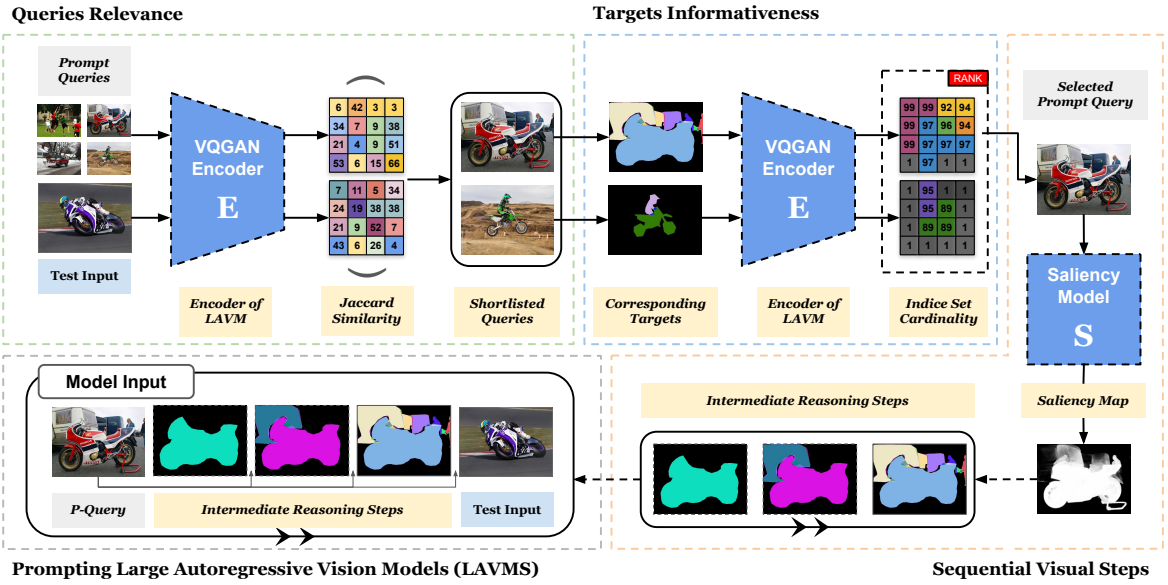

FIGURE 3.2. Illustration of Generating CoF Prompts. The framework can be viewed in two steps. First, CoF identifies a set of the most relevant queries to the test input and assesses the informativeness of their targets to filter out less informative prompt pairs. This step ensures that the prompts are highly relevant and informative to the test input. In the second step, CoF uses a saliency-based strategy to create intermediate steps for the answers to the query, which implicitly injects sequential visual cues into the prompt targets. CoF follows the general structure of Chain-of-Focus prompting, with improvements in automating the process of both prompt selection and intermediate steps generation.

ordered alternately, with the test query x_{tq} appended at the end of the sequence. We suggest that the optimal construction order depends on the pre-trained model itself. The pre-trained LAVMs (Hao et al., 2024; Bai et al., 2024) are primarily trained on visual sentences in the format of query and target pairs, thus its sequential prediction ability is restricted to paired representations. Similarly, finetuning the pre-trained LAVMs with natural reasoning steps allows for a different construction of prompts: $[x_{pq}, x^1_{pt}, x^2_{pt}, \ldots, x^m_{pt}, x_{tq}]$, which follows the natural sequential order. These intermediate reasoning steps decompose the complex answers into sub-pieces for understanding.

Visual Reasoning via Exploring Salient Regions To simulate a cognitive reasoning process, we generate a sequence of intermediate answers using visual saliency information. Given a visual query x_{pq} and its corresponding answer x_{pt}, where both x_{pq} and x_{pt} are images.

We utilize the salient regions within these images for constructing informative prompts. To quantitatively assess the salience of different regions within the images, we compute a saliency score $\sigma(r)$ for each region r, where the regions are defined by the masks on objects of interest in the image. In tasks such as image segmentation and pose estimation, the auxiliary information on masks are often provided with the ground truth as their segmentation masks and bounding boxes. We use a pre-trained saliency detection model (Qin et al., 2020) to obtain a saliency probability map for the image. For each region, we compute the saliency score as:

$$\sigma(r) = \sum_{i,j} M_r(i,j) \cdot S(x_{pq}). \tag{3.1}$$

$M_r(i,j)$ is the mask for the region r, where $M_r(i,j) = 1$ if the pixel (i,j) is within the region r and $M_r(i,j) = 0$ for pixels that do not belong to the region. The function $S : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{h\times w}$ extracts the pixel-wise saliency probability scores from the prompt query x_{pq} and forms the probability map. The function $\sigma(r)$ computes the summed probability for the masked area as the region saliency score. We label the regions in an incremental manner using the saliency scores. For each intermediate step, the target is given by:

$$x_{pt}^{t+1} = x_{pt}^{t} \cup \{r \mid \sigma(r) > \tau_{t+1}\}, \tag{3.2}$$

where τ_t is a saliency threshold for step t, defining the minimum saliency required for a region to be included in the intermediate target x_{pt}^{t}. In the last step, all regions in the image will be labelled, providing a complete prompt target. This ordered introduction of information helps the LAVM to focus on relevant features at each step, allowing it to build context progressively. By leveraging saliency-based cognitive pathway, we aim to mimic the hierarchy focusing observed in human visual attention, enhancing the understanding of visual content through structured, human-like reasoning.

3.3.3 Informative Visual Prompts

In chain-of-thought (CoT) prompting, selecting relevant queries is crucial as it directly impacts the quality of the generated responses. Traditionally, CoT involves manually choosing prompt queries for each test input, a process that ensures alignment with desired outcomes but is

labor-intensive and prone to human bias. In our method, we aim to automate this process by selecting the most relevant and informative visual query and target pairs to the test input, thereby enhancing the in-context learning performance of LAVMs. The following details our strategy for selecting visual query and target pairs to serve as the prompts for inference.

Selection of Relevant Queries Given a test query x_{tq} and a candidate pool of prompt pairs consisting of prompt queries $x_{pq} \in X_{pq}$ and prompt targets $x_{pt} \in X_{pt}$, our goal is to first shortlist a subset of prompts contain queries that is similar to the test query. To this end, we employ the same VQGAN encoder from the LAVM framework to serve as the feature extractor for the prompt queries and the test query. The encoder transforms the queries into discrete codebooks $\{z_{tq}, z_{pq_1}, z_{pq_2}, \ldots, z_{pq_n}\}$. Each entry in the codebook is a discrete generative factor that corresponds to the pixel space, therefore, more aligned entries in the two codebooks of queries indicate that the two queries contain similar objects or scenes in the pixel space. Through manual testing, we find that the relative position of the objects and the number of the objects in the prompt query do not affect the performance of inference as long as the two queries are semantically aligned. Hence, we convert the codebooks into sets and measure the similarity of each encoded prompt query z_{pq} and z_{tq} using the Jaccard similarity index, which is defined as:

$$J(z_{tq}, z_{pq}) = \frac{|z_{tq} \cap z_{pq}|}{|z_{tq} \cup z_{pq}|}. \tag{3.3}$$

This measure counts the number of unique indices shared between z_{tq} and z_{pq} without considering the position of the indices in the codebook. The set operation also helps avoid over-representation of redundant and repeating background features that are not pertinent to the task. Through this process, we shortlist a subset of N prompts that have queries similar to the given test query.

Selection of Rich Targets Once we have selected the N most similar queries, we need to further refine our selection for target informativeness, that is to ensure the chosen answers are providing rich information for inference. As observed in tasks such as image segmentation and keypoint detection, the presence of diverse and richly annotated segmentation masks is crucial for effective in-context learning. We quantify the informativeness of a prompt target x_{pt} by assessing the diversity of its encoded discrete representation z_{pt}. The intuition behind

this involves ranking the prompt targets based on feature richness, where prompt targets with less information tend to have fewer variations in their features. For a given prompt from the shortlisted subset, we calculate the number of unique indices in its encoded targets z_{pt}. Formally, we maximize the function:

$$D_k(z_{pt}) = \arg\max_{z_{pt}}^{k} |z_{pt}|, \tag{3.4}$$

where $|z_{pt}|$ denotes the number of unique indices in x_{pt}'s codebook, and the x_{pt} are from the shortlisted subset. We select the top k prompts with the highest number of unique indices in their target codebooks, which ensures that the selected examples contain diverse annotations with varying meanings and structures. The final selection comprises the prompts with the most relevant queries and informative targets, which serve as our baseline prompts for the following visual reasoning step.

3.4 Experiments

In this section, we conduct evaluation on CoF prompting for LAVMs. In Section 3.4.1, we introduce our experiment settings, including dataset, pre-trained models, metrics, and other details. In Section 3.4.2, we report our main results on downstream visual tasks and present extensive quantitative and qualitative analyses. In Section 3.4.3, we conduct ablation experiments on the three major components in the CoF framework to study the contributions of each module and provide discussions. Due to page limitations, we have included additional results and analyses in the Appendix.

3.4.1 Experimental Setup

Tasks and Dataset For our experiments, we select four downstream visual tasks: image segmentation (Hong et al., 2024a), object detection (Zheng et al., 2022), image inpainting and pose estimation. Image segmentation involves partitioning an image into multiple segments or regions. The primary objective of this task is to label each pixel in the image with a class label, identifying the object to which it belongs. Pose estimation refers to the task of

determining the configuration of the body in a given image by predicting the locations of keypoints or joints. The goal here is to detect and classify the keypoints representing the positions of body parts. To facilitate these tasks, we employ the MS-COCO dataset (Lin et al., 2014), adhering to the settings outlined in Bai et al., 2024; Guo et al., 2024. Our experimental protocol involves extracting 50,000 training images and their corresponding target annotations to form the candidate prompt pool, and we rigorously test our methods on the entire validation dataset. Note that, the pre-trained LLaMA-300M and LLaMA-1B only support the image segmentation and pose estimation tasks, while LLaMA-7B supports all four downstream tasks.

Pre-trained Models We utilize pre-trained LAVMs from (Bai et al., 2024) and (Hao et al., 2024) for in-context learning. Specifically, we employ the VQ-GAN model as proposed by (Chang et al., 2023) to generate discrete visual representations of 2048 dimensions. For the autoregressive network, we leverage pre-trained LLaMA models (Touvron et al., 2023b; Touvron et al., 2023a) at different scales, including LLaMA-300M, LLaMA-1B, and LLaMA-7B for sequence modeling. Additionally, we incorporate an off-the-shelf saliency detection model from U^2-Net (Qin et al., 2020), which takes RGB images as input and outputs a saliency probability map of the same height and width as the input image.

Visual ICL Baselines We compare our method with existing visual in-context learning approaches, specifically SupPR (Zhang et al., 2023c) and SegGPT (Wang et al., 2023b). SupPR is a general prompt retrieval framework that extracts prompt pairs that contain images similar to the test input. SegGPT is a prompting method designed for segmentation tasks. We only adopt its central idea of using the same color mask for the same object class when prompting for segmentation tasks.

Post-processing and Evaluation Metrics Following (Guo et al., 2024; Zhang et al., 2023c), We utilize Intersection over Union (IoU) and Pixel accuracy (P-ACC) as our evaluation metrics for segmentation and pose estimation. We convert the predicted outputs into binary pixel masks and compare them against the binary ground truth masks. IoU measures the overlap between the predicted and ground truth regions by dividing the area of intersection by the area of union. The P-ACC calculates the proportion of correctly classified foreground

Method / Model	Image Segmenation					
	LLaMA-300M (Hao et al., 2024)		LLaMA-1B (Hao et al., 2024)		LLaMA-7B (Bai et al., 2024)	
	IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)
Random Selection	26.31 ± 0.8	42.96 ± 1.1	27.21 ± 0.4	41.88 ± 1.0	45.69 ± 1.4	59.06 ± 2.2
SegGPT (Wang et al., 2023b)	26.52 ± 1.4	42.54 ± 2.7	26.39 ± 1.2	42.71 ± 1.6	45.38 ± 0.8	60.72 ± 1.9
SupPR (Zhang et al., 2023c)	27.05 ± 1.1	43.52 ± 1.4	27.94 ± 0.9	42.16 ± 1.2	49.41 ± 1.7	65.04 ± 1.1
CoF Prompting (Ours)	**28.35**± 0.6	**46.36** ± 0.8	**28.79** ± 0.3	**44.75** ± 1.0	**52.53** ± 0.3	**67.05** ± 0.7

TABLE 3.1. Segmentation results of CoF prompting on LLaMA-300M, LLaMA-1B and LLaMA-7B.

Method / Model	Pose Estimation					
	LLaMA-300M (Hao et al., 2024)		LLaMA-1B (Hao et al., 2024)		LLaMA-7B (Bai et al., 2024)	
	IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)
Random Selection	0.60 ± 0.07	1.44 ± 0.09	1.00 ± 0.05	2.96 ± 0.10	2.40 ± 0.07	10.23 ± 0.16
SupPR (Zhang et al., 2023c)	0.67 ± 0.04	1.65 ± 0.13	1.04 ± 0.02	2.93 ± 0.18	**2.87** ± 0.22	11.29 ± 0.21
CoF Prompting (Ours)	**0.68** ± 0.04	**1.75** ± 0.05	**1.09** ± 0.02	**3.29**± 0.07	2.80 ± 0.04	**13.34** ± 0.13

TABLE 3.2. Pose Estimation Results of CoF Prompting on LLaMA-300M, LLaMA-1B and LLaMA-7B.

Model	Random	CoF
LLaMA-300M w/ VQ-GAN	58.58 ± 1.8	**57.62** ± 0.6
LLaMA-1B w/ VQ-GAN	50.08 ± 2.5	**43.12** ± 1.9
LLaMA-7B w/ VQ-GAN	45.28 ± 1.1	**42.03** ± 0.5

TABLE 3.3. Failure Rates (↓) - Image segmentation

pixels in the binary prediction mask compared to the ground truth. For detection, since the model only outputs the visualised bounding box as in image, we cannot directly obtain the coordinates for evaluation. To address this, we employ a post-process network that intakes images with visualised bounding box and outputs the box coordinates. We then calculate the IoU of the bounding boxes to the ground truth. We name the metric as Learned IoU (L-IoU). For image inpainting, we report the MSE loss and LPIPS score. We also measure the failure cases of LAVMs in making predictions, that is when the LAVMs fail to output any meaningful prediction, where the output appears in pure black. Due to page limit, we put the analysis of object detection and image inpainting in the Appendix section B1.

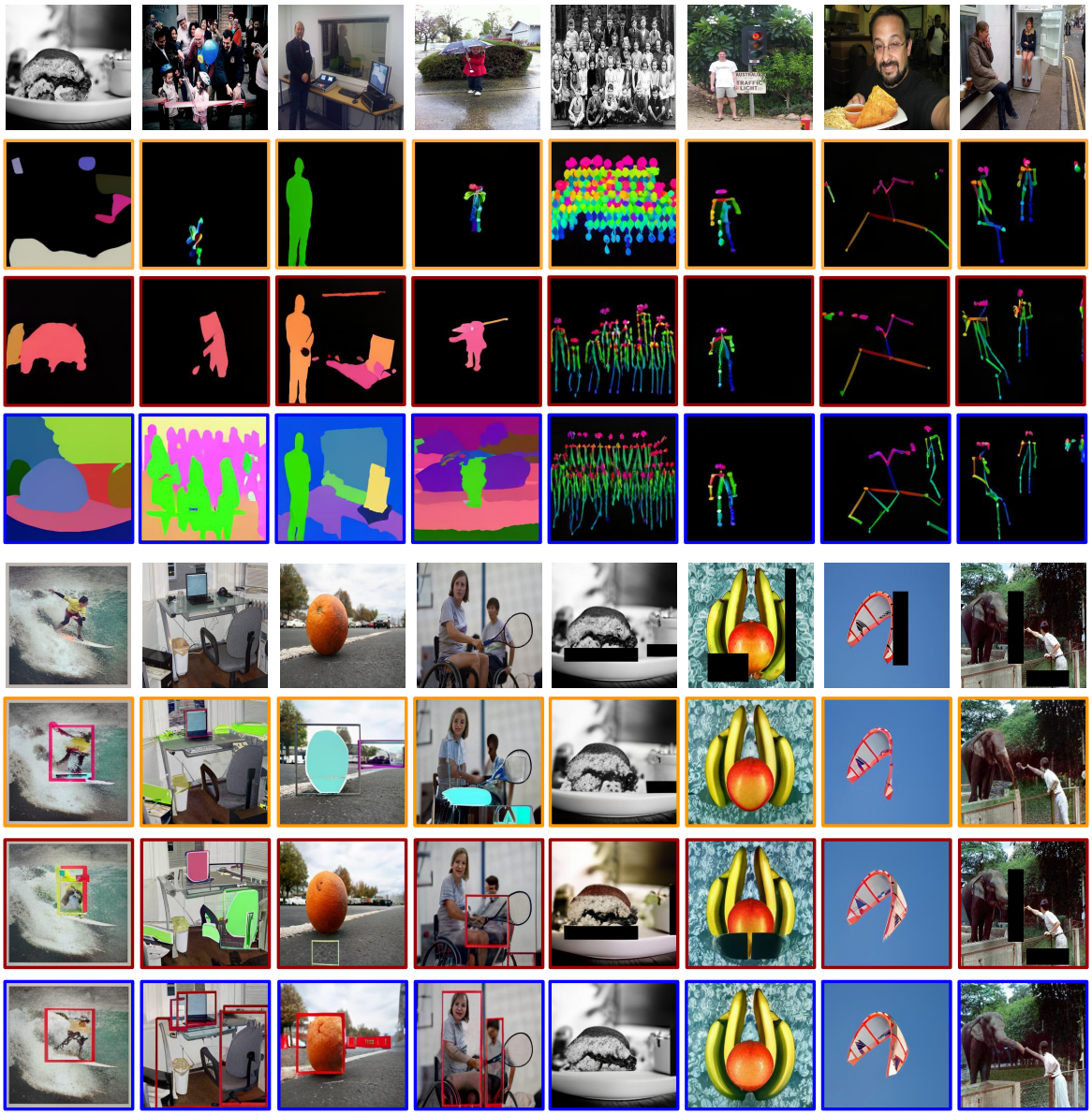

FIGURE 3.3. Results on LLaMA-7B Model. The first and fourth rows are the original test inputs for image segmentation, detection, inpainting and pose estimation, respectively. Orange boxes show the predictions given random prompts. Maroon boxes show the predictions using SupPR method (Zhang et al., 2023c). Blue boxes show the predictions using Chain-of-Focus prompting.

3.4.2 CoF Reasoning Results

Image Segmentation Table 3.1 reports the quantitative performance of CoF compared to random prompting, same colour masking (Wang et al., 2023b) and SupPR (Zhang et al.,

Model	Random	CoF
LLaMA-300M w/ VQ-GAN	53.49 ± 3.7	**52.38** ± 1.5
LLaMA-1B w/ VQ-GAN	43.67 ± 2.0	**42.54** ± 0.9
LLaMA-7B w/ VQ-GAN	41.74 ± 1.5	**35.62** ± 1.9

TABLE 3.4. Failure Rates (%↓) - Pose Estimation

2023c). The CoF method demonstrates notable percentage increases compared to the second best performing methods across various metrics. For image segmentation with LLaMA-300M, the increases are approximately 4.81% in IoU and 4.77% in P-ACC, while for LLaMA-1B and 7B, the increment in proportion is 3.04% and 6.31% in IoU, and 6.14% and 3.10% in P-ACC, respectively. The results are reported with predictions that have black rate > 0.2. We report the failure cases for segmentation in Table 3.3. Compared to the baseline, using CoF eliminates the failure cases caused by the incapability of two LAVMs by 1.64%, 13.9% and 7.7%, respectively. Figure 3.3 demonstrate the predictions made by Random, SupPR and CoF prompting with LLaMA-7B model. We observe improvement in the objects that models successfully identified and the accuracy of masking. The models using CoF prompting also demonstrate better scene understanding ability, which outputs complete masks for the same objects. These suggest that the in-context object discovery and segmentation ability of LAVMs can be enhanced by prompting them with our method.

Pose Estimation A similar trend is also found in the pose estimation task, where, as illustrated in Table 3.2, CoF prompting outperforms the other methods by a noticeable margin. For pose estimation using LLaMA-300M, compared to the second highest scores, the increases are approximately 1.49% in IoU and 6.06% in P-ACC. Moreover, LLaMA-1B shows a larger improvement, with an increase of 4.81% in IoU and 11.15% in P-ACC. For LLaMA-7B, the P-ACC is increased by 18%, but the IoU is 2.5% lower then the second highest method. The failure rates are reported in Table 3.4. Despite pose estimation being a challenging task for LAVMs, CoF prompting reduces the failure rate on both LLaMA-300M, LLaMA-1B and LLaMA-7B by 2.08% and 2.59%, 14.7% respectively. We qualitatively compare the results of the LLaMA-7B model in the top three rows in Figure 3.3. CoF prompting demonstrates better performance compared to the baselines, with improvements in the completeness of the skeletons, the accuracy of pose detection, and the number of human targets that the models

Model	CR	QR	AD	Image Segmentation		Pose Estimation	
				IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)
LLaMA-300M w/ VQ-GAN	✓			27.92	46.17	0.65	1.63
		✓		26.32	42.99	0.59	1.41
			✓	26.13	41.92	0.61	1.44
	✓	✓		28.13	45.32	0.68	1.77
		✓	✓	26.95	41.72	0.63	1.55
	✓		✓	28.21	46.10	0.65	1.71
LLaMA-1B w/ VQ-GAN	✓			28.63	45.07	1.04	2.87
		✓		26.14	43.05	0.99	2.73
			✓	27.39	43.02	1.01	2.84
	✓	✓		27.90	44.16	1.09	3.30
		✓	✓	27.33	42.19	1.07	3.01
	✓		✓	28.50	44.35	1.12	3.22
LLaMA-7B w/ VQ-GAN	✓			51.74	67.01	2.91	13.46
		✓		50.98	65.07	2.66	11.62
			✓	47.13	61.26	2.41	10.26
	✓	✓		52.01	65.83	2.88	12.89
		✓	✓	50.34	64.00	2.67	11.62
	✓		✓	51.97	66.41	2.64	12.55

TABLE 3.5. Ablation Study on the three major components involved in CoF pipeline. CR represents Cognitive Reasoning, which creates intermediate reasoning steps for the prompt target. QR represents Query relevance, which measures the similarity between the prompt queries and the test input. AD is Annotation Diversity, which involves accessing the diversity of indices within the targets' codebooks.

successfully identify in the given test input, demonstrating the effectiveness of our method. More results are provided in the appendix.

3.4.3 Ablation Studies

Intermediate Step and Prompt Selections To understand the impact of various components in our CoF prompting method, we conduct a series of ablation studies. Specifically, the designed experiments isolate and evaluate the contribution of individual components by systematically removing or modifying specific parts of the model and observing the resulting performance changes. Through this analysis, we seek to identify the critical factors that drive the success of our method and provide insights into potential areas for further improvement. We divide our entire framework into three parts: cognitive reasoning (CR), query relevance

(QR), and annotation diversity (AD). Cognitive reasoning involves generating intermediate reasoning steps using object saliency. When removing CR, we directly prompt the LAVMs using the query and its complete target. Query relevance involves selecting prompts by measuring their relevance to the test input. When removing QR, we randomly sample the candidate set of prompts. Annotation diversity involves evaluating the prompt target. When it is removed, the CoF does not access the prompt target for prompt selection.

We present the results of our ablation experiments in Table 3.5. In the image segmentation task, for all models, we observe that the primary performance improvement originates from cognitive reasoning, which incorporates intermediate steps for the prompt targets. The standalone performance of the prompt retrieval component does not significantly benefit the LAVMs, as evidenced by the predictive performance, which is comparable to the random selection baselines. However, the integration of prompt selection with cognitive reasoning shows a marked improvement, with both CR + QR and CR + AD combinations achieving better results than cognitive reasoning alone. A similar trend is observed in pose estimation, where cognitive reasoning remains the most crucial component, demonstrating a significant enhancement when applied. Notably, in pose estimation, prompt selection can also achieve good performance independently, without the aid of cognitive reasoning. This provides insight into the contribution of each component within the framework, highlighting cognitive reasoning as the most critical strategy, with the two steps involved in prompt selection seamlessly enhancing the efficacy of the reasoning strategy.

Number of Reasoning Steps Here we exam the influence of different number of reasoning steps for CoF prompting. Our setting includes using $[0, 1, 2]$ intermediate steps in between the prompt queries and the prompt target. Due to the maximum input length to the autoregressive model employed in (Hao et al., 2024), injects two intermediate steps before the final targets is the maximum for in-context learning using the model. We use the same prompt queries and original target for all three experiments to avoid influence from the prompt selection. Figure 3.4 demonstrates our results, where both models show improved performance with an increasing number of reasoning steps, indicating that more reasoning steps enhance their capabilities. However, in the case of the 300M model, the scores decrease when increasing

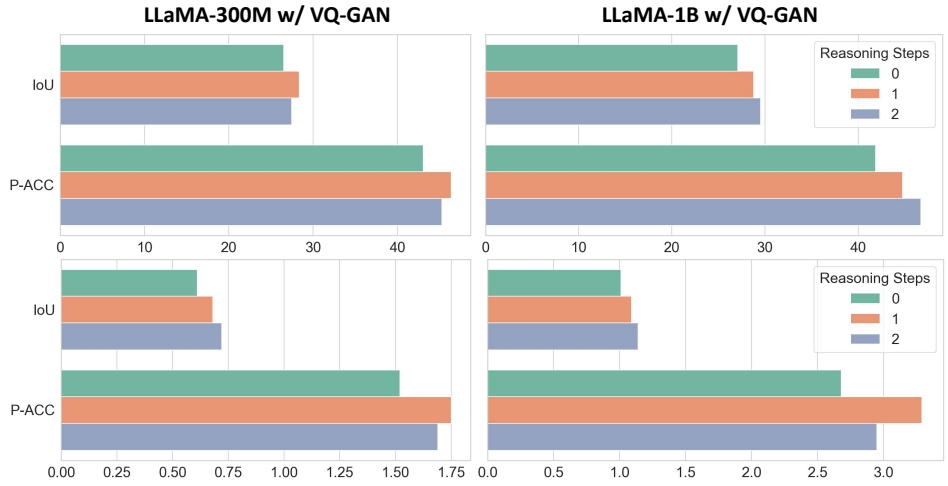

FIGURE 3.4. Comparison of using different reasoning steps. The first row of figures captures the performance measures of the image segmentation task, and the second row captures the performance measures of the pose estimation task.

from one intermediate step to two intermediate steps in image segmentation. Conversely, the LLaMA-1B model exhibits a more stable linear increment compared to LLaMA-300M, demonstrating that the larger model benefits more significantly from reasoning steps. These results highlight the importance of CoF prompting in achieving better performance.

3.5 Conclusion

The paper introduces Chain-of-Focus (CoF) prompting, a novel method designed to replicate the sequential steps of Chain-of-Thought prompting in the visual domain by bridging the gap between symbolic reasoning in language models and perceptual reasoning in vision models. CoF automates prompt design by selecting the most relevant and informative prompts from existing candidates and addresses the inherent challenge of the lack of explicit symbolic structure in images by utilizing visual saliency to create intermediate reasoning steps for prompt targets, capturing the intrinsic logic of the human perceptual system. By leveraging this hierarchical information, COF allows Large Autoregressive Vision Models (LAVMs) to process and understand visual information progressively, thus enhancing their sequential predictive performance on various downstream vision tasks. Our experiments on image

segmentation and pose estimation using LLaMA-300M, 1B and 7B w/ VQ-GAN models demonstrate that embedding visual reasoning into prompts significantly improves the model's inference capabilities. CoF prompting represents a significant advancement in visual in-context learning, with potential for broader applications in machine learning and computer vision.

CHAPTER 4

Interpretable Cross-Modal Latent Modeling for Multimodal Generation

---

This chapter extends the thesis from single-modality representation control and reasoning to interpretable latent modeling across modalities, focusing on text-to-audio-video generation. Building on earlier chapters that establish selective invariance and structured latent organization as key principles, this chapter addresses a fundamental limitation of existing multimodal generative models: the assumption of full correspondence between audio and visual representations. It introduces Selective Audio-Visual Alignment (SAVA), a framework that explicitly identifies and aligns only the latent components shared across modalities while filtering out modality-specific factors that introduce semantic or temporal noise. By grounding cross-modal alignment in a causal latent structure and enforcing selective, interpretable alignment, the chapter demonstrates how controllable latent modeling enables robust and synchronized multimodal generation, completing the thesis narrative from task-specific learning to large-scale multimodal synthesis.

4.1 Introduction

Recent advances in multimodal generative models (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023b; Ruan et al., 2023; Sun et al., 2024; Wu et al., 2024b; Team et al., 2024) have enabled high-quality content creation across text, image, audio, and video modalities. While notable progress has been made in text-to-video (Blattmann et al., 2023; Hong et al., 2023; Khachatryan et al., 2023; Hu, 2024; Singer et al., 2023) and text-to-audio (Ghosal et al., 2023; Liu et al., 2023a; Liu et al., 2024a; Majumder et al., 2024) generation individually, they are typically studied in isolation, leaving joint audiovisual generation from text largely

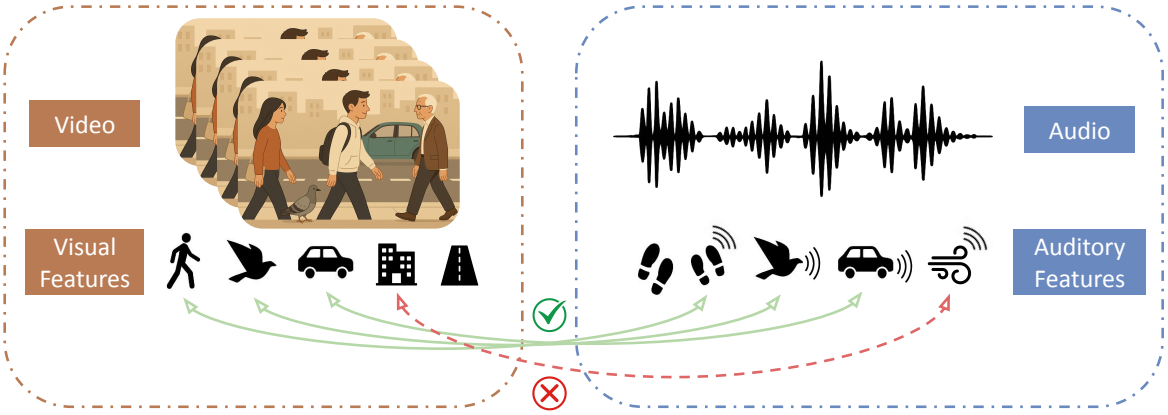

FIGURE 4.1. Visual-auditory feature alignment is essential in text-to-audio-video (T2AV) generation, yet assuming full correspondence between audio and visual modalities is often problematic. For example, visual elements like roads or buildings may not produce sound, while audio events such as wind may lack visual presence. Aligning such mismatched features introduces semantic noise, resulting in reduced cross-modal consistency and temporal mismatch in the generated outputs.

underexplored. Text-to-Audio-Video (T2AV) generation addresses this gap by aiming to synthesize audio and video streams that are both semantically and temporally aligned, conditioned on a single text prompt. This involves not only generating high-quality content for each modality, but also ensuring that the output audio and video remain contextually consistent and synchronized.

Achieving this requires modeling cross-modal alignment, where both audio and visual representations capture the informative content conveyed by the other modality. To facilitate such alignment, existing approaches often project multimodal features into a shared embedding space (Mao et al., 2024; Tang et al., 2023; Xing et al., 2024). This facilitates the model to capture joint semantics across modalities. However, forcing all audio and visual features to align can be problematic. In real-world settings, audio and visual streams may exhibit only partial alignment: audio may describe only parts of a visual scene, or visual frames may contain elements absent from the audio (See Figure 4.1). Enforcing full alignment under such conditions introduces mismatched information into the joint representations, resulting in semantically inconsistent or temporally desynchronized outputs during T2AV generation.

To address the challenge of partial correspondence between modalities, we introduce *SAVA*, a framework for **S**elective **A**udio-**V**isual **A**lignment in text-to-audio-video generation. Comparing with existing approaches (Luo et al., 2023; Mao et al., 2024; Tang et al., 2023; Wang et al., 2024b; Xing et al., 2024) that assume full alignment between audio and visual features, *SAVA identifies and aligns only those latent components that are jointly predictive across modalities*, while disregarding modality-specific information that could otherwise introduce noise or conflict. The overall pipeline, as shown in Figure 4.3, proceeds in three stages: *Align and Fine-tune*, it learns to map multimodal latents by selectively filtering out irrelevant dimensions in the latent space using a learnable mask, allowing the model to focus only on features that contribute meaningfully to both modalities. The alignment is learned through adapter networks applied to pretrained encoders. Then, we fine-tune the generator using the aligned multi-condition inputs. *Inference*, it operates in a cascaded manner, projecting features from video and audio into an aligned subspace, and conditioning the corresponding generator on both the text and the aligned video/audio signals. This design enables synchronized and consistent audio-visual generation while preserving efficiency and modularity.

SAVA is grounded in a causal view of multimodal generation, where audio and visual signals are generated from a mixture of shared and distinct latent factors. We provably show that the masked alignment objective recovers the minimal set of shared latent variables (those which constitute the true semantic interface between modalities). This not only ensures interpretability and robustness but also mitigates the entanglement issues observed in prior alignment-based models. Our empirical results across diverse benchmarks confirm that *SAVA* significantly improves semantic alignment and temporal synchronization in T2AV generation, outperforming existing baselines.

4.2 Related Works

Text-to-Audio-Video Generation Text-to-Audio-Video (T2AV) generation aims to synthesize audio and video streams that are semantically and temporally aligned, conditioned on a single text prompt. The task extends beyond text-to-video (T2V) and text-to-audio (T2A) generation

by requiring consistency across modalities. Recent advances in T2V (Chen et al., 2023b; Guo et al., 2023; Khachatryan et al., 2023; Wu et al., 2023a; Zhang et al., 2023b) and T2A (Agostinelli et al., 2023; Huang et al., 2023a; Liu et al., 2024a; Majumder et al., 2024; Tan et al., 2024) have enabled high-quality content generation in each modality. However, generating them independently often results in misaligned outputs, as the modalities are not conditioned on each other. A simple alternative is a cascaded approach, where one modality (e.g., video) is generated first and used to condition the other (e.g., audio). While this improves synchronization, it may propagate errors and lead to inconsistencies with the original text. To address these issues, recent T2AV methods propose joint modeling strategies. CoDi (Tang et al., 2023) unifies generation across multiple modalities in a single diffusion framework via aligning prompt encoders (text, image, video, audio) into a shared input space using contrastive learning, with text as the central bridging modality. (Xing et al., 2024) aligns pretrained T2V and T2A models via a shared semantic space using ImageBind. TAVDiffusion (Mao et al., 2024) adopts a two-stream latent diffusion model and addresses alignment via cross-attention and contrastive learning. Nevertheless, joint modeling of audio and visual modalities requires careful alignment of representations to preserve both semantic consistency and temporal synchronization. Our framework complements existing approaches by introducing a targeted alignment mechanism that mitigates the impact of noisy or partial correspondences, leading to more faithful and consistent T2AV generation.

Cross-Modal Alignment Cross-modal alignment is crucial for integrating information from different modalities, facilitating tasks such as retrieval, classification (Zheng et al., 2022; Hong et al., 2024a), and generation (Huang et al., 2023b; Zheng et al., 2024; Zheng et al., 2025). The goal is to project modality-specific features into a shared embedding space where semantically related inputs are closely aligned. In the vision-language domain (Huang et al., 2024b; Huang et al., 2025a; Huang et al., 2025b), CLIP (Radford et al., 2021) has become a standard framework, while CLAP (Wu et al., 2023b) and CAVP (Luo et al., 2023) extend contrastive alignment to audio-language and vision-audio pairs, respectively. ImageBind (Girdhar et al., 2023) further generalizes this approach to unify multiple modalities in a single embedding space. Such alignment modules are integral to conditional generative models (Luo et al., 2023). While early approaches trained modality encoders from scratch,

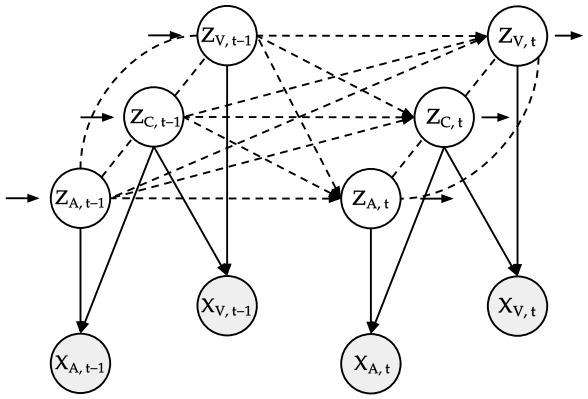

FIGURE 4.2. The data generative process of audiovisual data. Audio features Z_A are selectively derived from visual features Z_v guided by a learned mask m. Each modality-specific latent combines with residual noise ϵ to produce the outputs.

recent work shows that frozen foundation models can be effectively adapted using lightweight projectors (Houlsby et al., 2019; Mokady et al., 2021). In video-to-audio generation, V2A-Mapper (Wang et al., 2024b) learns a projection from CLIP to CLAP features using a simple MLP, enabling audio generation conditioned on vision without retraining large-scale models. Despite these advances, aligning the correct semantic content across modalities remains challenging. Representations often entangle modality-specific and irrelevant information, leading to noisy alignment. SmartCLIP (Xie et al., 2025) identifies this issue in vision-language models, showing that CLIP embeddings often entangle unrelated concepts due to coarse-grained alignment. These findings underscore a broader challenge in multimodal generation: how to align information across modalities such that the learned representations do not introduce inconsistencies in the generated outputs. Our work addresses this by introducing a masked adapter module that enables efficient and selective alignment between pretrained modality-specific encoders. By focusing alignment on semantically relevant regions, our method mitigates noisy correspondence and improves consistency in T2AV generations.

4.3 Problem Formulation

Text-to-Audio-Video (T2AV) generation requires accurate alignment between audio-visual representations to preserve meaningful cross-modal correspondence. In particular, semantic

misalignment, where visual and auditory components do not reflect the same underlying content, can mislead generative models and degrade the consistency of the resulting outputs. Our objective is to enable selective and reliable alignment by identifying and preserving only the semantically relevant components across modalities during training. To this end, we begin by reviewing the text-to-audio-video generative process.

Data Generative Process As illustrated in Figure 4.2, we model the audiovisual data generation process using a structured causal model composed of three latent variable sets: video-specific latent variables Z_V, audio-specific latent variables Z_A, and cross-modal latent variables Z_C, which encode shared content factors underlying both modalities. Z_C may include semantically grounded, temporally evolving entities that manifest in both the visual and auditory domains (e.g., a barking dog or a moving vehicle). In contrast, the modality-specific latents Z_V and Z_A capture factors that are unique to the video and audio domains, respectively. The latent variables are causally connected and evolve over time, with each group at time step t potentially influenced by their own past states and the past states of other groups. Formally, the evolution of these latent variables follows:

$$Z_V^t \leftarrow \{Z_V^{\text{t-1}}, Z_A^{\text{t-1}}, Z_C^{\text{t-1}}\},\ Z_A^t \leftarrow \{Z_V^{\text{t-1}}, Z_A^{\text{t-1}}, Z_C^{\text{t-1}}\},\ Z_C^t \leftarrow \{Z_V^{\text{t-1}}, Z_A^{\text{t-1}}, Z_C^{\text{t-1}}\}, \tag{4.1}$$

where the superscript 'past' denotes historical latent states (e.g., from time $t-1$), and the arrows represent causal influence. These relationships reflect the potential bidirectional statistical and causal dependencies (Von Kügelgen et al., 2021) across modalities.

The observable variables: video X_V and audio X_A, are generated from their corresponding modality-specific latent variables in conjunction with the cross-modal latent:

$$X_V \leftarrow \{Z_V, Z_C\},\ X_A \leftarrow \{Z_A, Z_C\}. \tag{4.2}$$

This formulation reflects that while Z_C captures semantically aligned and temporally correlated content, Z_V and Z_A may contain orthogonal information that should not be forced into alignment. Therefore, when attempting to recover cross-modal structure, it is critical to distinguish shared factors from modality-specific ones.

Problem Setup Let $X_V \in \mathcal{X}_V$ and $X_A \in \mathcal{X}_A$ denote observed video and audio inputs, generated from modality-specific latent variables $Z_V, Z_A \in \mathbb{R}^d$ and shared cross-modal latent variables $Z_C \in \mathbb{R}^d$. We use pretrained encoders to extract latent representations $\hat{Z}_V, \hat{Z}_A \in \mathbb{R}^d$, serving as proxies for Z_V and Z_A. To allow flexible transformation, we apply learnable reparameterizations $q_V, q_A : \mathbb{R}^d \to \mathbb{R}^d$, yielding $\tilde{Z}_V = q_V(\hat{Z}_V)$ and $\tilde{Z}_A = q_A(\hat{Z}_A)$. These transformed embeddings are then used to identify shared cross-modal structure. We learn binary mask functions $M_V, M_A : \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}^d$, which output masks $M_V(\tilde{Z}_V, \tilde{Z}_A)$ and $M_A(\tilde{Z}_V, \tilde{Z}_A)$, indicating dimensions in each modality that align with shared semantics or temporal structure.

Let $S_V \subseteq [d]$ and $S_A \subseteq [d]$ denote the support of these masks, i.e., the selected dimensions where the mask equals 1. The masked representations are denoted $\tilde{Z}_V^{S_V}$ and $\tilde{Z}_A^{S_A}$, and we write $\overline{S} = [d] \setminus S$ for the complement. Let $S_V^\dagger$ and $S_A^\dagger$ represent the (unknown) ground-truth index sets corresponding to the shared latent dimensions (i.e., a Markov blanket of X_A in Z_V, and vice versa, under reparameterization).

4.4 Masked Latent Adaptation and Cascaded Diffusion Generation

Selective Latent Alignment Let $x_v \in X_V,\ X_V \subseteq \mathbb{R}^{T \times H \times W \times 3}$ be a video clip and $x_a \in X_A,\ X_A \subseteq \mathbb{R}^{T' \times M}$ its corresponding audio spectrogram. We extract frozen modality-specific embeddings $\hat{z}_v = f_v(x_v), \hat{z}_a = f_a(x_a),$, where f_v is a pretrained video VAE encoder (Hong et al., 2023) and f_a is a pretrained audio diffusion encoder (Wu et al., 2023b; Liu et al., 2023a). These raw embeddings may contain modality-specific noise and are not guaranteed to lie in a common semantic subspace. To expose the shared latent structure $Z_C \subseteq \mathbb{R}^d$, we apply learnable reparameterizations (i.e., adapter networks):

$$\tilde{z}_v = q_V(\hat{z}_v), \qquad \tilde{z}_a = q_A(\hat{z}_a). \tag{4.3}$$

This projection step adapts the output of each frozen encoder and is subsequently trained to isolate cross-modal features. We then introduce two mask networks:

$$M_V, M_A : \mathbb{R}^{2d} \to [0,1]^d, \tag{4.4}$$

each taking both $\tilde{z}_v$ and $\tilde{z}_a$ as input. Due to the semantic ambiguity and contextual diversity in audio-visual alignment, the relevant latent dimensions within the visual representation can vary depending on the specific context. For example, a single video clip may be paired with different types of audio, such as background music or voiceover narration, each requiring attention to distinct visual regions or semantic features. Accordingly, the masking function should be conditioned on both video and audio inputs. Conditioning on only one modality impairs the model's ability to disambiguate cross-modal variations, leading to suboptimal or unstable mask learning.

Cross-Modal Reconstruction Let $S_V \subseteq [d]$ and $S_A \subseteq [d]$ represent the indices of dimensions selected by the soft masks. After applying a thresholding operation, we obtain binary supports $\tilde{S}_V \subseteq [d]$ and $\tilde{S}_A \subseteq [d]$, which indicate the dimensions where the mask value equals 1 (i.e., $\tilde{S}_V = \{i \in [d] \mid M_V(\tilde{z}_v, \tilde{z}_a)_i = 1\}, \tilde{S}_A = \{i \in [d] \mid M_A(\tilde{z}_v, \tilde{z}_a)_i = 1\}$). These binary masks are then used to construct the masked latent representations by retaining only the selected dimensions:

$$\tilde{z}_v = M_V(\tilde{z}_v, \tilde{z}_a) \odot \tilde{z}_v, \qquad \tilde{z}_a = M_A(\tilde{z}_v, \tilde{z}_a) \odot \tilde{z}_a, \tag{4.5}$$

and decode each using the corresponding latent diffusion decoders g_A and g_V:

$$\hat{x}_a = g_A(\tilde{z}_v), \qquad \hat{x}_v = g_V(\tilde{z}_a). \tag{4.6}$$

To ensure that the masked latent remains informative, we reconstruct each modality from the masked latent of the other. This reconstruction objective acts as a constraint that prevents degenerate masking solutions. Without it, the sparsity loss alone would encourage all-zero masks, as they trivially minimize the L1 penalty while discarding all information (Kong et al., 2022; Xie et al., 2023). We reconstruct across modalities (i.e., audio from masked video latent and video from masked audio latent) rather than within the same modality. This cross-modal reconstruction forces the mask to preserve only the latent dimensions that are predictive of

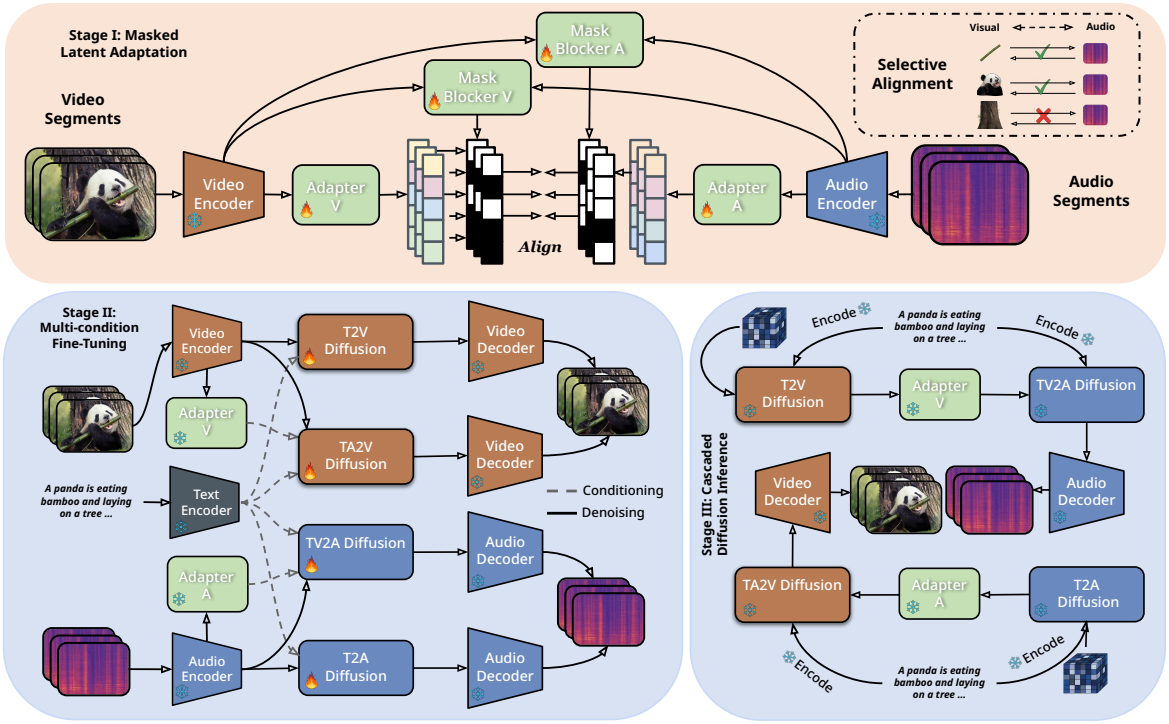

FIGURE 4.3. Overview of our proposed T2AV framework. The system involves training a learnable mask that selectively aligns the latents of each modality, filtering out irrelevant visual content (e.g., tree trunks) while preserving meaningful cues (e.g., bamboo being eaten). The aligned representations are then used to fine-tune the generator, adapting the multimodal conditions alongside the text condition, followed by generation through a latent diffusion model.

the other modality, thereby isolating the shared semantic structure. As a result, the model learns compact and meaningful latent supports that are truly cross-modally informative.

4.4.1 SAVA-Diffusion

In this section, we present the implementation of the proposed SAVA-Diffusion framework for Text-to-Audio-Video (T2AV) generation (See Figure 4.3). The framework consists of three stages: (1) a masked latent adaptation stage in which we train aligned projections of video-audio latents via selective masking, and (2) a fine-tuning stage that adapts the aligned latents as a joint condition alongside the text condition. (3) a cascaded diffusion generation

stage, where high-fidelity audio and video outputs are synthesized using latent diffusion models in sequential orders.

Stage I: Masked Latent Adaptation In first stage, our method implements selective cross-modal alignment by learning to isolate the latent dimensions that are predictive of the other modality. We first obtain reparameterized embeddings $\tilde{z}_v$ and $\tilde{z}_a$ from the frozen encoders (Yang et al., 2024; Liu et al., 2023a) and adapters. These are passed to the modality-specific mask functions, each conditioned on both modalities, to produce binary masks that filter the latent features. The masked latents are then decoded to reconstruct the opposite modality. To encourage a consistent embedding geometry between modalities (Xie et al., 2025), we further include a direct alignment loss between $\tilde{z}_v$ and $\tilde{z}_a$ prior to masking. This stabilizes training and promotes representational coherence across modalities. The total objective is:

$$\mathcal{L}_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}\left[\ell_A(x_a^{(i)}, g_A(M_V \odot \tilde{z}_v^{(i)})) + \ell_V(x_v^{(i)}, g_V(M_A \odot \tilde{z}_a^{(i)})) + \alpha \cdot \mathcal{L}_{\text{align}}(\tilde{z}_v^{(i)}, \tilde{z}_a^{(i)})\right] + \lambda\left(\|M_V\|_1 + \|M_A\|_1\right), \tag{4.7}$$

where ℓ_A and ℓ_V are cross-modal reconstruction losses, and the ℓ_1 regularization encourages sparsity in the learned supports. $\mathcal{L}_{\text{align}}$ measures the distance (e.g., normalized ℓ_2) between the unmasked latent representations, and α controls the alignment strength.

Stage II: Multi-condition Fine-tuning In Stage II, the TV2A and TA2V diffusion backbones are fine-tuned separately, each to use the cross-modal latent learned in Stage I, while keeping the decoders g_V, g_A frozen. For audio, given the text embedding $z_t = f_t(x_t)$ and the aligned visual latent $\tilde{z}_v = q_V(z_v^T)$, we form $c_A = [\, z_t;\ \phi_A(\tilde{z}_v)\,]$ and adapt the conditioning pathway via LoRA (Hu et al., 2022), similarly, for video we form $c_V = [\, z_t;\ \phi_V(\tilde{z}_a)\,]$ from the aligned audio latent $\tilde{z}_a = q_A(z_a^T)$, the objective is the diffusion loss:

$$\mathcal{L}_{\text{FT-A}} = \mathbb{E}_{x_a,\epsilon,t}\left\|\epsilon - \epsilon_{\theta_A}\big(x_a^{(t)}, t, c_A\big)\right\|_2^2 \qquad \mathcal{L}_{\text{FT-V}} = \mathbb{E}_{x_v,\epsilon,t}\left\|\epsilon - \epsilon_{\theta_V}\big(x_v^{(t)}, t, c_V\big)\right\|_2^2. \tag{4.8}$$

with LoRA parameterization $W' = W + BA$ on selected cross-attention or FiLM layers. The total objective is not coupled during optimization, instead, we run distinct trainings.

LoRA only on conditioning layers (mid-block and a few down/up blocks), and all backbone convolutions and decoders frozen to preserve pretrained priors while teaching each model, in isolation, to respond to its new cross-modal condition. In addition, we fine-tune the individual T2A and T2V models to further enhance generation quality across modalities for the inference pipeline.

Stage III: Cascaded Diffusion Inference Building on the aligned latent representations from stage I and fine-tuned diffusion models from stage II, we generate video and audio in a cascaded manner using independently finetuned single-modal diffusion models (T2A, T2V) (Yang et al., 2024; Liu et al., 2023a) and multi-model diffusion models (TA2V, TV2A). As illustrated in Figure 4.3, the process begins by generating a video from a text prompt using a T2V diffusion model. The resulting visual latent z_v^T is then adapted through a lightweight projection network $\mathcal{P}_\theta$, producing an audio-guiding latent $\tilde{z}_a$ that encodes visually grounded cues. This latent conditions the subsequent audio generation, serves as a supervision signal to finetune the audio diffusion model for improved semantic coherence. Similarly, we can generate video conditioned on both audio and text latents. By structuring the process in this cascaded fashion, we ensure that the audio is aligned with the generated visual/audible content. The detailed formulation of the reverse diffusion process for both modalities is provided in Appendix C. Notably, the pretrained diffusion encoders remain frozen during Stage I, and only the adapters are updated; fine-tuning of the diffusion models is performed in Stage II to enhance generation quality while maintaining modularity and efficiency.

4.5 Theoretical Analysis

In this section we show that our masked cross-modal reconstruction with an ℓ_1-penalty provably recovers exactly the shared latent factors between video and audio, i.e. the minimal Markov blankets on the pretrained features, even when those features are entangled. By faithfulness and d-separation on the latent DAG (Peters et al., 2017) over (Z_V, Z_A, X_V, X_A), there exist unique index sets $S_V^\dagger \subseteq [d],\ S_A^\dagger \subseteq [d]$, which are the minimal Markov blankets of X_A in Z_V and of X_V in Z_A, respectively. Equivalently, $S_V^\dagger$ is the smallest subset satisfying

that conditioning on $\{Z_{V,i} : i \in S_V^\dagger\}$ renders all other latent coordinates irrelevant to X_A. The analogous property holds for $S_A^\dagger$.

To make precise what it means for two sets of latent factors to capture all and only the shared information, we introduce the following definition.

DEFINITION 1 (Minimum Sufficient Latents). *Given index sets $\tilde{S}_V, \tilde{S}_A \subseteq [d]$, we say that the pairs $\left(\tilde{Z}_V^{\tilde{S}_V}, \tilde{Z}_A^{\tilde{S}_A}\right)$ are* Minimum Sufficient Latents *if they satisfy*

$$I\left(\tilde{Z}_V^{\tilde{S}_V}; X_A\right) = I\left(Z_V^{S_V^\dagger}; X_A\right),\ I\left(\tilde{Z}_{V_j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V}\right) = 0\ \ \forall j \notin \tilde{S}_V,$$
$$I\left(\tilde{Z}_A^{\tilde{S}_A}; X_V\right) = I\left(Z_A^{S_A^\dagger}; X_V\right),\ I\left(\tilde{Z}_{A_j}; X_V \mid \tilde{Z}_A^{\tilde{S}_A}\right) = 0\ \ \forall j \notin \tilde{S}_A.$$

Key Assumptions We require four conditions (see Appendix B for formal definitions):

(1) **(DAG & d-Separation)** There is a latent DAG over (Z_V, Z_A, X_V, X_A) whose minimal Markov blankets $S_V^\dagger, S_A^\dagger$ correspond to the truly shared factors.
(2) **(Block-wise Reparameterization)** The class of invertible maps q_V, q_A is rich enough that there exists a reparameterization under which the shared block $S_V^\dagger$ (resp. $S_A^\dagger$) becomes an axis-aligned subset $\tilde{S}_V^\dagger$ (resp. $\tilde{S}_A^\dagger$) of the coordinates.
(3) **(Decoder Universality)** The decoder families Q_{g_A}, Q_{g_V} can approximate any conditional distribution, so that minimizing cross-entropy is equivalent to minimizing true conditional entropy.
(4) **(Mask Universality & Penalty-Range)** The masks can implement any support selection per example, and the sparsity weight λ lies strictly between the smallest shared-factor contribution and the largest non-shared contribution (see Assumption 4).

Above assumptions are commonly used. First, a latent-variable DAG with faithfulness (Assumption 1) underlies most generative models in vision and audio, and the Markov blanket then exactly characterizes the shared information. This is the fundamental assumption in Causality (Peters et al., 2017). Second, block-wise reparameterization (Assumption 2) merely requires that our invertible networks q_V, q_A have sufficient capacity to "whiten" or disentangle the small block of truly shared latents; in practice modern normalizing-flow and

invertible-residual architectures easily satisfy this. Third, decoder universality (Assumption 3) is standard in representation learning deep decoders with enough width and nonlinearity can approximate any conditional density arbitrarily well, so cross-entropy minimization recovers true conditional entropy. Mask universality implies stipulate that our mask networks are expressive enough to pick any subset of coordinates per example. All these universality have been supported by universal approximation theory of deep learning methods (Huang et al., 2024a). Finally, penalty-range requirement (Assumptions 4) implies that the sparsity weight λ can be chosen (e.g. via cross-validation) to lie between the minimal utility of a shared factor and the maximal spurious contribution of a non-shared factor. In practice, we can just make λ be sufficiently small. Together, these common assumptions ensure our theoretical guarantees apply to many practical architectures.

LEMMA 1 (Sufficientness of Reconstruction). *Fix any invertible q_V. Under Assumptions 1-4, any mask–decoder pair (M_V, g_A) that minimizes $\mathbb{E}[-\log Q_{g_A}(X_A \mid M_V \odot \tilde{Z}_V)]$ must satisfy, for every example, $I\big(\tilde{Z}_V^{S_V(\tilde{Z}_V,\tilde{Z}_A)}; X_A\big) = I\big(\tilde{Z}_V; X_A\big)$. In other words, the selected coordinates form a sufficient statistic for X_A.*

This lemma shows that if we only optimize the reconstruction loss (cross-entropy) then the learned mask necessarily keeps *all* the information in $\tilde{Z}_V$ that is relevant to predicting X_A. In other words, the selected subset of coordinates forms a sufficient statistic for the audio modality, capturing every bit of shared information from the video embedding.

LEMMA 2 (Sparsity-Induced Minimality). *Fix any invertible q_V. Under Assumptions 3-4, the joint minimizer $(M_V^*, g_A^*) = \arg\min_{M_V, g_A} \Big\{\mathbb{E}\big[-\log Q_{g_A}(X_A \mid M_V \odot \tilde{Z}_V)\big] + \lambda\, \mathbb{E}[\|M_V\|_1]\Big\}$ satisfies, for every example,*

$$S_V^*(\tilde{Z}_V, \tilde{Z}_A) = \tilde{S}_V^\dagger,\ I\big(\tilde{Z}_{V,j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V^\dagger}\big) = 0\ \forall j \notin \tilde{S}_V^\dagger. \tag{4.9}$$

That is, the mask prunes away non-shared coordinates, recovering exactly the minimal shared block.

This lemma establishes that once we add a sparsity penalty on the mask, the model discards every coordinate that does not uniquely contribute to cross-modal reconstruction. The result

Method	VGGSound+				AudioCaps			
	FVD ↓	**FAD** ↓	**AVHScore** ↑	**CAVPSIM** ↑	**FVD** ↓	**FAD** ↓	**AVHScore** ↑	**CAVPSIM** ↑
Two-Streams	768.5	6.29	0.058	0.104	961.4	7.36	0.041	0.165
CasC-Diff	768.5	7.53	0.144	0.126	961.4	9.51	0.092	0.192
TAVDiff (Mao et al., 2024)	956.3	8.94	0.162	0.098	1131.9	8.43	0.105	0.182
CoDi (Tang et al., 2023)	709.4	8.36	0.108	0.149	902.5	9.07	0.098	0.211
JavisDiT (Liu et al., 2025)	697.4	6.17	0.153	0.140	**801.2**	7.55	0.104	0.207
Unidirectional	**662.9**	**5.49**	0.206	0.165	**817.6**	**7.32**	0.142	0.230
Bidirectional	701.4	-	**0.217**	**0.183**	852.4	-	**0.157**	**0.242**

TABLE 4.1. Quantitative comparison. Our method outperforms existing baselines in both generative quality metrics and alignment metrics, demonstrating improvements in fidelity as well as cross-modal consistency. For the unidirectional setting, we directly adopt the fine-tuned T2V model for video generation. The generated audio for both the bidirectional and unidirectional settings is identical.

is the *minimal* subset of features which are precisely the shared latent block. Therefore no redundant or modality-specific information remains.

THEOREM 1 (Global Block-Alignment and Recovery). *Under Assumptions 1, 2, 3, and 4 the global minimizer of Objective 4.7 yields* $\left(\tilde{Z}_V^{\tilde{S}_V^*}, \tilde{Z}_A^{\tilde{S}_A^*}\right)$ *that satisfies Definition 2.*

This theorem combines the sufficiency and minimality results in both directions (i.e., video to audio and audio to video) and shows that our simple mask-and-reconstruct framework provably extracts exactly the shared latent variables and eliminates all modality-specific components.

4.6 Experiments

4.6.1 Experiment Setup

Dataset We conduct experiments on two benchmark datasets: VGGSound (Chen et al., 2020a) and AudioCaps (Kim et al., 2019). VGGSound comprises approximately 200K 10-second video clips spanning 310 sound classes, with strong audio-visual correspondence ensured by the presence of visible sound sources. Following the protocol in (Xing et al., 2024), we sample 5k and 3K clips from the train and test split, respectively, and annotate them with

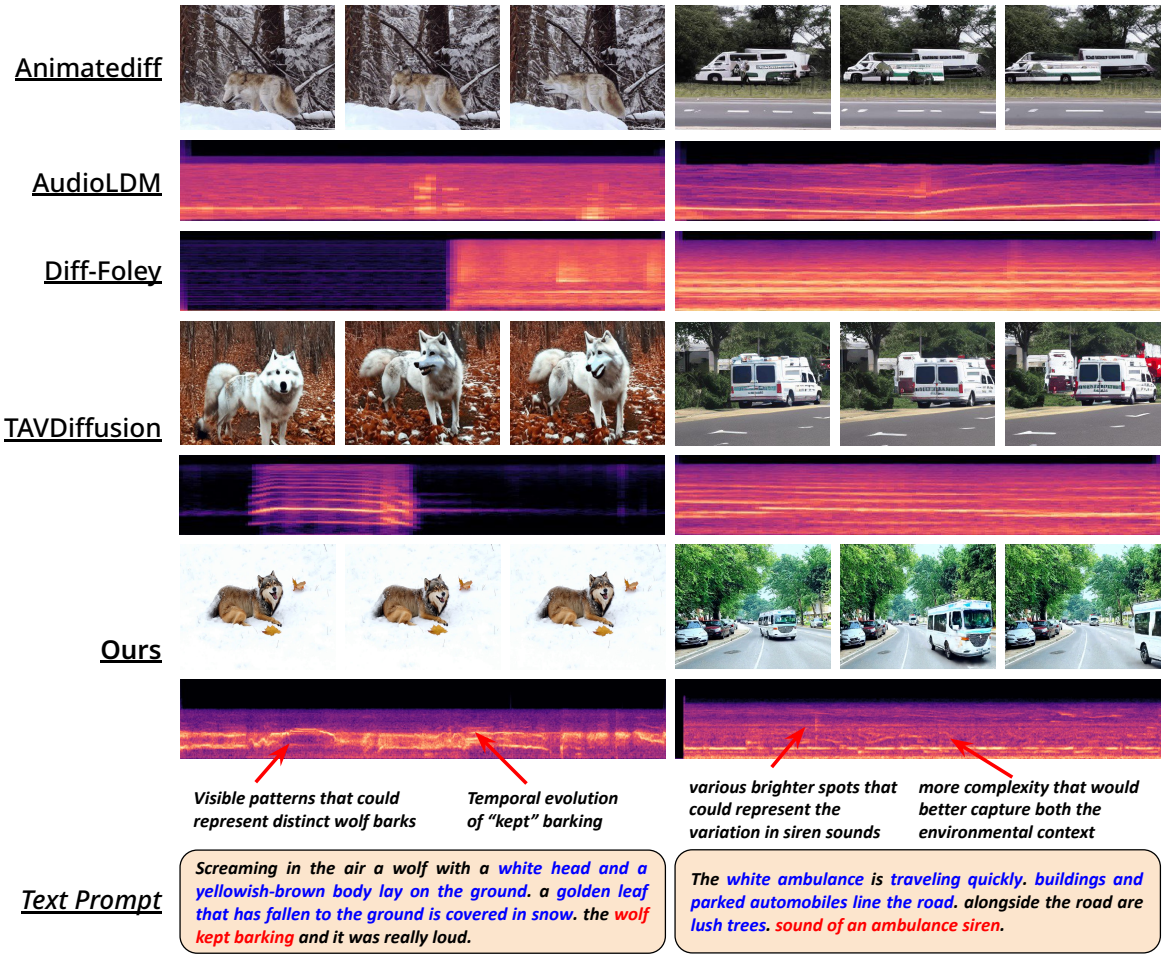

FIGURE 4.4. Text-to-Audio-Video generation results. We use the same text prompt as in (Mao et al., 2024) for our demonstration and compare our method against multiple baselines (Animatediff (Guo et al., 2023), AudioLDM (Liu et al., 2023a), Diff-Foley (Luo et al., 2023), and TAVDiffusion (Mao et al., 2024)). Compared to prior methods, our approach (unidirectional setting as illustrated) produces higher quality and aligned video and audio content.

text prompts using VideoBlip (Yu et al., 2024), as adopted in (Mao et al., 2024). AudioCaps consists of 46K audio clips paired with human-written captions sourced from AudioSet, and serves as a standard benchmark for audio-language grounding. We also sample 5K paired clips from the training split. To facilitate alignment learning and fine-tuning, we merge the training sets of both datasets, and perform evaluation separately on each test set.

Implementation Details To adapt the diffusion models to the target data domains, we first fine-tune the video and audio diffusion models independently using the training set, respectively.

For video generation, we employ the pretrained CogVideoX1.5 (Yang et al., 2024), and extract latent representations using its VAE encoder. For audio, we adopt AudioLDM (Liu et al., 2023a), which integrates a pretrained CLAP encoder (Wu et al., 2023b) for audio feature extraction. The latent dimensionality of aligned embeddings for audio generation is fixed at 512. Each generated sample has a duration of 10 seconds, with video rendered at 16 frames per second and audio sampled at 48 kHz. Our adapter and masking modules are implemented as multilayer perceptrons. For the masking mechanism, we evaluate both soft masks (sigmoid outputs as weights) and hard masks, obtained by thresholding at 0.5. The loss weights λ_1 and λ_2 are empirically set to 5 and 0.1, respectively.

Evaluation Metrics We assess perceptual quality of the generated video and audio using Fréchet Video Distance (FVD) (Unterthiner et al., 2019) and Fréchet Audio Distance (FAD) (Ruan et al., 2023), respectively. Cross-modal semantic alignment is measured by AVHScore (Mao et al., 2024), while CAVP similarity (Luo et al., 2023) evaluates temporal synchronization. For V2A performance, we adopt the evaluation protocol from (Xing et al., 2024), including KL divergence, Inception Score (ISc), Fréchet Distance (FD), and FAD.

Baselines We compare our method against two state-of-the-art T2AV approaches: TAVDiffusion (Mao et al., 2024) and CoDi (Tang et al., 2023), using the same text prompts for all models. Additionally, we include: (1) a Cascaded pipeline that uses Animatediff (Guo et al., 2023) for video generation followed by V2A-Mapper (Wang et al., 2024b) for audio synthesis, and (2) a Two-Stream approach in which video and audio are independently generated from the same text prompt using Animatediff and AudioLDM. For V2A generation, we also compare with the contrastive alignment method in (Xing et al., 2024) and SpecVQGAN (Iashin and Rahtu, 2021), a spectrogram-based audio generator employing vector quantization.

4.6.2 Main Results

T2AV Generative Quality Table 4.1 presents the quantitative comparison of our proposed method against existing T2AV generation baselines on the text-labeled VGGSound and AudioCaps datasets. Our method consistently achieves the best performance across all

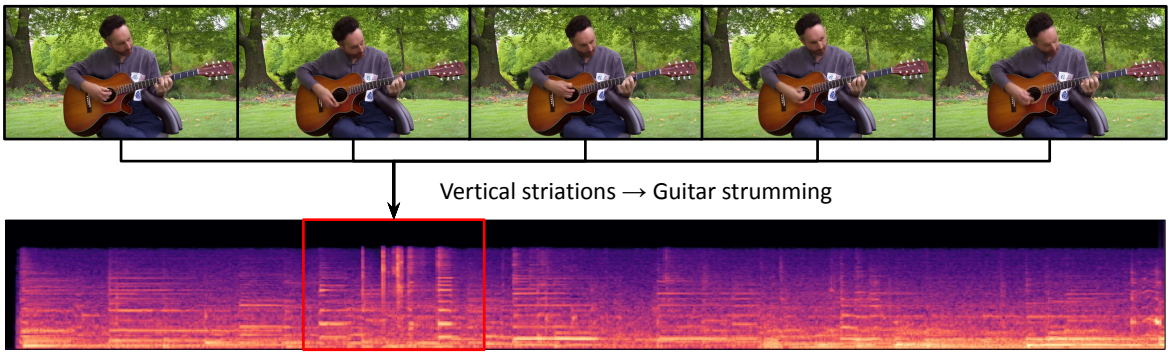

FIGURE 4.5. Temporal alignment between visual motion and acoustic patterns. The strumming motion of the guitarist's hand aligns with vertical striations in the spectrogram, indicating synchronized transient audio events.

reported metrics. On VGGSound, it reduces FVD and FAD to 662.9 and 5.49, respectively, reflecting significant improvements in both video and audio generation fidelity. Compared to the best-performing baseline (Tang et al., 2023), our method achieve a relative reduction of 6.5% in FVD and 16.0% in FAD. In terms of semantic alignment, our model achieves the highest AVHScore of 0.206 and 0.142 on VGGSound and AudioCaps, respectively, demonstrating improved correspondence between generated content and the input descriptions. A similar trend is observed in the qualitative results shown in Figure 4.4. For example, in the wolf scenario, our generated video better captures key semantic attributes such as the white head and yellow-brown body, while the visual background and ambient objects more faithfully reflect the textual prompt. Likewise, the ambulance scene displays correct object types, vehicle motion, and contextual elements like roadside greenery and traffic, showing high semantic fidelity across modalities.

Semantic and Temporal Alignment As shown in Table 4.1, our method achieves the highest CAVPSIM scores on both VGGSound and AudioCaps, indicating improvement on cross-modal temporal alignment. This is further illustrated in Figure 4.4, where the temporal evolution of audio patterns (e.g., barking or sirens) closely corresponds with visual events. In the wolf example, distinct spectrogram patterns align with repeated barking motions, while the ambulance scenario shows dynamic spectrogram textures matching siren intensity and vehicle motion. To further highlight this property, Figure 4.5 visualizes a man strumming a

Method	KL↓	ISc↑	FD↓	FAD↓
SpecVQGAN (Iashin and Rahtu, 2021)	3.290	5.108	37.269	7.736
SeeHear-Vani (Xing et al., 2024)	3.203	5.625	40.457	6.850
SeeHear-Full (Xing et al., 2024)	2.619	5.831	**32.920**	7.316
Ours	**2.128**	**5.677**	39.534	**6.155**

TABLE 4.2. Video-to-Audio Generation Results. Our method outperforms existing V2A baselines across most evaluation metrics, demonstrating noticeable improvements in audio fidelity.

Mask	FVD ↓	FAD ↓	AVHScore ↑	CAVPSIM ↑
□	662.9	6.95	0.175	0.141
◯	662.9	6.08	0.192	0.144
△	662.9	5.49	0.206	0.165

TABLE 4.3. Ablation study on masking input modalities. □: no masking, direct alignment, ◯: only takes video modality embeddings as the input, △: takes both video and audio modality embeddings as the input.

guitar, where the rhythmic hand motion aligns with vertically striated spectrogram features indicative of transient guitar strokes. These results collectively confirm that our method not only generates high-quality content but also preserves temporal synchronization across modalities. Additional examples are provided in Appendix D.

V2A Generation To further evaluate the effectiveness of cross-model alignment, we assess the video-to-audio (V2A) generation performance using a subcomponent of our model. As shown in Table 4.2, our approach outperforms existing V2A baselines: SpecVQGAN (Iashin and Rahtu, 2021) and SeeHear (Xing et al., 2024) across most metrics, achieving better KL, ISc and FAD. These results indicate that our model effectively captures the shared semantic and temporal information between video and audio, enabling high-quality cross-modal generation. The performance of this subcomponent further validates the robustness of our alignment strategy.

4.6.3 Ablation Study

Study on Mask Input We conduct an ablation study to assess the impact of different mask input configurations on cross-modal generation quality in Table 4.3. We observe that when no masking is applied (□), performance is significantly lower across all metrics, indicating that direct alignment without filtering introduces noise and misalignment. Conditioning the mask on video alone (◯) yields moderate improvements, suggesting that video features contain partial cues for predicting shared content. However, the best performance is achieved when

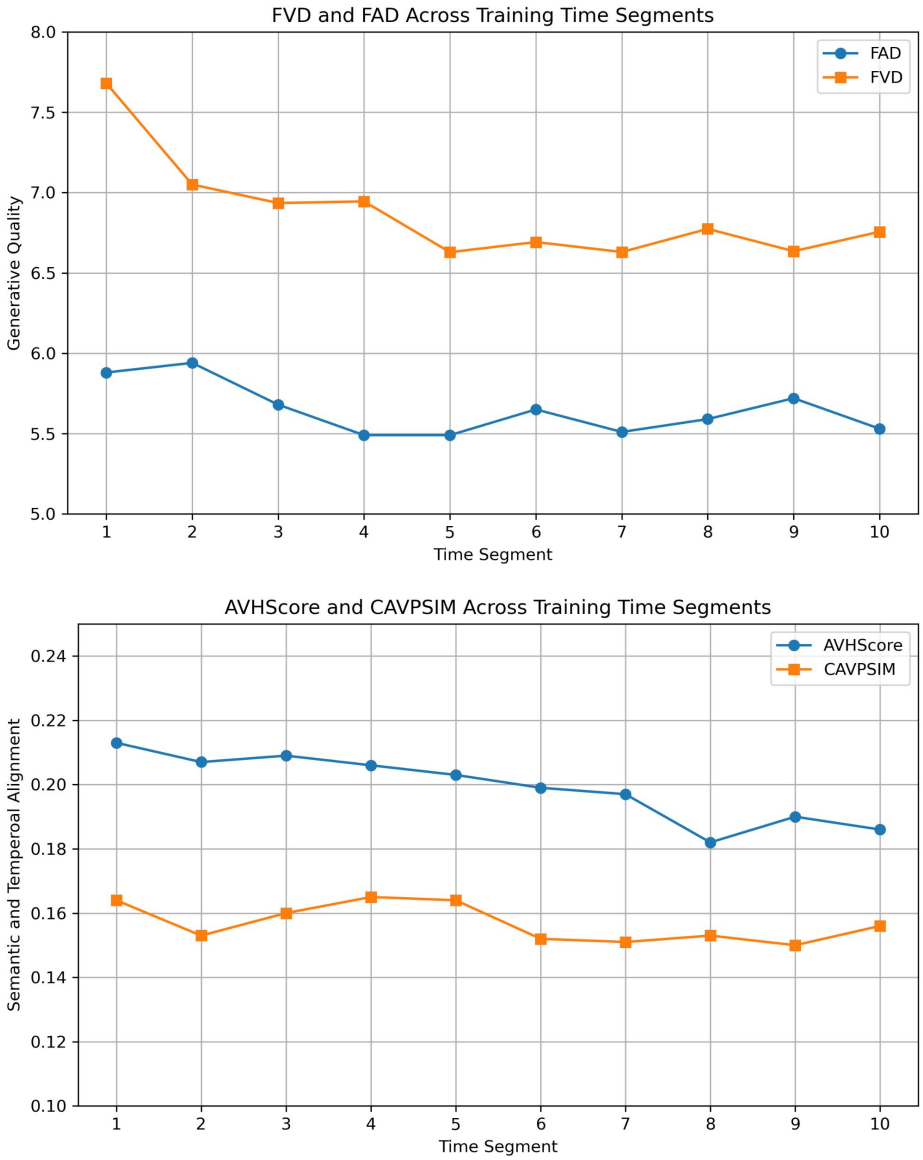

FIGURE 4.6. Ablation on different time segment lengths. We find that longer segments improve generative quality, while shorter segments benefit alignment.

the mask is conditioned on both video and audio embeddings (△), resulting in the lowest FAD and highest AVHScore and CAVPSIM. This confirms our hypothesis that observing both modalities enables the mask to more accurately isolate cross-modally relevant dimensions, thereby enhancing semantic consistency and temporal alignment in the generated outputs.

Effect of Temporal Segmentation We conduct an ablation study to investigate the impact of different temporal segmentation strategies on the performance of T2AV generation. Specifically, given a 10-second video/audio clip, we divide the content into sub-clips of varying lengths (ranging from 1s to 10s) and use the aligned segments for fine-tuning the pretrained diffusion models and training the alignment modules, including the masking functions and adapters. As illustrated in Figure 4.6, longer segment durations consistently improve generative quality, as measured by FAD and FVD, likely due to providing richer temporal context for fine-tuning the diffusion backbones. In contrast, shorter segments yield stronger performance in alignment metrics such as AVHScore and CAVPSIM, suggesting that temporally concise segments reduce misalignment and noise during cross-modal training. Based on this trade-off, we select a 5-second segment length as a balanced choice that supports both high generative fidelity and accurate audio-visual alignment.

4.7 Conclusion

We presented a multi-stage framework for text-to-audio-video (T2AV) generation that addresses the challenge of semantic and temporal misalignment between modalities. Our method introduces a masked latent adaptation mechanism that selectively aligns video representations with audio embeddings using a learnable adapter and relevance mask. During inference, we leverage a cascaded diffusion structure in which video is generated from text, and audio is subsequently synthesized conditioned on both text and the adapted video latent. This design ensures coherence across modalities while maintaining flexibility by reusing single-modal diffusion models. Extensive experiments demonstrate that our approach improves cross-modal consistency and achieves state-of-the-art results on multimodal generation benchmarks.

CHAPTER 5

Interpretable Latent Workflow Modeling for Agentic Reasoning

This chapter broadens the thesis beyond representation and generation to agentic reasoning in complex, real-world decision workflows, focusing on automated medical coding. Positioned as the final technical contribution, it reframes coding not as a static prediction problem but as a structured reasoning process composed of interacting modules and tools. The chapter identifies a key limitation of existing agentic systems: their reliance on manually designed workflows, which constrains both performance and adaptability. To address this, it introduces a framework that models workflow design itself as a learnable, interpretable latent process, enabling automated discovery and refinement of agentic pipelines through iterative design, execution, and reflection. By demonstrating how structured, interpretable workflow representations can be optimized under domain constraints, the chapter completes the thesis trajectory from controllable latent representations, to structured reasoning, to interpretable latent workflows for complex agentic decision-making.

5.1 Introduction

Medical coding is the process of translating unstructured clinical notes into standardized diagnostic and procedural codes, most commonly following the World Health Organization's International Classification of Diseases (ICD) standard (Dong et al., 2022a). These codes underpin critical functions in healthcare, including billing and reimbursement, hospital resource planning, and epidemiological research (Campbell and Giadresco, 2020). Unlike simple classification, manual coding requires coders to engage in a multi-step workflow: identifying relevant mentions in free text, consulting multiple resources such as the alphabetic

index and tabular index, applying coding guidelines, and cross-checking for consistency across diagnoses and procedures. This structured but intricate process makes medical coding highly labour-intensive and error-prone, contributing to global coding backlogs and clinical risks when errors occur (Alonso et al., 2020; Douglas et al., 2025; Gan et al., 2025).

To reduce the burden of manual coding, recent research has leveraged large language models (LLMs) as the foundation for automated systems (Boyle et al., 2023; Yang et al., 2023c; Falis et al., 2024; Baksi et al., 2025). Beyond their core ability to map free-text clinical notes to candidate codes, LLMs possess reasoning and tool-use capabilities that are particularly well-suited to the structured, rule-based nature of the coding process (Kwan, 2024; Mustafa et al., 2025).

Building on this, emerging work has proposed agentic coding workflows (Li et al., 2024c; Motzfeldt et al., 2025), where multiple interacting agents cooperate to mirror the steps taken by human coders: extracting relevant terms, consulting indexes and guidelines, and verifying consistency across diagnoses and procedures. This paradigm has demonstrated competitive performance and enhanced interpretability compared to treating coding as a flat multi-label classification problem.

Despite these advances, most existing agentic frameworks for medical coding remain manually crafted, relying on human experts to specify the modules and execution steps within a workflow (Motzfeldt et al., 2025). Yet this *design problem* is particularly challenging, as correct code assignment with LLM-based systems often hinges on two factors: (i) the quality of module definitions (e.g., tool calls, inference strategies), and (ii) the effectiveness of interactions across modules (i.e., how tools and strategies are combined and ordered). Manually fixing these design choices risks overlooking more effective coordination patterns, thereby limiting the potential of agentic approaches (Li et al., 2024b; Zhang et al., 2025b). For instance, automated search may discover non-obvious yet beneficial strategies, such as applying a *contrastive screening* step to prune near-duplicate ICD codes based on description similarity, whereas manual designs may fail to recognize such refinements. This motivates the need for automated workflow learning, which can flexibly search for and refine workflow designs rather than constraining systems to static, expert-defined pipelines.

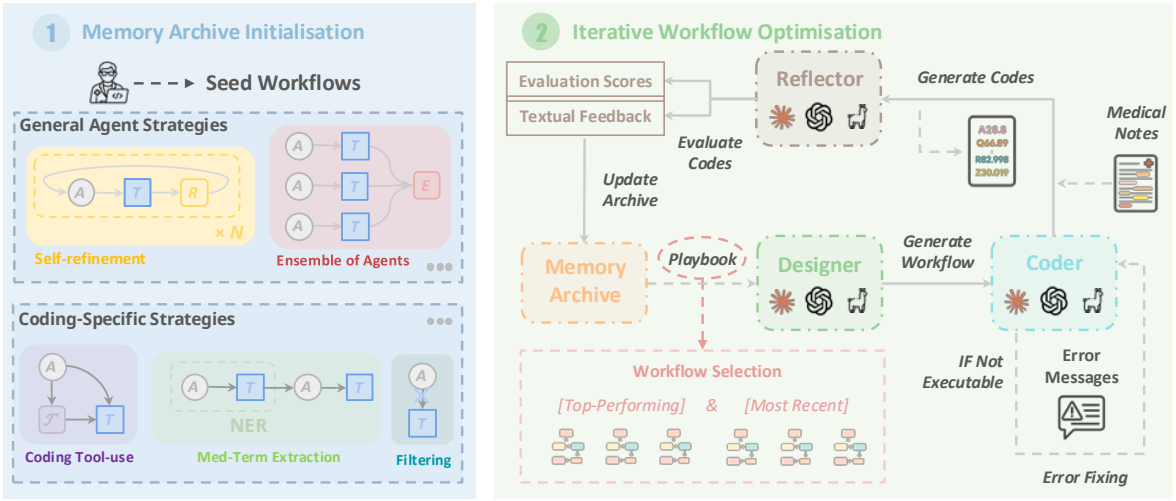

FIGURE 5.1. Overview of the MedDCR framework. (1) The *memory archive* is initialised with general reasoning strategies (e.g., self-refinement, multi-agent ensembles, chain-of-thought prompting) and coding-specific strategies (e.g., medical term extraction, weak code filtering, ICD tool use), together with other optional seed workflows. (2) In each optimisation loop, the *Designer* proposes new workflows, the *Coder* compiles and executes them (with self-fixing if needed), and the *Reflector* provides both evaluation scores and textual feedback. The memory archive stores all past workflows, enabling reuse, progressive refinement, and workflow selection from top-performing and recent designs. This closed-loop process discovers effective coding workflows under guideline constraints.

To address this challenge, we propose MedDCR (As demonstrated in Figure 5.1), an automated framework that optimises workflows for medical coding. Instead of relying on a single fixed pipeline, MedDCR treats workflow design as a learning problem (Hu et al., 2025; Zhang et al., 2025a; Zhang et al., 2025b; Zhou et al., 2025), where workflows are proposed, executed, and evaluated in an iterative loop.

Within this loop, a **D**esigner agent generates workflow plans by inventing or combining coding tools and strategies. A **C**oder agent then translates these plans into executable pipelines in the form of concrete programs, which carry out operations such as tool calling, validation, and reconciliation to predict medical codes. Finally, a **R**eflector agent evaluates the predictions, providing both performance scores and textual feedback on the effectiveness of the workflow design. The Designer uses this feedback to refine its proposals, enabling workflows to evolve and improve over time (Wang et al., 2025b). Beyond this loop, MedDCR maintains a memory

archive of prior designs, enabling workflows to be reused, progressively refined, or augmented with expert-crafted pipelines (Ji et al., 2024; Li et al., 2024c; Motzfeldt et al., 2025).

Our experiments demonstrate that the workflows discovered by MedDCR significantly outperform both state-of-the-art, hand-designed baselines and pretrained language model approaches. Specifically, MedDCR achieves a 6.2% improvement in F1 score on the MDACE (Cheng et al., 2023a) dataset and a 7.4% gain on ACI-BENCH (Yim et al., 2023), compared to the second-best performing baselines.

5.2 Related Works

Automated Medical Coding. Automated medical coding seeks to accelerate the mapping of free-text clinical notes to standardized medical codes using computer-assisted tools. Early systems were rule-based (Farkas and Szarvas, 2008; Kavuluru et al., 2015), but the availability of large EHR datasets such as MIMIC-III and MIMIC-IV (Johnson et al., 2016; Johnson et al., 2020) established deep learning models as the dominant approach, typically framing coding as an extreme multi-label classification task. Encoder–decoder architectures with label-wise attention (Mullenbach et al., 2018; Li and Yu, 2020; Vu et al., 2021) were extended with textual descriptions, synonyms, or co-occurrence signals to better align notes and codes (Cao et al., 2020; Dong et al., 2021; Yuan et al., 2022). More recently, pre-trained language models fine-tuned for ICD prediction (PLM-ICD) achieved state-of-the-art performance (Huang et al., 2022a; Edin et al., 2024; Douglas et al., 2025). However, these models struggle with the extremely large ICD label space, rare codes, and long clinical notes.

To address these challenges, researchers have explored LLMs for generative coding. Zero and few-shot prompting (Yang et al., 2023b; Boyle et al., 2023; Gero et al., 2023) showed flexibility but underperformed PLM-based classifiers. This has motivated agentic approaches, where LLMs interact with external tools, indexes, and validation routines in multi-agent workflows that mimic human coding processes (Li et al., 2024c; Kwan, 2024; Motzfeldt et al., 2025). While promising, these workflows remain manually designed and fixed, making them potentially suboptimal.

Our work addresses this gap by introducing MedDCR, which treats workflow design as a learning problem and automatically searches for effective agentic workflows for medical coding, enhancing both predictive performance and interpretability.

Agentic Workflow Design. Agentic workflows frame problem solving as a coordinated process across one or more LLM-based agents, each assigned specific roles or equipped with external tools. Recent advances enrich these systems with prompting strategies (Chen et al., 2023a), chain-of-thought planning (Wei et al., 2022; Zhang et al., 2023d), reflection and refinement (Madaan et al., 2023; Shinn et al., 2023; Dhuliawala et al., 2024), tool use (Nakano et al., 2021; Schick et al., 2023; Qin et al., 2024), and role assignment for multi-agent cooperation (Shanahan et al., 2023; Li et al., 2023a; Wu et al., 2024a). Multi-agent topologies vary from parallel setups for exploration (Wang et al., 2023c) to serial pipelines with reflective refinement (Madaan et al., 2023), and even debate-style structures that improve reliability (Du et al., 2023; Liang et al., 2024).

Building on these foundations, the community has begun exploring automated design of agentic systems. Most works focus on automating prompt optimization (Yang et al., 2023a; Fernando et al., 2024; Khattab et al., 2024), role definition (Li et al., 2023a; Wu et al., 2024a), or specific topology search (Zhou et al., 2025; Wang et al., 2025b). Others attempt to expand the search space to workflows (Hu et al., 2025; Zhang et al., 2025a; Zhang et al., 2025b), where both the definition of workflow components and their topological organisation are jointly optimised, but in practice, many components remain fixed, making discovered agents less flexible.

Our work advances this line by introducing MedDCR, which treats agentic workflow design as a learning problem. Unlike prior approaches, MedDCR employs a meta-agent architecture with a Designer, Coder, and Reflector, reinforced by a memory archive that supports reuse and refinement of workflows. Coupled with medical coding tools and guideline-driven strategies, MedDCR provides the first domain-specific framework for automated optimisation of medical coding workflows, offering stronger performance and greater interpretability.

5.3 Interpretable Workflow Construction for Medical Coding

5.3.1 Problem Formulation

We formalize automated workflow optimisation problem for medical coding as follows. Let $\mathcal{X}$ denote the space of clinical notes and $\mathcal{C}$ the set of admissible ICD codes. A dataset is

$$D = \{(x_i, Y_i)\}_{i=1}^{n}, \quad x_i \in \mathcal{X},\ Y_i \subseteq \mathcal{C}. \tag{5.1}$$

A *workflow* W induces a coder function

$$f_W : \mathcal{X} \to 2^{\mathcal{C}}, \tag{5.2}$$

that maps a note to a predicted set of codes.

Workflows are constructed from a component library $\mathcal{L}$ consisting of: (i) a tool set $\mathcal{T}$ (e.g., ICD index retrieval, guideline validators), (ii) reasoning/strategy primitives Σ (e.g., extraction, validation, reconciliation), and (iii) LLM modules $\mathcal{M}$ with parameter configurations Θ.

We represent a workflow as a directed acyclic graph

$$W = (G = (V, E), \phi), \tag{5.3}$$

where each node $v \in V$ instantiates a component $c_v \in \mathcal{L}$ with parameters $\phi_v \in \Theta$, and edges E define data/control flow. The *search space* of feasible workflows is

$$S = \left\{ W = (G, \phi) \;\middle|\; \begin{array}{l} G \text{ is a DAG over } \mathcal{L}, \\ W \text{ is executable} \end{array} \right\}. \tag{5.4}$$

Medical coding is constrained by official guidelines. Let Γ denote the guideline resource (ICD Alphabetic/Tabular Index, coding rules). A compliance oracle $V_\Gamma(W, x) \in [0, 1]$ measures the fraction of guideline checks that pass for workflow W on input x. We evaluate a workflow with a task metric $g(\cdot, \cdot)$ (e.g., micro/macro F_1) and define the objective on a validation set

D_{val}:

$$G(W; D_{\text{val}}, \Gamma) = \mathbb{E}_{(x,Y)\sim D_{\text{val}}}\Big[g(f_W(x), Y) - \lambda_{\text{viol}}\big(1 - V_\Gamma(W, x)\big)\Big] - \lambda_{\text{cost}}\, C(W), \tag{5.5}$$

where $C(W)$ measures resource usage (e.g., tool calls, latency, tokens) and $\lambda_{\text{viol}}, \lambda_{\text{cost}} \geq 0$ control the trade-offs between predictive performance, guideline compliance, and efficiency.

The automated workflow optimisation problem for medical coding is then:

$$W^\star \in \arg\max_{W\in\mathcal{S}} \; G(W; D_{\text{val}}, \Gamma). \tag{5.6}$$

In the remainder of this section, we detail how MedDCR performs the iterative search for $W^\star$ guided by the objective G and the constraints Γ.

5.3.2 MedDCR Framework Overview

MedDCR framework instantiates the workflow optimisation problem defined in Section 5.3.1 by organising the search process into an iterative loop of design, coding (execution), and reflection. The framework aims to automatically discover effective workflows for medical coding without relying on fixed, manually crafted pipelines.

The loop begins with the initialisation of a memory archive $\mathcal{H}_0$, which contains a small collection of seed workflows (See Figure 5.1). These seeds capture both general reasoning strategies, such as chain-of-thought (Wei et al., 2022; Zhang et al., 2023d), self-refinement (Madaan et al., 2023), and multi-agent debate (Du et al., 2023) and domain-specific medical coding strategies, including medical entity extractions (Douglas et al., 2025), tabular index similarity checks, and reconciliation heuristics (Gan et al., 2025), among others. The archive also stores a list of coding tools, such as alphabetic index search, parent–child code lookup, and code–description extraction modules. This initial set provides the system with a foundation of plausible behaviours, but the search is not limited to them, where the archive can expand with newly discovered strategies, including tool calls proposed dynamically by LLMs.

Once initialised, the search process proceeds iteratively through a *Design–Execute–Reflect* cycle. At iteration t, a new workflow plan π_t is proposed by an LLM-based Designer agent (Li et al., 2023a; Wu et al., 2024a), drawing on the memory archive that contains both the seeded workflows and prior designs $\mathcal{H}_{0:t-1}$. Each plan specifies a combination of tools, strategies, and reasoning steps, as well as their execution order. The plan is then compiled into an executable workflow W_t, translated into runnable code by a Coder agent, and subsequently executed to produce medical code predictions.

The execution results are assessed with an evaluation function $G(W_t)$, which integrates predictive performance (e.g., F1 score), guideline compliance (Ann Barta, 2009), and computational cost. Alongside the scalar score, the evaluation produces textual feedback produced by a *Reflector* agent, diagnosing workflow errors such as ineffective tool combinations or guideline violations. Both the score and feedback are appended to the archive as a record (π_t, W_t, s_t, r_t).

The memory archive thus grows with every iteration, providing two key signals for the search: (i) high-performing past workflows, which act as exemplars to imitate or refine, and (ii) diverse recent workflows, which encourage exploration. New proposals are therefore shaped both by the accumulated experience in the archive and by the evaluation feedback from prior iterations (Hu et al., 2025). Over time, the system learns to discover increasingly effective workflows by combining coding tools and strategies in novel ways.

The search loop terminates once the budget of iterations is reached or when performance converges. The final output is the best workflow $W^\star$ found in the archive, which can be used directly for automated medical coding or supplied as a strong starting point for further optimisation.

5.3.3 Meta-Agent Architecture

In this section, we present the details of the meta-agents within the MedDCR framework.

Designer Agent. The Designer agent is responsible for generating candidate workflow plans π. At iteration t, it outputs

$$\pi_t = \text{Designer}(\mathcal{H}_t, \mathcal{L}, \Gamma) \in \Pi, \tag{5.7}$$

where $\mathcal{H}_t$ is the memory archive, $\mathcal{L}$ is the component library (tools, strategies, and LLM modules), and Γ encodes coding guidelines. The Designer operates under a structured *meta-prompt* that (i) specifies constraints on how workflows must be formatted and executed, (ii) enumerates the available tool and strategy signatures, and (iii) appends informative exemplars from the memory archive.

In particular, the prompt shows the top-k best-performing workflows and the most recent n designs, ensuring that the Designer can exploit strong prior solutions while maintaining exploration.

Coder Agent. The Coder agent transforms abstract workflow plans into executable programs (e.g., Python-like codes). Given a plan π_t, it compiles the abstract plan into

$$W_t = \text{Coder}(\pi_t) \in \mathcal{S}, \tag{5.8}$$

where W_t is an operational program that can be executed on clinical notes to generate medical codes.

In practice, however, generated code may contain *syntax or execution errors*, which would otherwise block evaluation. To address this, the Coder incorporates a *self-fixing loop*: the executor first checks whether the compiled code runs successfully, if an error occurs, the Coder attempts to repair the code automatically and retries execution until a valid workflow is obtained or a retry budget is exceeded (Joshi et al., 2023; Zhang et al., 2024a).

This mechanism ensures that the search process is not derailed by syntactic inconsistencies (Olausson et al., 2024). The separation of roles is thus preserved: the Designer explores high-level workflow planning, while the Coder guarantees that plans are translated into syntactically correct and executable implementations.

Reflector Agent. The Reflector evaluates executed workflows and provides targeted feedback. Specifically, for each workflow W_t, it outputs

$$(s_t, r_t) = \text{Reflector}(W_t, D_{\text{val}}), \tag{5.9}$$

where $s_t = G(W_t; D_{\text{val}}, \Gamma)$ is a scalar score that integrates predictive accuracy (e.g., F1 score or precision/recall), guideline compliance, and computational efficiency, and r_t is a textual critique.

To generate this feedback, the Reflector collects the intermediate outputs of each LLM call within the workflow and interprets them in the context of the corresponding operation. For example, it may highlight when an entity extraction step misses key mentions, when a guideline validator rejects a candidate code, or when reconciliation produces redundant outputs. The tuple (π_t, W_t, s_t, r_t) is then appended to the archive, which is updated as

$$\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(\pi_t, W_t, s_t, r_t)\}. \tag{5.10}$$

This combination of quantitative scoring (Khattab et al., 2024) and fine-grained textual feedback (Yang et al., 2023a; Pryzant et al., 2023) ensures that subsequent iterations improve not only raw performance but also the robustness and interpretability of workflows.

5.3.4 Memory Archive and Plug-and-Play

A central component of our proposed framework is the memory archive, which records the history of all explored workflows. At iteration t, the archive $\mathcal{H}_t$ contains tuples

$$\mathcal{H}_t = \{(\pi_j, W_j, s_j, r_j)\}_{j<t}, \tag{5.11}$$

where π_j is the workflow plan proposed by the Designer, W_j is the compiled executable workflow, s_j is the score, and r_j is the textual feedback.

This archive plays a dual role: it enables *reuse* of effective workflows by presenting the top-k highest scoring exemplars to the Designer, and it supports *exploration* by also presenting

the n most recent workflows, which prevent the search from collapsing into repetitive local optima (Zhu et al., 2023).

In this way, the Designer is encouraged to learn from successful past patterns while avoiding premature convergence to a single family of workflows (Zhong et al., 2024). As the search progresses, the archive grows into a structured memory of workflow designs, making it an adaptive optimiser that improves continuously over time.

Beyond storing internal proposals, the memory archive also supports a *plug-and-play* mode. Here, external expert-crafted workflows, denoted $\mathcal{W}_{\text{seed}}$, are injected into the initial archive $\mathcal{H}_0$. These workflows may come from prior research, human-designed pipelines, or established clinical coding heuristics (Gan et al., 2025; Douglas et al., 2025; Motzfeldt et al., 2025). Once seeded, MedDCR treats them like any other entry in the archive: they can be directly reused, refined through feedback, or combined with newly generated strategies.

This property makes our framework workflow-agnostic: it can either discover new workflows from scratch or improve upon existing designs, depending on the application scenario. The final output is therefore not limited to the best design discovered during search, but can also include optimised variants of expert workflows, providing a bridge between automated search and human expertise.

5.4 Experiments

5.4.1 Experimental Setup

Dataset. We evaluate our framework on two publicly available ICD-10 coding benchmarks. MDACE (Cheng et al., 2023a) provides expert-verified ICD-10 annotations to a mix of inpatient and professional-fee charts drawn from MIMIC-III (Johnson et al., 2016), including discharge summaries, radiology reports and physician notes. ACI Benchmark (Yim et al., 2023) is a synthetic dataset of clinical notes paired with ICD-10 codes for benchmarking

automated coding systems. Together, they enable evaluation of both coding accuracy and explainability in real-world settings.

Baseline. We benchmark MedDCR against a broad set of existing approaches spanning *three* paradigms. First, we consider pre-trained language model methods that frame coding as multi-label classification and require re-training on coding datasets, including PLM-ICD(Huang et al., 2022a), and PLM-CA (Douglas et al., 2025). Second, we include expert-designed medical workflows that rely on heuristic rules or structured coding pipelines, represented by RRS (Kwan, 2024), MAC (Li et al., 2024c), and CLH (Motzfeldt et al., 2025). Finally, we evaluate against general agentic strategies and automated search frameworks that build on large language models with structured reasoning or multi-agent coordination, such as Chain-of-Thought (Wei et al., 2022), Self-Consistency (Wang et al., 2023d), Multi-Debate (Du et al., 2023), Self-Refine (Shinn et al., 2023), and NER (Goel, 2025), as well as recent optimisation frameworks like ADAS (Hu et al., 2025). GPT-4o (Hurst et al., 2024) serves as the backbone model for all baseline agents. Together, these baselines span model-based, rule-based, and agentic paradigms, enabling a comprehensive comparison.

Implementations. We use GPT-based models as the backbone for MedDCR, evaluating both GPT-4o (Hurst et al., 2024) and GPT-5 (OpenAI, 2025). The search loop is run for 100 iterations, with the Designer conditioned on the top-5 highest scoring and 3 most recent workflows from the archive at each step. To initialise the archive, we include all baselines from the aforementioned agentic workflow group, ensuring a diverse starting pool. Coding tools are modified and adapted from the simple-icd-10 library, covering both ICD-10-CM and PCS codes. More details are in the Appendix D4.

Evaluation Metrics. We report standard measures for extreme multi-label coding: micro-F1, precision, and recall (Huang et al., 2022a; Edin et al., 2024; Douglas et al., 2025). Micro-F1 balances overall precision and recall across the large code space, providing a primary measure of workflow effectiveness. Precision reflects the system's ability to avoid spurious code assignments, while recall captures its capacity to recover the full set of relevant codes. Together, these metrics indicate how well the proposed workflows navigate the

Method Category (Label Space)	Model	MDACE			ACI-BENCH		
		Precision	Recall	F1	Precision	Recall	F1
PLM (1K)	ICD (Huang et al., 2022a)	**0.49**	0.47	**0.48**	0.43	0.41	0.42
	CA (Douglas et al., 2025)	**0.46**	0.45	0.45	**0.44**	0.42	0.43
Coder Workflow (1K)	RRS (Kwan, 2024)	0.24	0.30	0.27	0.26	0.52	0.35
	MAC (Li et al., 2024c)	0.27	0.31	0.29	0.23	0.50	0.31
	CLH (Motzfeldt et al., 2025)	0.45	0.40	0.42	**0.44**	0.39	0.41
Agentic Method (1K)	CoT (Wei et al., 2022)	0.30	0.31	0.30	0.35	0.50	0.41
	CoT-SC (Wang et al., 2023c)	0.39	0.43	0.41	0.36	0.59	0.44
	MulDe (Du et al., 2023)	0.21	0.40	0.28	0.16	0.65	0.25
	Judge (Zheng et al., 2023b)	0.26	0.53	0.35	0.22	0.64	0.33
	MNER (Goel, 2025)	0.13	0.48	0.20	0.15	**0.67**	0.25
	ADAS (Hu et al., 2025)	0.37	0.51	0.43	0.28	0.59	0.43
	MedDCR-GPT-4o	0.41	**0.55**	0.47	0.36	0.65	**0.46**
	MedDCR-GPT-5	**0.46**	**0.59**	**0.51**	**0.43**	**0.67**	**0.52**

TABLE 5.1. Main results on MDACE and ACI-BENCH datasets. The best results are highlighted in bold, and the second-best results are shown in gray bold. Methods are grouped into three categories: Pretrained Language Models, expert-designed coding workflows, and agentic workflow strategies (including agent-based search methods).

Method	**MDACE**			**ACI-BENCH**		
	Search Tokens	Exec. Tokens	Cost (USD)	Search Tokens	Exec. Tokens	Cost (USD)
MedDCR	19,994	11,545,586	**$17.09**	14,294	2,233,623	**$5.72**

TABLE 5.2. Computation Cost Comparison. Token usage and projected cost in USD per 100 inference samples per search loop on MDACE and ACI-BENCH datasets. While effective, our method remains cost-efficient.

trade-off between exploration (capturing diverse correct codes) and exploitation (maintaining accuracy).

5.4.2 Main Results

As shown in Table 5.1, MedDCR consistently outperforms all baselines on both MDACE and ACI-BENCH. On MDACE, MedDCR-GPT-5 achieves a Micro-F1 of 0.51, improving over the strongest PLM baseline (ICD) and the best agentic baseline (CoT-SC) by 6%. On ACI-BENCH, MedDCR-GPT-5 reaches 0.52 for F1, again surpassing PLM and agentic baselines. These results demonstrate that automated workflow search yields more effective

Seed Setting	MDACE		
	Precision	Recall	Micro-F1
CoT-SC (seed only)	0.37	0.48	0.42
+ Quality-Diversity	0.38	0.48	0.42
+ Multi-Debate	0.37	0.53	0.44
+ NER	0.35	0.55	0.43
+ All auxiliary	**0.41**	**0.55**	**0.47**
No seed (scratch)	0.39	0.51	0.44

TABLE 5.3. Plug-and-play validation: MedDCR initialised with CoT-SC as the primary seed, optionally augmented with auxiliary seeds (Self-Refine, Multi-Debate, Med-NER). Optimisation consistently improves the baseline CoT-SC workflow and outperforms search from scratch.

	Precision	Recall	Micro-F1
MedDCR-GPT-4o	0.41	0.55	0.47
– No Top-k	0.35	0.42	0.38
– No Recent-n	0.33	0.34	0.34
– No Score Feedback	0.39	0.47	0.43
– No Text Feedback	0.44	0.50	0.47
– No Guideline in Meta	0.40	0.57	0.47
– No Exemplar in Meta	0.36	0.51	0.42

TABLE 5.4. Ablation study on MDACE dataset. Each row removes one core component of MedDCR. The performance drops confirm the importance of the workflow exemplars, reflector feedback, guideline constraints and few-shot exemplar in achieving the full performance.

coding strategies than either retrained language models or expert-crafted agentic pipelines. Importantly, Table 5.2 demonstrates that these improvements are achieved with modest overhead. Despite the iterative search process, MedDCR remains cost-efficient: a single search loop requires fewer than 20k search tokens on MDACE and 15k on ACI-BENCH, with total costs of $17.09 and $5.72 per 100 inference samples, respectively. Notably, the vast majority of cost arises from execution tokens rather than search tokens, showing that workflow optimisation overhead is modest relative to workflow execution. Finally, Table 5.3 validates MedDCR's plug-and-play property. Using CoT-SC as the primary seed, optimisation alone improves F1 from 0.41 to 0.42. When augmented with auxiliary seeds, performance increases further, reaching 0.47 F1 with all auxiliaries. Compared to search from scratch, seeded

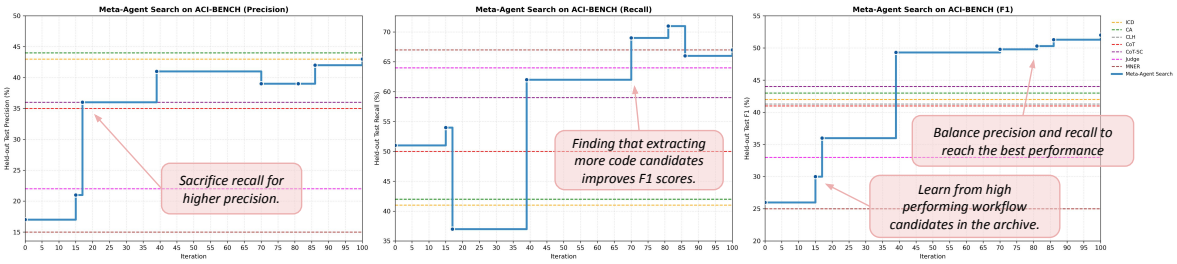

FIGURE 5.2. Case study of the search process on ACI-Bench. The blue line tracks the best workflow discovered at each iteration, measured by F1. The figure illustrates how performance improves as the system explores diverse candidates, learns from high-performing workflows, and balances precision and recall to refine the final design.

optimisation also achieves higher final scores. This confirms that MedDCR can flexibly refine and use existing workflows and diverse strategies, making it practical as a general workflow optimiser for medical coding.

5.4.3 Ablation Studies

Table 5.4 reports ablations on MDACE, where we systematically remove key components of MedDCR. Removing the top-k or recent-n workflows causes the largest performance drops of 0.09 and 0.13 in F1, confirming that both exploitation of strong designs and diversity from recent ones are essential for effective search. Feedback from the reflector is also critical: without score feedback, F1 falls to 0.43, while removing textual feedback prevents further improvements despite retaining numeric scores. Guideline constraints show a trade-off, slightly increasing recall but offering no net F1 gain when excluded. Finally, removing all exemplars from the meta-prompt reduces F1 to 0.42, underscoring their role in stabilizing the workflow generation. Overall, the results highlight that MedDCR's gains arise from the complementary contributions of memory selection, reflective feedback, guideline grounding, and exemplar conditioning.

5.4.4 Case Study

Figure 5.2 illustrates how our framework improves over the course of the search in terms of precision, recall, and F1, with the blue line indicating the best workflow discovered at each iteration as measured by F1. In the early stages, the search explores relatively simple strategies, while later iterations integrate more complex tool use and validation steps, resulting in steady performance gains. This progression highlights the role of memory and reflective feedback in escaping suboptimal designs and converging toward stronger workflows. The final best workflow combines multiple reasoning strategies and coding tools in a coordinated sequence, exemplifying the kind of designs that MedDCR can automatically uncover. Due to space, the best workflow structure is provided in the Appendix D3.

CHAPTER 6

Conclusion

This thesis has examined how latent representations and latent structures can be deliberately organised to support interpretability and controllability in vision and closely related multimodal learning systems. Motivated by the limitations of opaque and entangled latent spaces in contemporary deep learning models, the work set out to understand how task-relevant semantic information can be distinguished from task-irrelevant variation, and how such structure can be incorporated into learning objectives and model design.

The primary contribution of this thesis is the articulation and empirical validation of a structured latent modeling perspective, in which interpretability and controllability are treated as design objectives rather than incidental properties. Across a range of learning paradigms, the thesis demonstrates that latent spaces benefit from exposing internal organisation aligned with task semantics. By moving beyond monolithic latent embeddings toward selectively structured representations, the proposed approaches enable more transparent analysis, targeted intervention, and predictable model behaviour.

Central to this perspective is the principle of selective modeling. Instead of enforcing uniform invariance or alignment across latent dimensions, the thesis shows that selectively preserving information that is semantically or causally relevant to a task, while suppressing irrelevant variation, leads to latent representations that are both more interpretable and more controllable. This principle is shown to be applicable across diverse settings, including perceptual learning, visual reasoning, and multimodal generation, providing a unifying framework for understanding how latent structure can be shaped to meet task-specific requirements.

Beyond internal feature representations, this thesis further establishes that latent structure can arise at the level of reasoning processes and computational workflows. By treating such workflows as latent procedural constructs that can be learned, constrained, and analysed, the thesis demonstrates that the same modeling principles extend naturally to complex reasoning systems. This generalisation highlights that controllable and interpretable latent modeling is not limited to representation spaces, but can also govern how models organise and execute sequences of operations.

Throughout the thesis, emphasis has been placed on approaches that integrate naturally with existing deep learning frameworks. The proposed methods rely on principled objectives, constraints, and training strategies, rather than heavy supervision or extensive architectural modification. This design choice demonstrates that meaningful latent structure can be induced without sacrificing practicality or compatibility with modern large-scale models.

In summary, this thesis has contributed a coherent framework for understanding and designing latent representations and latent processes that support interpretable and controllable model behaviour. By explicitly structuring latent spaces and workflows around task-relevant semantics, the work advances a principled approach to latent modeling that applies across vision, multimodal learning, and agentic reasoning. Together, these contributions clarify how latent structure can be leveraged to move beyond opaque end-to-end systems toward learning models whose internal behaviour can be systematically understood and controlled.

6.1 Future outlook

The findings and methodologies developed in this thesis establish a foundation for several promising avenues of future research.

Causal and Interventional Latent Modeling. The approaches developed in this thesis focus on structuring latent representations and workflows according to task semantics, but they do not explicitly model causal relationships among latent factors. Future work could incorporate causal assumptions or interventional objectives to further disentangle latent variables and

to support stronger guarantees under distribution shift. Integrating causal inference with selective latent modeling may provide deeper insight into how latent factors influence model predictions and how controlled interventions can be performed reliably.

Interactive and Continual Learning. The proposed methods are evaluated primarily in static, offline learning scenarios. An important direction for future research is to extend structured latent modeling to interactive, continual, or lifelong learning environments, where task requirements and data distributions evolve over time. In such settings, maintaining interpretable and controllable latent structure while accommodating continual adaptation poses additional challenges, particularly in avoiding catastrophic forgetting and uncontrolled drift in latent representations.

Latent Modeling in Large-Scale Agentic Systems. This thesis demonstrates that latent modeling principles can be extended to agentic workflows, but the scope is limited to relatively constrained reasoning pipelines. Future work could explore how structured latent representations and workflows scale to more complex multi-agent or hierarchical systems, where coordination, memory, and long-term planning play a central role. Understanding how interpretability and controllability can be preserved as agentic systems grow in complexity remains an open problem.

Human-Centred Evaluation. The interpretability of latent representations and workflows in this thesis is primarily assessed through structural and behavioural analyses. Future research could incorporate human-centred evaluation protocols to better understand how these latent structures align with human expectations, reasoning processes, and trust. Such studies may help bridge the gap between technical notions of interpretability and practical usability in real-world applications.

APPENDIX A

Appendix of Chapter 2

A1 Ablation Analysis

Number of Augmented Instances. We conduct an experiment to assess the sensitivity of the models to the number of augmented instances. Specifically, we adjust the number of views from the set [1,3,5,10] and 'dynamic', and exam the performance on POE w/ OCP-CL. As illustrated in Figure A.1, the models' performance remains relatively stable when the number of augmented instances ranges between 3 and 5. However, over-supplementing the data with augmented instances can lead to a degradation in model performance. Interestingly, we find that a dynamic number of augmentations depending on the class could benefit the models. This is particularly relevant for ordinal regression datasets that suffer from class imbalance. Specifically, by increasing the number of instances in underrepresented classes, we observe an improvement in the overall performance of the methods.

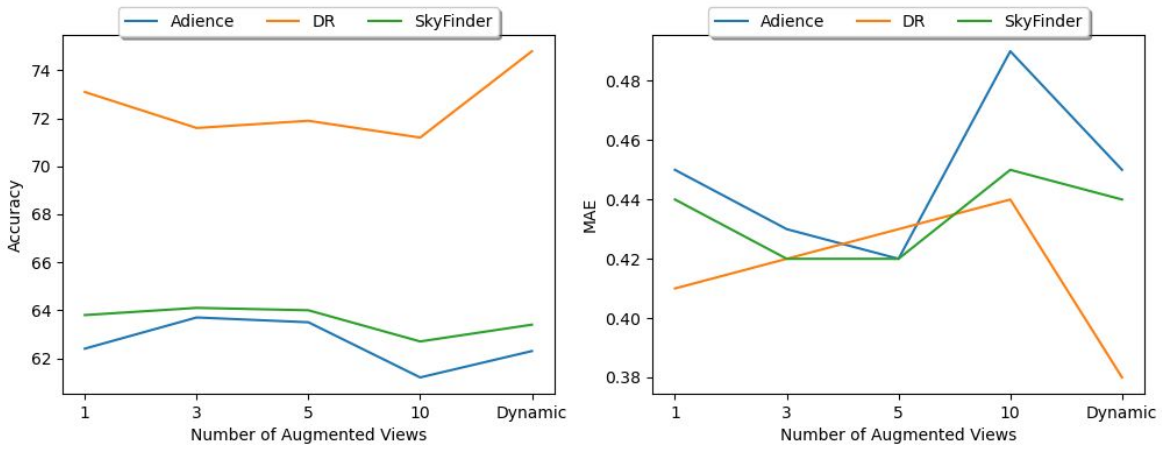

FIGURE A.1. Ablation Study on the Number of Augmented Views.

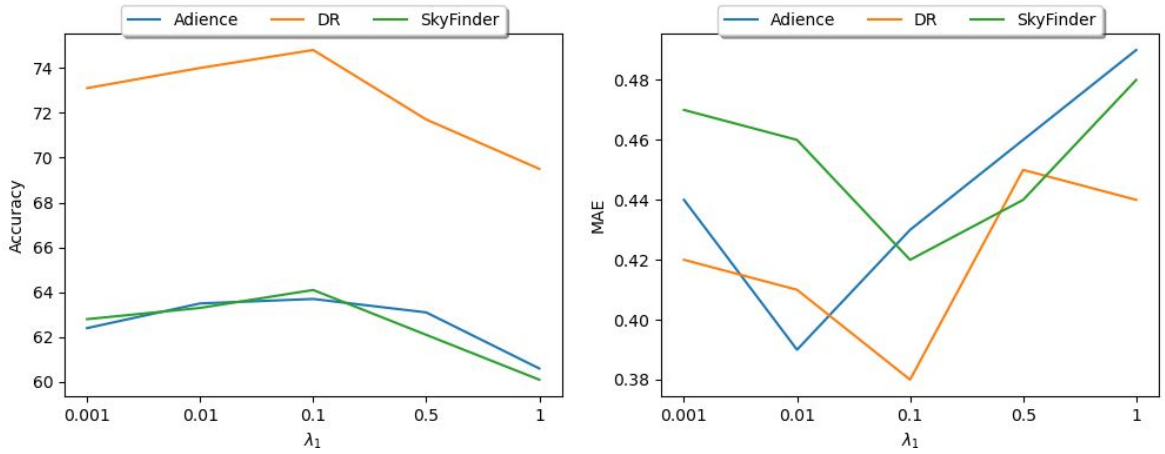

FIGURE A.2. Sensitivity Analysis on λ_1 ratio.

Sensitivity Analysis on λ_1. We conduct experiments to assess the impact of the mask sparsity ratio, denoted as λ_1, on the performance of the ordinal regression model across three downstream tasks. For this purpose, we utilize POE w/ OCP-CL to examine sensitivity to changes in λ_1. We test five different sparsity ratios within the range of 1e-3 and 1. The results, presented in Figure A.2, indicate that the ordinal regression model achieves optimal performance when λ_1 is set to 0.1 for all downstream tasks, and its performance decreases linearly with increases in the ratio beyond 0.1.

A2 Additional Related Works

In this section, we provide additional background discussions relevant to our work. Specifically, we discuss recent advancements in Generative Data Augmentation, Disentangled Representation Learning, and Nonlinear ICA and explore their relationship with our research.

Generative Models for Data Augmentations. Instead of generating data augmentations using predefined transformations, Generative Data Augmentation (GDA) employs an alternative approach that leverages Deep Latent Variable Models (DLVMs) to generate new synthetic views from existing samples, based on conditional generative processes. Antoniou et al., 2017 and Tran et al., 2017 propose the use of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Xia et al., 2022) to create a broader set of augmented data. More recently, Diffusion Models (DMs) (Sohl-Dickstein et al., 2015) have been utilized to alter high-level

semantic attributes, thereby addressing the problem of lack of diversity along key semantic axes in data augmentation (Trabucco et al., 2023). While these approaches can generate impressive results that appear both realistic and novel, most of them are not guaranteed to maintain the invariance of the original data. Our proposed method fulfills the need for controllable generative data augmentation, offering a more trustworthy GDA approach.

Disentangled Representation Learning. The objective of Disentangled Representation Learning (DRL) is to construct a model proficient in recognizing and isolating the latent factors concealed within observable data (Wang et al., 2022a). This isolation into semantically meaningful factors enhances the model's ability to produce interpretable data representations, thereby simulating the cognitive processes humans employ in understanding objects or relationships. In the context of generative modeling, Higgins et al., 2017a introduce a β-penalty coefficient for the KL divergence term in the evidence lower bound of a Variational Autoencoder (VAE) (Kingma and Welling, 2013; Li et al., 2022b; Huang et al., 2022b; Hong et al., 2024b; Lin et al., 2023) to balance latent channel capacity and independence constraints with reconstruction accuracy. Subsequently, various modifications to VAE have been introduced to improve its capability for disentanglement (Chen et al., 2018; Kumar et al., 2017). These include the incorporation of either implicit or explicit inductive biases as well as the utilization of diverse regularization techniques. On the other hand, InfoGAN (Chen et al., 2016) was the first to address the problem of disentangling latent factors in Generative Adversarial Networks, introducing an extra variational regularization of mutual information. Lin et al., 2019 introduced InfoGAN-CR, an unsupervised extension of InfoGAN that includes a contrastive regularizer to infer latent dimensions. Zhu et al., 2021 present PS-SC GAN, which builds upon InfoGAN and features a Spatial Constriction (SC) strategy to extract significant areas influenced by each latent dimension, along with a Perceptual Simplicity (PS) approach to make the latent factors more unambiguous. Wei et al., 2021 propose a method known as Orthogonal Jacobian Regularization (OroJaR) aimed at enhancing disentanglement in generative models. OroJaR uses the Jacobian matrix to examine how output alterations correspond to changes in input variables, specifically the latent dimensions. Our methods are parallel to GAN-based DRL methods, wherein we disentangle the latent factors by introducing the principle of minimal change.

Nonlinear ICA. Nonlinear independent component analysis (ICA) theoretically addresses the problem of disentangling latent factors when a nonlinear invertible transformation function exists, mapping independent samples to the latent space Hyvarinen and Morioka, 2016; Hyvarinen and Morioka, 2017. Recent developments Locatello et al., 2020; Zimmermann et al., 2021; Xie et al., 2022; Kong et al., 2022 indicate that, within a conditional generative process, the true latent factors might become identifiable when auxiliary information is provided. Khemakhem et al., 2020 demonstrate that the joint data and latent space distributions can be recovered, up to a simple transformation in the latent space, provided the generative process conditions on a variable observed alongside the data. Von Kügelgen et al., 2021 employ two views of the same image to disentangle the latent factors into ordinal content and non-ordinal components, with only the content component associating with the image's semantics. Our generative model leverages Nonlinear ICA theories to theoretically justify the disentanglement of latent factors. By manipulating the ordinal content variable, while keeping the ordinal content factors consistent, our model can generate augmented views that preserve ordinal content.

A3 Intuition of Why Augmenting non-ordinal factors

Here, we provide further discussion on why augmenting non-ordinal factors in images can benefit the training of ordinal regression/classfication models. In general, data augmentation aims to modify the styling factors in original examples that are not related to the predictive objectives of downstream tasks (Von Kügelgen et al., 2021).

In computer vision tasks, by changing the styles in images, we add more variety to the training data. This helps the model not to focus too much on the specific styles it sees in the training images. It teaches the model to recognize objects or features in images, no matter how the style of the image changes. This is important because in the real world, images can come in many different styles. So, adding style changes in training helps the model perform well on all kinds of images

In the context of ordinal regression, style information is referred to as non-ordinal information, governed by underlying non-ordinal factors. Our method also aims to enrich style diversity by altering non-ordinal information. This is achieved by randomly sampling non-ordinal factors while maintaining the ordinal content factors, then generating examples based on these factors. By preserving the ordinal content factors, we can change the image's style while keeping its ordinal content unchanged, thereby generating synthetic (counterfactual) images not seen in the training data. This approach enables neural networks to access more samples with diverse styles, thereby improving the generalization capabilities of ordinal regression models for unseen samples. As demonstrated in Figure 7, by randomly sampling non-ordinal factors while maintaining the ordinal content factors, we can alter various aspects of the image's style, such as people's dressing, background, and camera angles, etc., ensuring the ordinal content remains unchanged.

It is also important to emphasize two major advantages of our data augmentation methods: Firstly, existing image augmentation strategies do not guarantee the preservation of ordinal information during the augmentation process. For example, color jittering can change an image's color, potentially altering white hair to yellow, which could obscure the age of the person in the image. Secondly, our proposed data augmentation method is general. Our approach can be broadly applied to automatically infer ordinal content from other types of information and generate new examples with guarantees. While primarily tested on image data, our method's framework should be adaptable to non-image data. This adaptability is not achievable with traditional data augmentation methods, which mainly focus on image data. For instance, applying rotations to non-image data is not feasible.

A4 Visualisation of Data Augmentation

We provide additional visualizations of the augmentation results. As shown in Figure A.3, we conducted experiments on three ordinal regression datasets. Our findings indicate that our augmentation method effectively retains age-related and weather-related features, while simultaneously introducing significant stylistic variations. In the case of diabetic retinopathy

instances, the differences between the original and augmented views are subtle. Without domain-specific knowledge, it is challenging to conclusively determine whether the ordinal content has been preserved. Figure A.4 demonstrates the generative results for altering the ordinal factors. For each instance, we fix the non-ordinal factors and replace the ordinal factors with age-specific ordinal factors (i.e., representing different age groups). The age-specific ordinal factor is extracted from training images of the corresponding age group. By visualizing the results, we can observe that the age of the individuals has changed following augmentation, while the styling information from non-ordinal factors remains similar. This effectively illustrates the efficacy of our method in disentangling ordinal and non-ordinal factors. We also present image generation results from a conventional GAN (Karras et al., 2020) in Figure A.5. It is important to note an advantage of our method over conventional GANs: the inability of conventional GANs to disentangle ordinal factors from non-ordinal factors. This means they cannot guarantee the preservation of an image's semantic information. Additionally, our model focuses more on fine-grained details when constructing novel samples. This is evident in the detailed modeling of age-related components in facial images. While conventional GANs can achieve high generative quality, they sometimes fail to accurately represent certain age-related features in the images, such as generate hair for infants and child face for seniors.

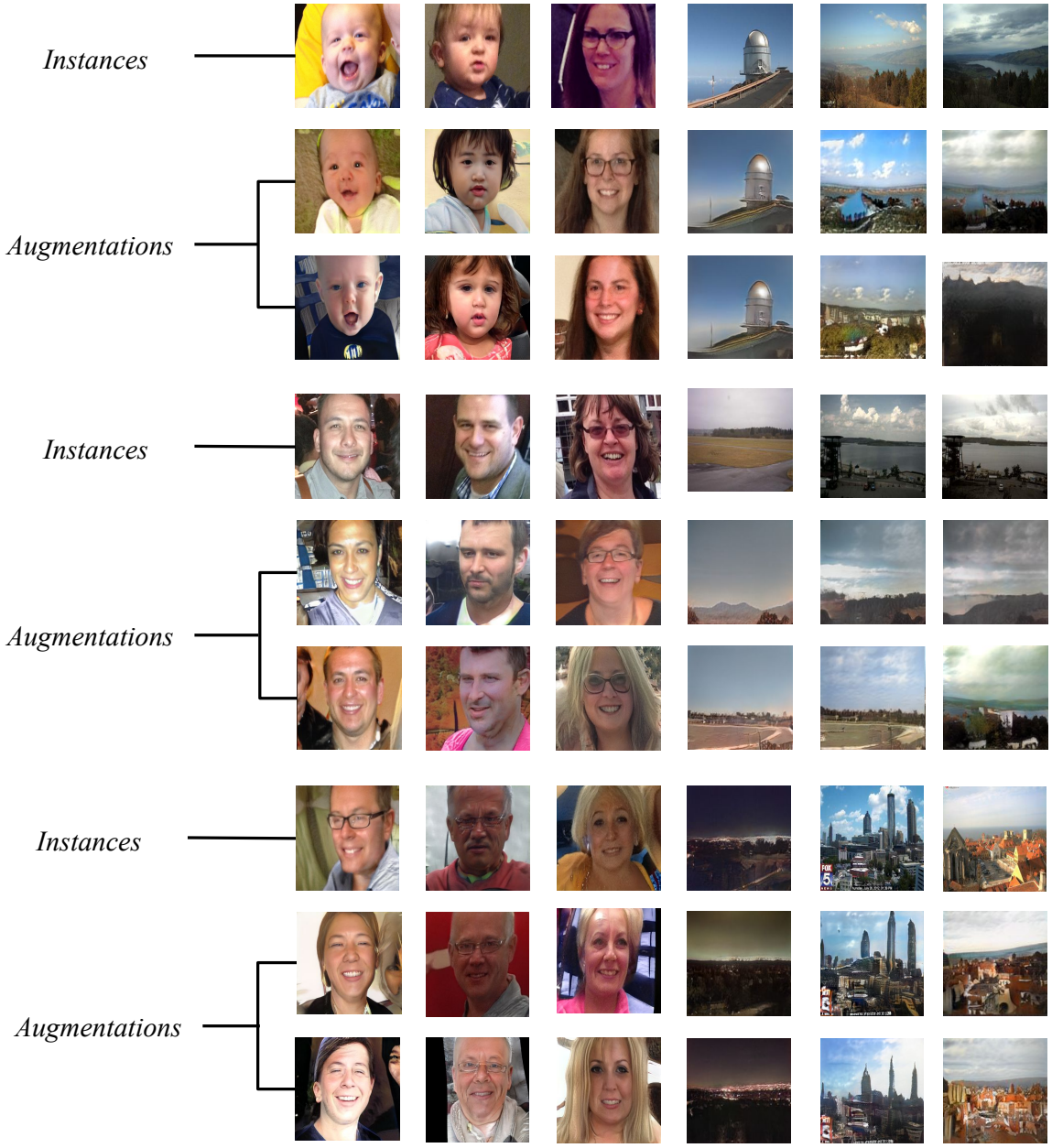

FIGURE A.3. Generated augmentations by augmenting the non-ordinal factors $\hat{z}_n$.

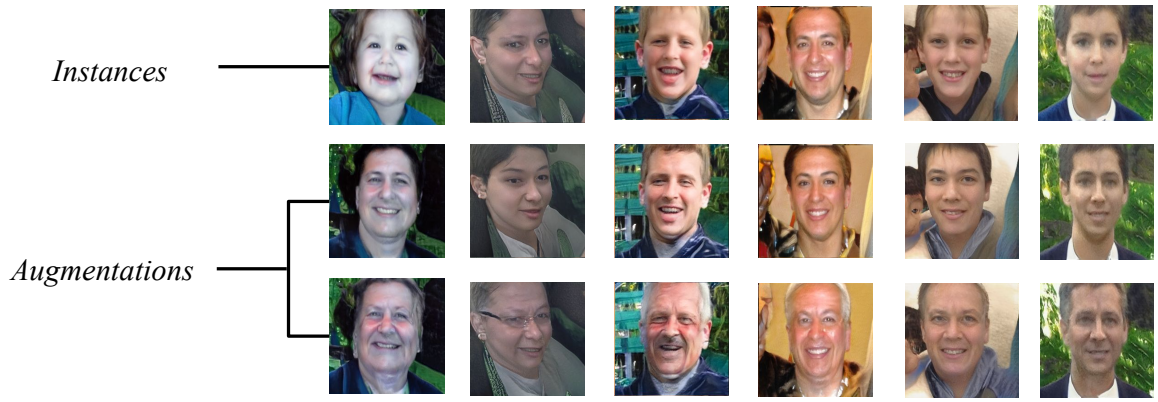

FIGURE A.4. Generated augmentations by augmenting the ordinal factors $\hat{z}_o$ with age-specific factors.

FIGURE A.5. Unconditional image generation results of conventional GAN (the first Row) and Our method (the second Row).

APPENDIX B

Appendix of Chapter 3

B1 Analysis of Object Detection and Image Inpainting

Method	Object Detection	Image Inpainting	
	L-IoU (%↑)	MSE (%↓)	LPIPS (%↓)
Random	17.19 ± 0.6	0.91 ± 0.04	0.64 ± 0.01
SupPR (Zhang et al., 2023c)	19.65 ± 2.9	0.87 ± 0.06	0.55 ± 0.07
CoF Prompting (Ours)	**19.74**± 0.8	**0.61** ± 0.01	**0.47** ± 0.02

TABLE B.1. Object Detection and Image Inpainting Results of CoF Prompting on LLaMA-7B.

Object Detection Table B.1 presents the quantitative performance of our CoF method compared to baselines. CoF achieve increment over the random on the L-IoU by 14.8%. While the quantitative results of SupPR and CoF are very similar in this tasks, with CoF slightly higher in the metric. However, by observing the qualitative results in Figure 3.3, we can still observe the difference in between the two methods, where CoF are more accurate in locating the boxs and reconstruct the original input. We additional calcuate the failure rate, where the predicted bounding boxs are completely disjoint to the ground truth boxs. Failure cases for detection are detailed in Table B.2, where CoF reduces failures by 11.9% on LLaMA-7B for object detection.

Image Inpainting As shown in Figure 3.3, the overall qualitative performance of LAVM on inpainting task is exceptional. However, they still benefit from proper prompting. By applying CoF prompting, the generated patches are more natural and of higher quality compared to the baselines. Table B.1 (Right) shows that our CoF method quantitatively outperforms the

Model	Random	CoF
LLaMA-7B w/ VQ-GAN	57.49 ± 3.7	**51.63** ± 0.8

TABLE B.2. Failure Rates (↓) - Object Detection

baselines, achieving a 4.3% and 1.7% improvement in MSE and a improvement in LPIPS over the second-highest prompting method.

B2 Thresholding Performance Analysis

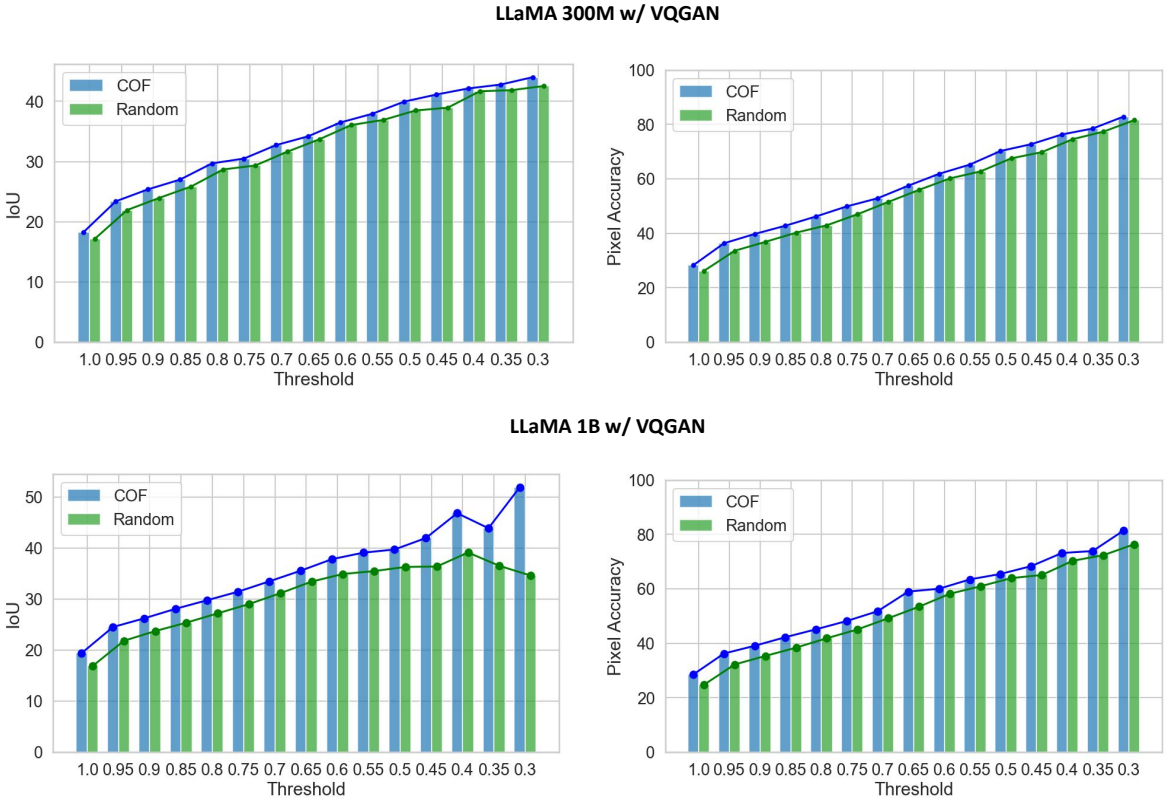

FIGURE B.1. Image Segmentation and Pose Estimation Results for various black rate thresholding. Our method consistently outperforms the baselines on different pre-trained models across various threshold rates, demonstrating the stable performance of CoF prompting.

In this section, we analyze segmentation performance by thresholding the black rate of the prediction. The black rate represents the proportion of the black area in the predicted results. We assess the performance of COF prompting at different sizes of the predictable object to

ensure its contribution is stable to the LAVMs. Figure B.1 demonstrates the performance comparison between COF prompting and the Random baseline across two metrics. Across varying thresholds, COF consistently outperforms the Random baseline. This showcases that COF prompting maintains robust performance in enhancing the predictive capabilities of LAVMs regardless of the size of the predictable object.

B3 Visualisation of Results of LAVM w/ LLaMA-1B

Here we present qualitative results of image segmentation and pose estimation using LAVM with LLaMA-1B. As demonstrated in Figure B.2, using CoF prompting significantly improves the accuracy of object mask identification for image segmentation. Similarly, the quality of the estimated skeletons is also better when applying CoF prompting.

B4 Reversing Order of intermediate Reasoning Steps

Prompting Method	LLaMA-7B			
	Image Segmentation		Pose Estimation	
	IoU (%↑)	P-ACC (%↑)	IoU (%↑)	P-ACC (%↑)
COF	52.53	67.05	2.80	13.34
COF-reversed	49.65	65.37	2.71	10.52

TABLE B.3. Reversed Intermediate Reasoning Steps with the LAVM w/ LLaMA-7B

The core of CoF prompting is to generate a series of intermediate reasoning steps for sequentially prompting the LAVMs. The reasoning path is created based on the saliency paths we identify within individual images. Here, we explore whether the order of reasoning steps will affect the in-context learning of LAVMs. To this end, we present a qualitative comparison in Figure B.3 and a quantitative comparison in Table B.3. It is observed that reversing the order of the intermediate steps can impact the in-context predictions of LAVMs; however, compared to the predictions in Figure 3.3, we can conclude that reverse sequential prompting is still better than directly showing the LAVMs the full target.

FIGURE B.2. Results on LLaMA-1B Model. The first and fourth rows are the original test inputs for image segmentation and pose estimation, respectively. Orange boxes show the predictions given random prompts. Blue boxes show the predictions using Chain-of-Focus prompting.

B5 Dependency on Saliency Detectors

Measuring saliency (Huang et al., 2023b) is an important step in our method. Here, we assess the sensitivity of our prompting method to variations in saliency detectors, we further employed a different approach: GradCAM (Selvaraju et al., 2017) to compute saliency scores. Figure B.4 demonstrate the different attention maps visualized from GradCAM and U2-Net, respectively. Table B.4 shows the results for the LLaMA-7B LAVM on the four tasks, comparing U2-Net and GradCAM. Notably, switching the method for measuring saliency scores does not result in significant differences in performance. Based on the observation, we

FIGURE B.3. Qualitative Results of reversing intermediate reasoning steps with the LAVM w/ LLaMA-7B. The second row shows the CoF prompting output. The third row show the results of using the same prompt, but reversing the order in intermediate steps.

conclude that both approaches effectively detect salient regions, and the consistent in-context learning performance further highlights the robustness of our approach.

Prompting Method\Task	**Segmentation**	**Pose Estimation**	**Object Detection**	**Image Inpainting**
	IoU (↑)/ P-ACC (↑)	IoU (↑)/ P-ACC (↑)	L-IoU (↑)	MSE (↓) / LPIPS (↓)
CoF Prompting w/ GradCAM	52.68 / 67.14	2.77 / 13.16	19.77	0.63 / 0.51
CoF Prompting w/ U2-Net	52.53 / 67.05	2.80 / 13.34	19.74	0.61 / 0.47

TABLE B.4. Comparison of CoF Prompting with different saliency detectors.

B6 Additional Qualitative Results

In this section, we present additional visualizations of the image segmentation and pose estimation tasks to illustrate the in-context learning performance of COF prompting. For image segmentation, the results for the LLaMA-300M model are depicted in Figure B.6, while Figure B.7 showcases the outcomes for the LLaMA-1B model. For pose estimation, we provide visual evidence of the improved performance facilitated by COF prompting using LLaMA-300M (Figure B.8) and LLaMA-1B (Figure B.9). These results demonstrate the efficacy of our method in enhancing LAVMs' predictive ability on both tasks through structured, reasoning-based prompting. Based on these additional visualizations, as well as the results shown in Figure 3.3, we can observe that larger models have strong predictive

FIGURE B.4. Visual Attention of different saliency detectors. Green boxes show the results given by GradCAM. Yellow boxes show the results given by U2-net.

power on both tasks. This implies that, in order to achieve predictive capabilities similar to expert models, we will need to scale up the parameter size of the models as well as the size of the training data.

B7 Limitations and Future Directions

While visual prompting methods can potentially enhance the predictive performance of Large Vision Models, their limitations are constrained by the capacity of these models. Practical usage of LAVMs requires stronger and more robust pretrained models, along with the advancement of in-context learning methods. Instances where current LAVMs produce pure black predictions highlight their fundamental instability, raising concerns about their trustworthiness in real-world deployment.

We observed that the failure cases are primarily associated with two factors: model scale and prompt selection. Model scale is the major factor, as LAVMs with larger parameter sizes tend to exhibit fewer failure cases. Failure cases can also arise from the choice of

visual prompts, and our experiments demonstrate that the proposed prompt selection module effectively reduces the number of failures.

To further investigate, we identified prompts that previously caused failures in in-context predictions and were not encountered by the model during pre-training. We then switched the test input while using the same failure-inducing prompts, and the failure persisted across different test inputs. However, when we replaced the prompts with training samples from datasets used in the pre-training process, the success rate significantly increased. Based on these observations, we hypothesize that the root cause of failure cases is related to the model's out-of-distribution generalization ability with respect to the prompts. The model may fail to perform in-context learning if the prompts are unseen during pre-training or exhibit a domain gap.

The community as a whole desires a unified solution for all vision tasks. Therefore, the authors advocate for continued research into building robust large vision models.

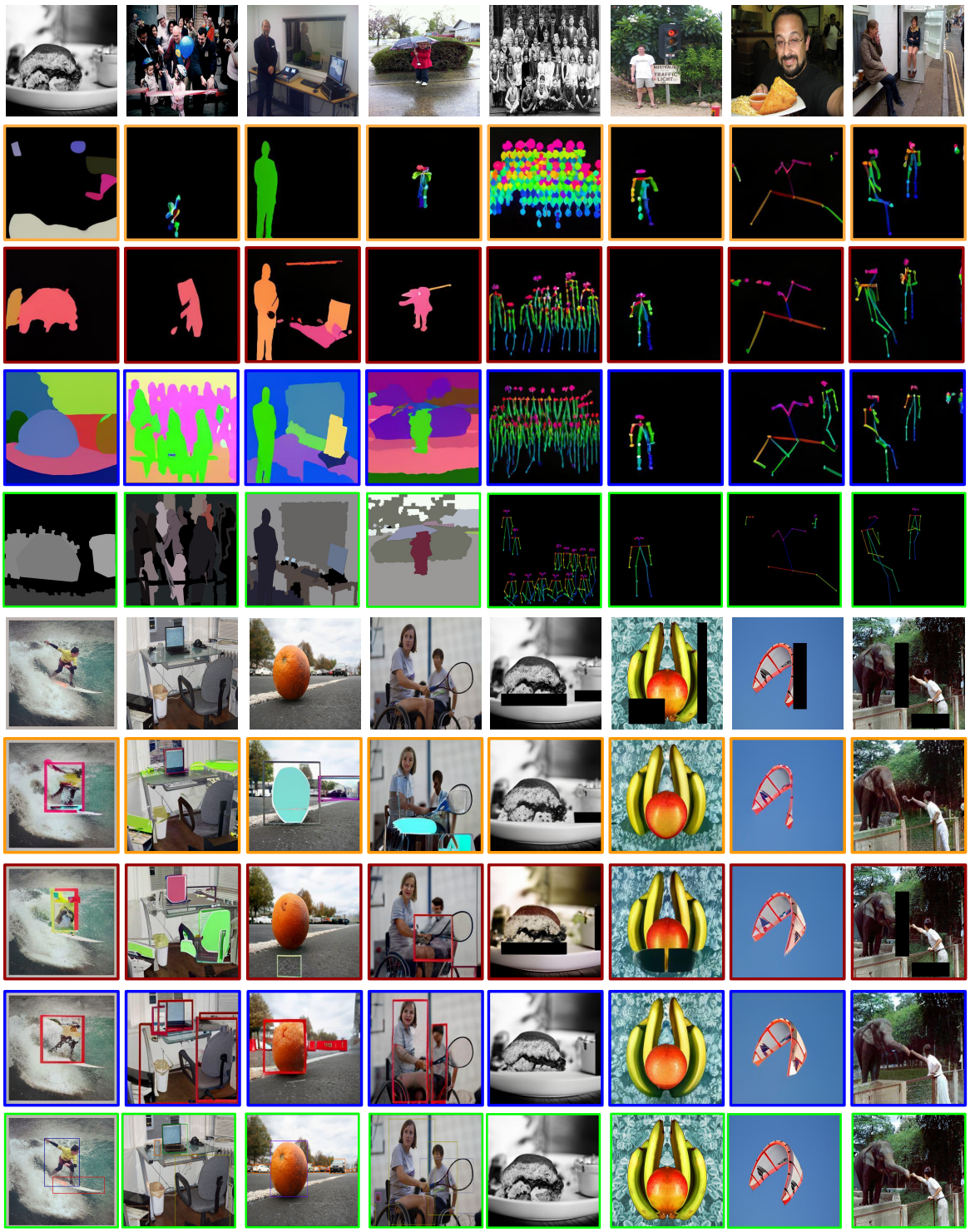

FIGURE B.5. Ground Truth Visualization for the test input.

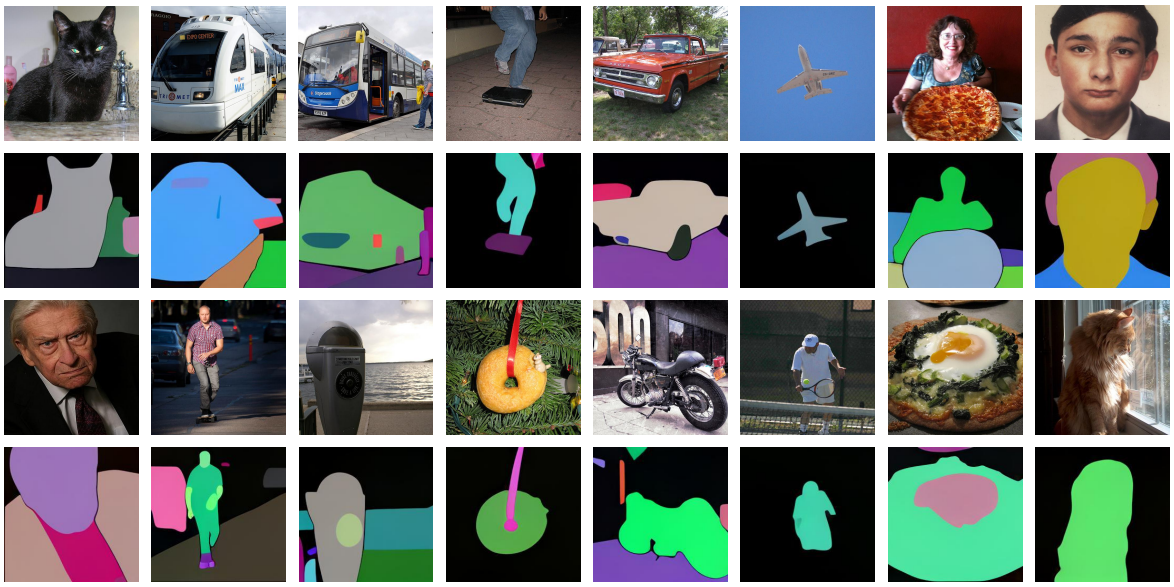

FIGURE B.6. Image Segmentation Results from LLaMA-300M w/ VQ-GAN using COF prompting.

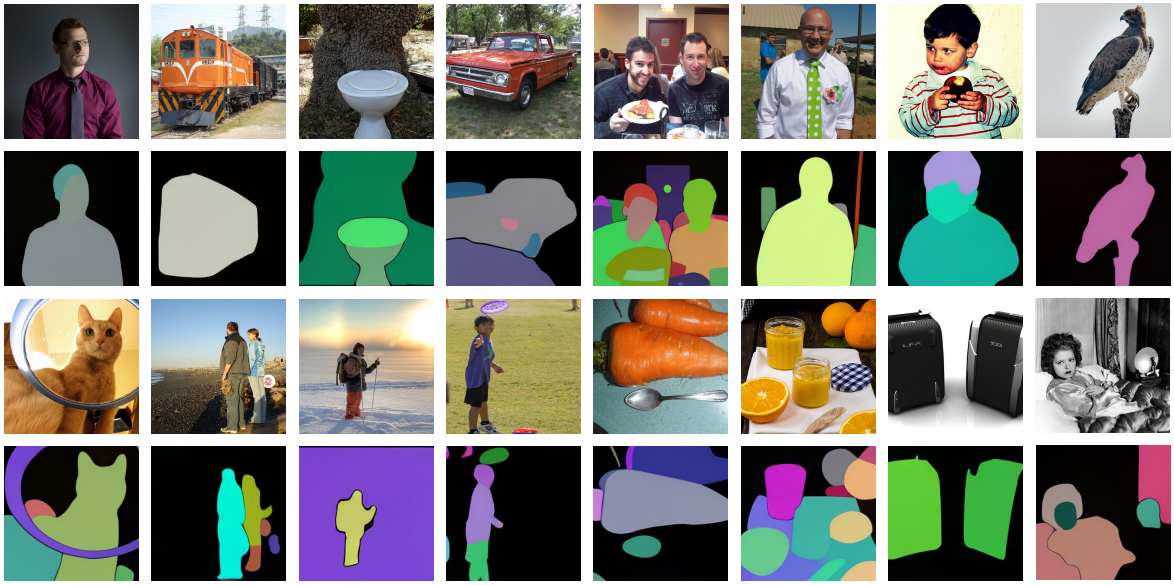

FIGURE B.7. Image Segmentation Results from LLaMA-1B w/ VQ-GAN using CoF prompting.

FIGURE B.8. Pose Estimation Results from LLaMA-300M w/ VQ-GAN using COF prompting.

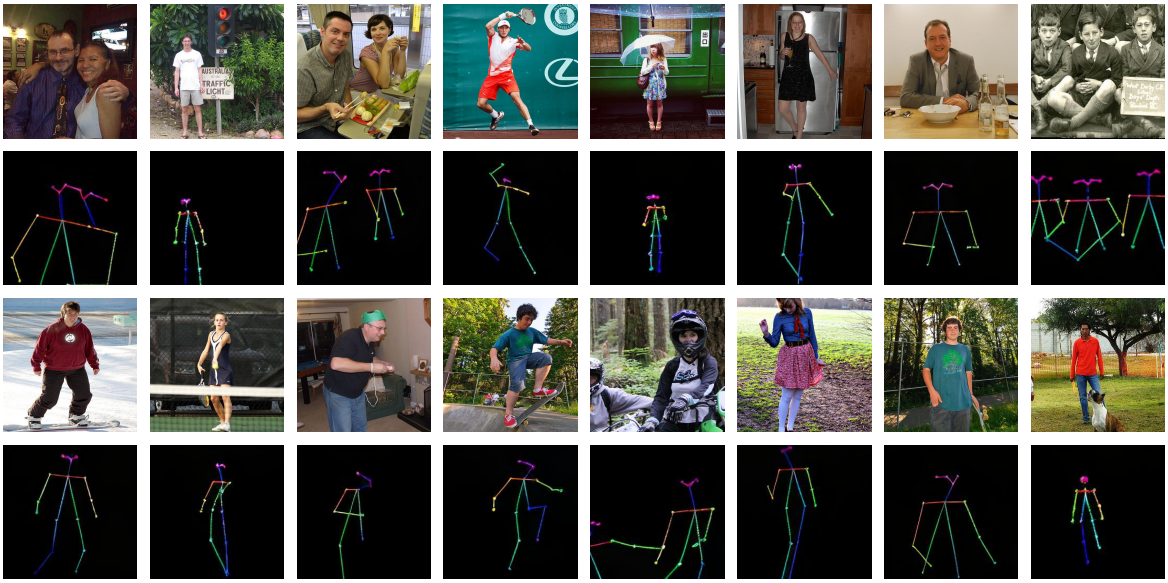

FIGURE B.9. Pose Estimation Results from LLaMA-1B w/ VQ-GAN using CoF prompting.

APPENDIX C

Appendix of Chapter 4

C1 Theoretical Results and Proofs

C1.1 Notations and Definitions

We introduce key notations and definitions as follows.

Let $X_V \in \mathcal{X}_V$ (video) and $X_A \in \mathcal{X}_A$ (audio) be two modalities. Pretrained encoders produce

$$\hat{Z}_V = (\hat{Z}_{V,1}, \ldots, \hat{Z}_{V,d}), \quad \hat{Z}_A = (\hat{Z}_{A,1}, \ldots, \hat{Z}_{A,d}) \in \mathbb{R}^d. \tag{C.1}$$

We introduce learnable invertible reparameterizations

$$q_V, q_A : \mathbb{R}^d \to \mathbb{R}^d, \tag{C.2}$$

and define

$$\tilde{Z}_V = q_V(\hat{Z}_V), \quad \tilde{Z}_A = q_A(\hat{Z}_A). \tag{C.3}$$

We learn *mask functions*

$$M_V : \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}^d, \quad M_A : \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}^d, \tag{C.4}$$

so that for each sample the model produces binary masks

$$M_V(\tilde{Z}_V, \tilde{Z}_A), \; M_A(\tilde{Z}_V, \tilde{Z}_A) \in \{0,1\}^d, \tag{C.5}$$

Denote their supports (or selected indices) by

$$\tilde{S}_V := S_V(\tilde{Z}_V, \tilde{Z}_A) = \{\, i : M_{V,i}(\tilde{Z}_V, \tilde{Z}_A) = 1\}, \tag{C.6}$$

$$\tilde{S}_A := S_A(\tilde{Z}_V, \tilde{Z}_A) = \{\, i : M_{A,i}(\tilde{Z}_V, \tilde{Z}_A) = 1\}. \tag{C.7}$$

Let $\tilde{Z}_V^{\tilde{S}_V} \subseteq \tilde{Z}_V$ and $\tilde{Z}_A^{\tilde{S}_A} \subseteq \tilde{Z}_A$ be the selected variables induced by learnable masks, i.e.,

$$\tilde{Z}_V^{\tilde{S}_V} = (\tilde{Z}_{V,i})_{i\in\tilde{S}_V}, \quad \tilde{Z}_A^{\tilde{S}_A} = (\tilde{Z}_{A,i})_{i\in\tilde{S}_A}, \quad \overline{\tilde{S}_V} = [d] \setminus \tilde{S}_V \text{ and } \overline{\tilde{S}_A} = [d] \setminus \tilde{S}_A, \tag{C.8}$$

for any set $\tilde{S}_V \subseteq [d]$ and set $\tilde{S}_A \subseteq [d]$.

Write $S_V^\dagger \subseteq [d]$ (resp. $S_A^\dagger$) for the *true* minimal Markov blanket of X_A in Z_V (resp. of X_V in Z_A). In other words, $S_V^\dagger$ is the smallest subset satisfying that conditioning on $\{Z_{V,i} : i \in S_V^\dagger\}$ renders all other latent coordinates irrelevant to X_A. The analogous property holds for $S_A^\dagger$.

Reconstruction quality is measured via true conditional entropy:

$$H(X_A \mid \tilde{Z}_V^{\tilde{S}_V}) = -\mathbb{E}\big[\log p(X_A \mid \tilde{Z}_V^{\tilde{S}_V})\big], \quad H(X_V \mid \tilde{Z}_A^{\tilde{S}_A}) = -\mathbb{E}\big[\log p(X_V \mid \tilde{Z}_A^{\tilde{S}_A})\big]. \tag{C.9}$$

Let $Q_{g_A}(X_A \mid \cdot)$ and $Q_{g_V}(X_V \mid \cdot)$ be decoder families. We optimize

$$\mathcal{L}_V(M_V, q_V, g_A) = \mathbb{E}\Big[-\log Q_{g_A}\big(X_A \mid M_V(\tilde{Z}_V, \tilde{Z}_A) \odot \tilde{Z}_V\big)\Big] + \lambda\, \mathbb{E}\big[\|M_V(\tilde{Z}_V, \tilde{Z}_A)\|_1\big], \tag{C.10}$$

and symmetrically $\mathcal{L}_A(g_V, q_V, M_A)$ for audio→video.

DEFINITION 2 (Minimum Sufficient Latents). *Given index sets $\tilde{S}_V, \tilde{S}_A \subseteq [d]$, we say that the pairs $\big(\tilde{Z}_V^{\tilde{S}_V}, \tilde{Z}_A^{\tilde{S}_A}\big)$ are* Minimum Sufficient Latents *if they satisfy*

$$I\big(\tilde{Z}_V^{\tilde{S}_V}; X_A\big) = I\big(Z_V^{S_V^\dagger}; X_A\big),\ I\big(\tilde{Z}_{V_j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V}\big) = 0 \ \ \forall j \notin \tilde{S}_V,$$

$$I\big(\tilde{Z}_A^{\tilde{S}_A}; X_V\big) = I\big(Z_A^{S_A^\dagger}; X_V\big),\ I\big(\tilde{Z}_{A_j}; X_V \mid \tilde{Z}_A^{\tilde{S}_A}\big) = 0 \ \ \forall j \notin \tilde{S}_A.$$

C1.2 Assumptions

We introduce the assumptions required by our method as follows.

ASSUMPTION 1 (DAG & d-Separation). *The joint distribution of* (Z_V, Z_A, X_V, X_A) *factors according to a DAG satisfying the global Markov property and faithfulness.*

Hence for any $S_V \subseteq [d]$*,*

$$X_A \perp Z_V^{\overline{S}_V} \mid Z_V^{S_V} \quad \Longleftrightarrow \quad I\big(Z_{V,i}; X_A \mid Z_V^{S_V}\big) = 0 \; \forall i \notin S_V, \tag{C.11}$$

and the same condition holds when V *and* A *are interchanged.*

ASSUMPTION 2 (Block-wise Reparameterization). *The joint function class for* (q_V, q_A) *is rich enough that there exist invertible maps*

$$q_V^*, q_A^* : \mathbb{R}^d \to \mathbb{R}^d \tag{C.12}$$

and index-sets $\tilde{S}_V^\dagger, \tilde{S}_A^\dagger \subseteq [d]$ *satisfying*

$$I\big(q_V^*(Z_V)^{\tilde{S}_V^\dagger}; X_A\big) = I\big(Z_V^{S_V^\dagger}; X_A\big), \quad I\big(q_V^*(Z_V)_j; X_A \mid q_V^*(Z_V)^{\tilde{S}_V^\dagger}\big) = 0 \;\; \forall j \notin \tilde{S}_V^\dagger. \tag{C.13}$$

ASSUMPTION 3 (Decoder Universality). *For any* $S \subseteq [d]$*,* $\min_{g_A} \mathbb{E}[-\log Q_{g_A}(X_A \mid \tilde{Z}_V^{\tilde{S}_V})] \to H(X_A \mid \tilde{Z}_V^{\tilde{S}_V})$ *and similarly for* $X_V \mid \tilde{Z}_A^S$*.*

ASSUMPTION 4 (Mask Universality). *The mask networks* M_V, M_A *are sufficiently expressive to realize any mapping* $\mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}^d$*, i.e. choose any support* $S \subseteq [d]$ *for each sample.*

ASSUMPTION 5 (Penalty-Range). *For any subset* $\tilde{S}_V \subseteq [d]$ *and index* $i \in [d]$*, define*

$$\Delta_{V,i}(\tilde{S}_V) = I\big(\tilde{Z}_{V,i}; X_A \mid \tilde{Z}_V^{(\tilde{S}_V \setminus \{i\})}\big),$$

which is the mutual information between $\tilde{Z}_{V,i}$ *and* X_A *conditioned on the remaining variables* $\tilde{Z}_V^{(\tilde{S}_V \setminus \{i\})} = \{\tilde{Z}_{V,j} : j \in \tilde{S}_V \setminus \{i\}\}$*. There exists a constant* λ *such that*

$$\max_{j \notin \tilde{S}_V^\dagger} \Delta_{V,j}([d]) \;<\; \lambda \;<\; \min_{i \in \tilde{S}_V^\dagger} \Delta_{V,i}(\tilde{S}_V^\dagger),$$

and the same condition holds when V *and* A *are interchanged.*

Note that above assumptions are common. Assumption 1 is a fundamental assumption in causality (Peters et al., 2017). Assumption 2 merely requires that our networks q_V, q_A

have sufficient capacity to "whiten" or disentangle the small block of truly shared latents. Assumption 3 assumes that deep decoders can approximate any conditional density arbitrarily well, so cross-entropy minimization recovers true conditional entropy. Assumption 4 implies stipulate that our mask networks are expressive enough to pick any subset of coordinates per example. All these assumptions 2, 3, 4 have been supported by universal approximation theory of deep learning methods (Huang et al., 2024a). Finally, Assumptions 5 implies that the sparsity weight λ can be chosen to lie between the minimal utility of a shared factor and the maximal spurious contribution of a non-shared factor. In practice, we can just make λ be sufficiently small.

C1.3 Theoretical Results

The following lemma shows that, by minimizing the cross-entropy loss of a decoder trained to reconstruct a short audio segments from the selected learned representations of video frames, one asymptotically recovers the conditional entropy of reconstructed short audio segments given those representations. The same result holds when swapping the roles of video V and the audio A.

LEMMA 3 (Cross-Entropy Reduction to Conditional Entropy). *Under Assumption 3, for any fixed mask function M_V and fixed q_V, we have*

$$\min_{g_A} \mathbb{E}\Big[-\log Q_{g_A}\big(X_A \mid M_V(\tilde{Z}_V, \tilde{Z}_A) \odot \tilde{Z}_V\big)\Big] \longrightarrow \mathbb{E}\Big[H\big(X_A \mid \tilde{Z}_V^{\tilde{S}_V}\big)\Big], \tag{C.14}$$

where $\tilde{Z}_V = q_V(Z_V)$ and $S_V(\tilde{z}_V, \tilde{z}_A) = \textit{support}\big(M_V(\tilde{z}_V, \tilde{z}_A)\big)$.

PROOF. Let

$$L(M_V, g_A) \;=\; \mathbb{E}\Big[-\log Q_{g_A}\big(X_A \mid M_V(\tilde{Z}_V, \tilde{Z}_A) \odot \tilde{Z}_V\big)\Big]. \tag{C.15}$$

By the interchange of minima,

$$\min_{M_V, g_A} L(M_V, g_A) = \min_{M_V}\Big[\min_{g_A} L(M_V, g_A)\Big]. \tag{C.16}$$

Fix any mask M_V. Then by Assumption 3 (Decoder Universality),

$$\min_{g_A} L(M_V, g_A) \;=\; \min_{g_A} \mathbb{E}\big[-\log Q_{g_A}(X_A \mid \tilde{Z}_V^{\tilde{S}_V})\big] \;\longrightarrow\; \mathbb{E}\big[\, H(X_A \mid \tilde{Z}_V^{\tilde{S}_V})\big]. \tag{C.17}$$

Note that $(X_A \mid M_V(\tilde{Z}_V, \tilde{Z}_A) \odot \tilde{Z}_V = Z_V^{\tilde{S}_V}$ by definition, since the mask selects exactly those components. Therefore, the proof is complete. □

The following lemma shows that any mask–decoder pair minimizing the cross-entropy reconstruction loss inevitably selects a subset of video representations that retains the full mutual information with the audio segment, i.e., it forms a sufficient statistic for the audio segment. The same result holds when swapping the roles of video V and the audio A.

LEMMA 4 (Sufficientness of Reconstruction). *Fix any invertible q_V. Under Assumptions 1–4, any mask–decoder pair (M_V, g_A) that minimizes $\mathbb{E}[-\log Q_{g_A}(X_A \mid M_V \odot \tilde{Z}_V)]$ must satisfy, for every sample,*

$$I\big(\tilde{Z}_V^{\tilde{S}_V}; X_A\big) \;=\; I\big(\tilde{Z}_V; X_A\big). \tag{C.18}$$

In other words, the selected coordinates form a sufficient statistic for X_A.

PROOF. For notational simplicity, we omit the arguments $(\tilde{Z}_V, \tilde{Z}_A)$ when writing $\hat{M}_V(\tilde{Z}_V, \tilde{Z}_A)$. Fix q_V and consider any minimizer $(\hat{M}_V, \hat{g}_A)$ of the cross-entropy. By Lemma 3, this pair also minimizes $\mathbb{E}[\, H(X_A \mid \tilde{Z}_V^{\tilde{S}_V})]$. Under Assumption 4, the mask M_V is expressive enough to choose the index set $\tilde{S}_V$ arbitrarily for each sample. Hence the minimization decomposes per sample: for each $(\tilde{z}_V, \tilde{z}_A)$, we pick

$$S_V(\tilde{z}_V, \tilde{z}_A) \in \arg\min_{\tilde{S} \subseteq [d]} H\big(X_A \mid \tilde{Z}_V^{\tilde{S}_V} = \tilde{z}_V^{\tilde{S}_V}\big). \tag{C.19}$$

Recall that for any fixed $\tilde{S}_v$,

$$H(X_A \mid \tilde{Z}_V^{\tilde{S}_V}) \;=\; H(X_A) - I(\tilde{Z}_V^{\tilde{S}_V}; X_A), \tag{C.20}$$

By the causal faithfulness and causal Markov properties (Peters et al., 2017) (analogous to Assumption 1, but applied to the variables produced by the neural network), it gives

$$I(\tilde{Z}_V^{\tilde{S}_V}; X_A) \le I(\tilde{Z}_V; X_A) \quad\Longleftrightarrow\quad H(X_A \mid \tilde{Z}_V^{\tilde{S}_V}) \ge H(X_A \mid \tilde{Z}_V). \tag{C.21}$$

Hence the unique minimizer of $H(X_A \mid \tilde{Z}_V^{\tilde{S}_V})$ is any S satisfying

$$H(X_A \mid \tilde{Z}_V^{\tilde{S}_V}) = H(X_A \mid \tilde{Z}_V), \tag{C.22}$$

which is equivalent to

$$I(\tilde{Z}_V^{\tilde{S}_V}; X_A) = I(\tilde{Z}_V; X_A). \tag{C.23}$$

Thus for each sample, $I(\tilde{Z}_V^{S_V(\tilde{z}_V, \tilde{z}_A)}; X_A) = I(\tilde{Z}_V; X_A)$, completing the proof. □

The following lemma shows that adding an ℓ_1-penalty on the mask encourages sparsity: the optimal mask discards all non-shared coordinates and exactly recovers the minimal shared block of the variables produced by the neural network.

LEMMA 5 (Sparsity-Induced Minimality). *Fix any invertible q_V. Under Assumptions 3–5, the joint minimizer*

$$(M_V^*, g_A^*) = \arg\min_{M_V, g_A} \Big\{ \mathbb{E}\big[-\log Q_{g_A}(X_A \mid M_V \odot \tilde{Z}_V)\big] \; + \; \lambda\, \mathbb{E}[\|M_V\|_1] \Big\} \tag{C.24}$$

satisfies, for almost every sample,

$$S_V^*(\tilde{Z}_V, \tilde{Z}_A) = \tilde{S}_V^\dagger, \quad I\big(\tilde{Z}_{V,j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V^\dagger}\big) = 0 \quad \forall j \notin \tilde{S}_V^\dagger. \tag{C.25}$$

That is, the mask prunes away all non-shared coordinates, recovering exactly the minimal shared block.

PROOF OF LEMMA 5 (SPARSITY-INDUCED MINIMALITY). Fix q_V. As before, by Decoder Universality (Lemma 3) the joint minimization over (M_V, g_A) is equivalent to

$$\min_{M_V} \; \mathbb{E}\Big[\, H\big(X_A \mid \tilde{Z}_V^{\tilde{S}_V}\big) + \lambda\, |\tilde{S}_V| \Big]. \tag{C.26}$$

Since M_V can choose $\tilde{S}_V$ per sample (Assumption 4), we solve for each $(\tilde{z}_V, \tilde{z}_A)$:

$$\min_{\tilde{S} \subseteq [d]} \; f(\tilde{S}) \quad \text{where} \quad f(\tilde{S}) = H\big(X_A \mid \tilde{z}_V^{\tilde{S}_V}\big) + \lambda\, |\tilde{S}|. \tag{C.27}$$

For any $j \notin \tilde{S}$, adding j changes f by

$$f(\tilde{S} \cup \{j\}) - f(\tilde{S}) = -I\big(\tilde{Z}_{V,j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V}\big) + \lambda. \tag{C.28}$$

By Assumption 5, $I(\tilde{Z}_{V,j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V}) \leq \Delta_{V,j}([d]) < \lambda$, so $f(S \cup \{j\}) > f(S)$ and no non-blanket index is added. Similarly, for any $i \in S$, dropping i changes f by

$$f(\tilde{S} \setminus \{i\}) - f(\tilde{S}) = I\big(\tilde{Z}_{V,i}; X_A \mid \tilde{Z}_V^{\tilde{S}\setminus\{i\}}\big) - \lambda, \tag{C.29}$$

and Assumption 5 ensures this is positive for all $i \in \tilde{S}_V^\dagger$. Hence the unique minimizer is $\tilde{S} = \tilde{S}_V^\dagger$, and $I(\tilde{Z}_{V,j}; X_A \mid \tilde{Z}_V^{\tilde{S}_V^\dagger}) = 0$ for $j \notin \tilde{S}_V^\dagger$, as required. □

The following theorem shows that, when jointly optimizing encoders, masks, and decoders with our bidirectional objective over both video and audio representations, the global minimizer precisely aligns and recovers the shared latent blocks—i.e. it achieves exactly the block-alignment specified in Definition 2.

THEOREM 2 (Global Block-Alignment and Recovery). *Under Assumptions 1, 2, 3, 4 and 5, the global minimizer of Objective 7 yields* $\big(\tilde{Z}_V^{\tilde{S}_V^*}, \tilde{Z}_A^{\tilde{S}_A^*}\big)$ *that satisfies Definition 2.*

PROOF. We decompose the total training objective into two symmetric parts,

$$\mathcal{L} = \mathcal{L}_{V\to A}(q_V, M_V, g_A) + \mathcal{L}_{A\to V}(q_A, M_A, g_V),$$

where, for instance,

$$\mathcal{L}_{V\to A}(q_V, M_V, g_A) = \mathbb{E}\Big[-\log Q_{g_A}\big(X_A \mid M_V(\tilde{Z}_V, \tilde{Z}_A) \odot q_V(Z_V)\big)\Big] + \lambda\, \mathbb{E}\big[\|M_V\|_1\big].$$

1. Existence of an optimal block-aligned configuration. By Assumption 2, there exist $(q_V^\dagger, M_V^\dagger)$ and $g_A^\dagger$ such that

$$M_V^\dagger(\tilde{z}_V, \tilde{z}_A) \equiv \tilde{S}_V^\dagger, \quad \mathcal{L}_{V\to A}(q_V^\dagger, M_V^\dagger, g_A^\dagger) = H\big(X_A \mid \tilde{Z}_V^{\tilde{S}_V^\dagger}\big) + \lambda\, |\tilde{S}_V^\dagger|.$$

The same argument applies to the audio-to-video term, yielding $(q_A^\dagger, M_A^\dagger, g_V^\dagger)$.

2. Optimality of the shared supports. Fix any candidate (q_V, M_V, g_A). First, for a fixed encoder q_V, Lemma 3 (Decoder Universality) shows that

$$\min_{g_A} \mathcal{L}_{V\to A}(q_V, M_V, g_A) = \mathbb{E}\big[H(X_A \mid \tilde{Z}_V^{\tilde{S}_V})\big] + \lambda\, \mathbb{E}\big[|\tilde{S}_V|\big].$$

Lemma 4 then implies any minimizer M_V must satisfy

$$I\big(\tilde{Z}_V^{\tilde{S}_V}; X_A\big) = I\big(\tilde{Z}_V; X_A\big).$$

Next, we allow q_V itself to vary. By Assumption 2, the encoder family contains some $q_V^\dagger$ that minimizes the conditional entropy $\mathbb{E}[H(X_A \mid \tilde{Z}_V^{\tilde{S}_V^\dagger})]$. Lemma 5 ensures the unique sparsest choice of $\tilde{S}_V$ is $\tilde{S}_V^\dagger$. Altogether, when λ is sufficiently small, the global minimizer of the video→audio term satisfy

$$q_V^* = q_V^\dagger, \quad M_V^*(\cdot) = \tilde{S}_V^\dagger.$$

An identical chain of reasoning on $\mathcal{L}_{A\to V}$ yields

$$q_A^* = q_A^\dagger, \quad M_A^*(\cdot) = \tilde{S}_A^\dagger.$$

Therefore, at the global optimum both pairs $\big(\tilde{Z}_V^{\tilde{S}_V^*}, \tilde{Z}_A^{\tilde{S}_A^*}\big)$ thus coincide with the unique minimal sufficient latents $\big(\tilde{Z}_V^{\tilde{S}_V^\dagger}, \tilde{Z}_A^{\tilde{S}_A^\dagger}\big)$, so they satisfy Definition 1. □

C2 Cascade Diffusion Model

Building on the aligned latent representations from Stage I, we generate video and audio in a cascaded manner using independently finetuned single-modal diffusion models. As illustrated in Figure 3 (II), the process begins by generating a video from a text prompt using a video latent diffusion model. The resulting visual latent z_v^T is then adapted through a lightweight projection network $\mathcal{P}_\theta$, producing an audio-guiding latent $\tilde{z}_a$ that encodes visually grounded cues. This latent, along with the original text embedding, conditions the subsequent audio generation. By structuring the process in this cascaded fashion, we ensure that the audio is temporally and semantically aligned with the generated video content. Notably, the pretrained diffusion backbones remain frozen during training, and only the adapters are updated, preserving modularity and enabling efficient adaptation to downstream multimodal generation tasks.

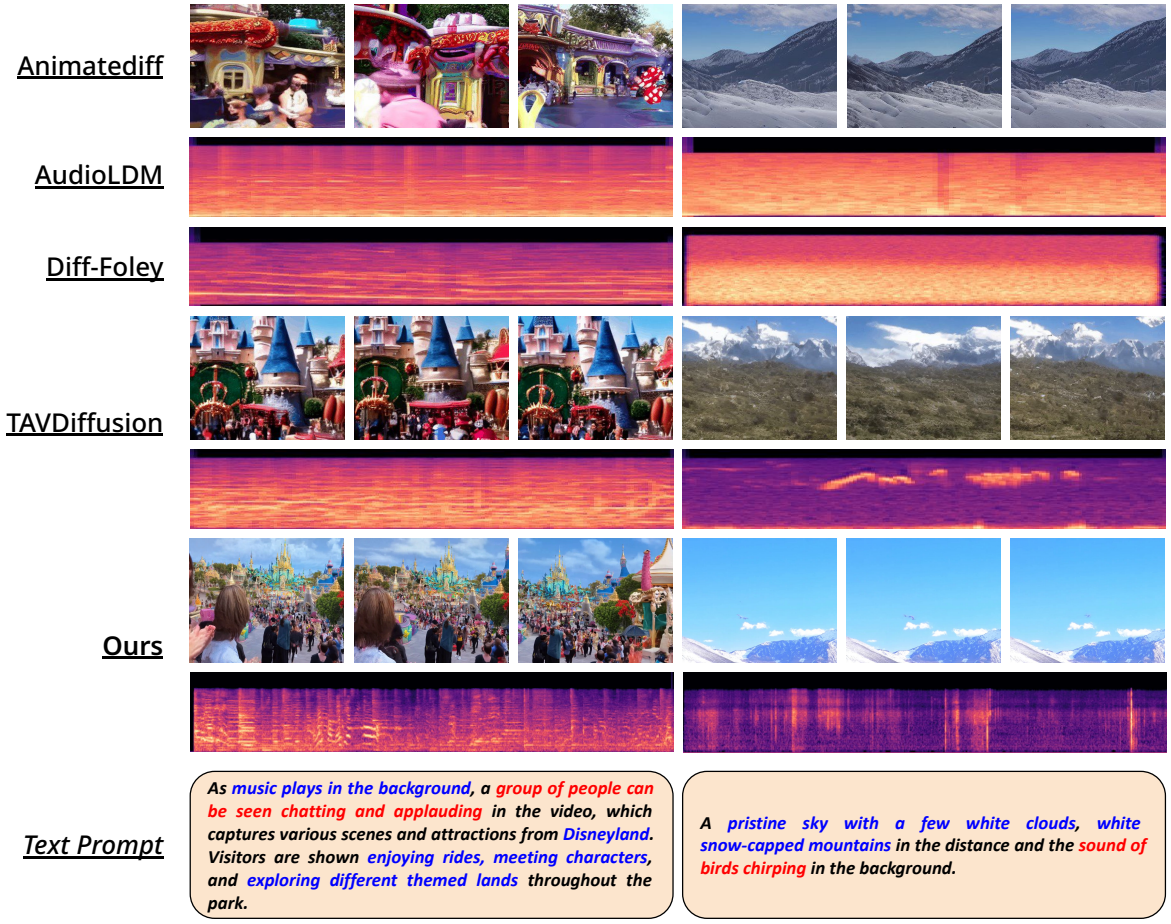

FIGURE C.1. Additional Text-to-Audio-Video generation results compared with other baselines. We use the same text prompt as in (Mao et al., 2024) for our demonstration and compare the method against multiple baselines (Animatediff (Guo et al., 2023), AudioLDM (Liu et al., 2023a), Diff-Foley (Luo et al., 2023), and TAVDiffusion (Mao et al., 2024)).

Diffusion Formulation Let x_t denote the input text prompt, which is encoded via a pretrained text encoder $f_t(\cdot)$ to obtain $z_t = f_t(x_t)$. The video generation begins by sampling Gaussian noise $z_v^0 \sim \mathcal{N}(0, I)$, which is progressively denoised through the reverse diffusion process:

$$z_v^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_v^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \epsilon_{\theta_v}(z_v^t, z_t, t) \right) + \sigma_t \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{C.30}$$

Once the final video latent z_v^T is obtained, it is projected into an audio-guiding latent $\tilde{z}_a = \mathcal{P}_\theta(z_v^T)$. Audio generation is then conditioned on both $\tilde{z}_a$ and z_t using an analogous denoising

process:

$$z_a^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_a^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \epsilon_{\theta_a}(z_a^t, \tilde{z}_a, z_t, t) \right) + \sigma_t \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{C.31}$$

The generated latents are subsequently decoded using pretrained decoders to obtain the final outputs: $\hat{x}_v = g_v(z_v^T)$ and $\hat{x}_a = g_a(z_a^T)$.

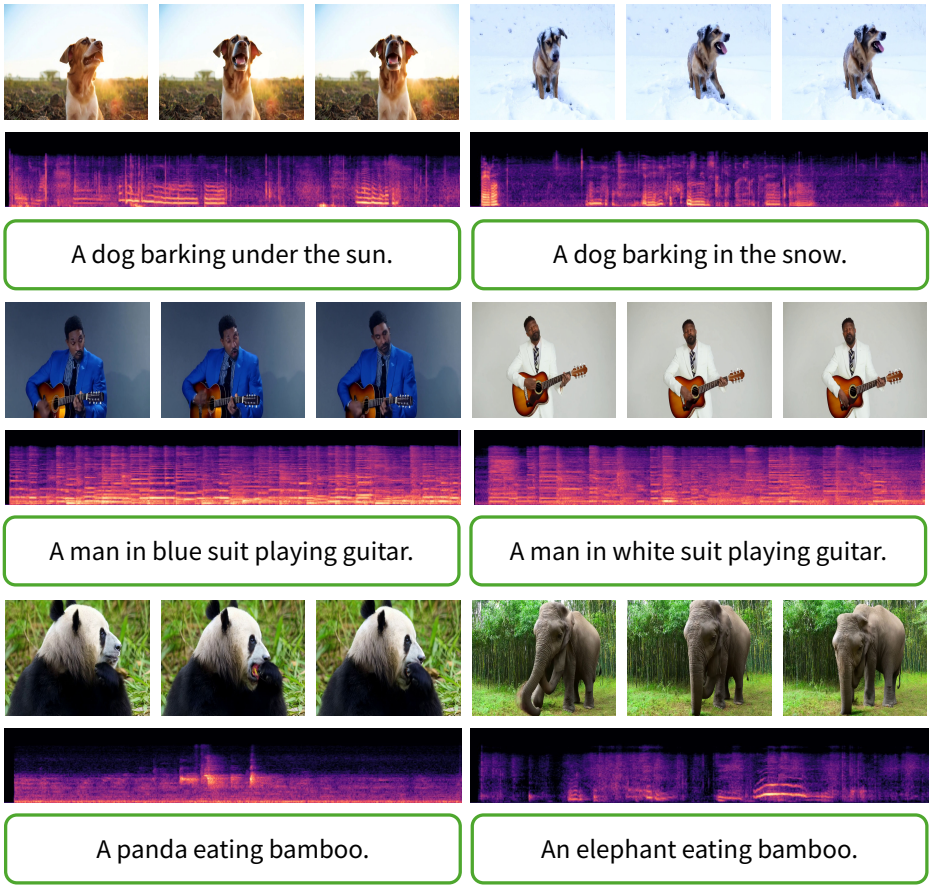

FIGURE C.2. Change of audible or non-audible attributes to the generative results.

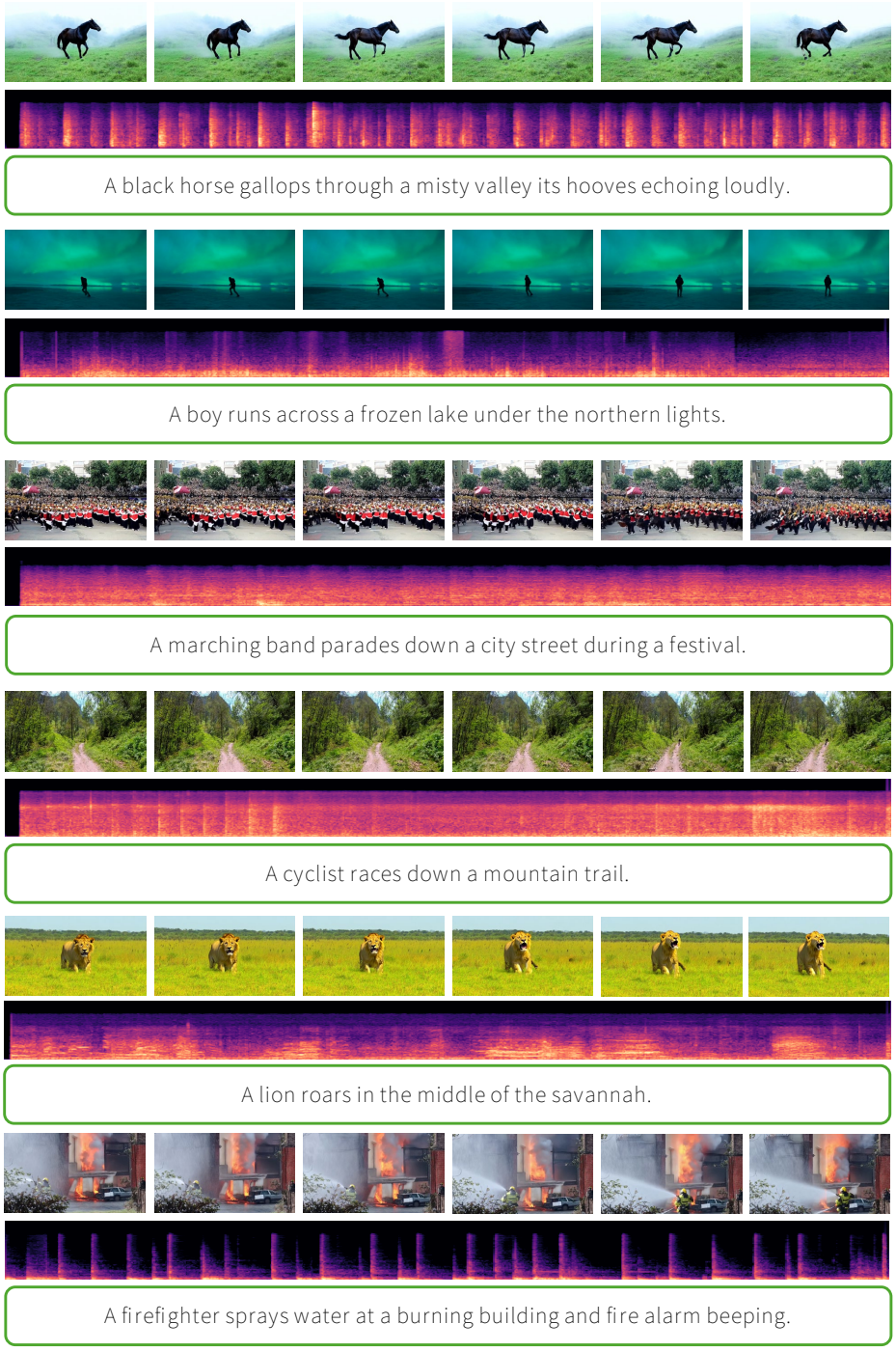

FIGURE C.3. Additional Text-to-Audio-Video generation results by our proposed framework.

C3 Additional Results

We present additional visualizations of our T2AV-generated results in Figure C.1. Using the same text prompts as those in (Mao et al., 2024), we generate audio-video pairs and compare them against existing baselines. In the *Disneyland* scene, our model produces vertically structured spectrogram features that plausibly correspond to discrete auditory events such as applause or exclamations, reflecting a diverse and dynamic soundscape. In the *sky with bird chirping* scene, we observe consistent, rhythmic patterns in the spectrogram that align with natural bird calls, indicating accurate temporal grounding of audio events. These results highlight the model's ability to synthesize semantically coherent and temporally aligned audio conditioned on visual content. Additional examples are provided in Figure C.1.

C4 Sensitivity Test

We conduct a sensitivity analysis to probe how the system responds to prompt variations that (i) alter non-audible attributes (e.g., background, static non-audible objects) while keeping the sound-causing event unchanged, and (ii) alter audible attributes (e.g., object/action that produces sound) while holding non-audible details fixed. As shown in Figure C.2, for each prompt pair, we generate matched video-audio samples under identical random seeds. Qualitatively, we visualize spectrograms alongside key video frames to inspect whether non-audible edits leave the soundtrack invariant and whether audible edits induce commensurate changes in rhythm and energy. Ideally, non-audible edits should produce minimal shifts in audio metrics, while audible edits should yield significant, directionally consistent changes. This analysis clarifies which aspects of the prompt our model treats as causally relevant for sound synthesis and highlights residual entanglements where visual-only edits still perturb the audio.

α	AVH↑	CAVP↑	FAD↓
0.001	0.174	0.150	5.52
0.01	0.197	0.159	5.60
0.1	0.206	0.165	5.49
1	0.199	0.154	5.31

TABLE C.1. Grid search summary for α.

λ	AVH↑	CAVP↑	FAD↓
1	0.208	0.161	7.03
5	0.206	0.165	5.49
10	0.174	0.150	7.42
100	0.172	0.141	6.97

TABLE C.2. Grid search summary for λ.

C5 Hyperparameter Studies

We study how the alignment weight α and sparsity weight λ affect performance in Table C.1 and C.2. We sweep $\alpha \in \{0.001, 0.01, 0.1, 1\}$ and $\lambda \in \{1, 5, 10, 100\}$ on a validation split, measuring AVHScore and CAVP similarity (alignment), FAD (audio quality). In unidirectional setting the FVD is not affected by alignment. We Find that: Reducing α too much weakens the latent coupling, leading to semantic drift between modalities (e.g., misaligned motion and sound). Excessively high λ over-prunes the mask. Disabling sparsity results in overly dense masks, which fail to isolate cross-modal signals and slightly reduce performance in AVHScore and CAVP similarity.

C6 Limitations

While our cascaded T2AV framework demonstrates strong performance in both generative quality and cross-modal alignment, several limitations remain. First, the reliance on sequential generation, where video is produced before audio, introduces unidirectional dependency that may constrain expressiveness in audio-visual co-synchronization, particularly for content requiring tight mutual feedback. Second, the alignment mechanism is trained on temporally segmented clips, which may limit its ability to generalize to complex or highly dynamic temporal structures in longer sequences. Third, although we finetune the diffusion models to better adapt to generated embeddings, the overall generation quality remains bounded by the capacity and resolution of the pretrained backbones, which are not jointly optimized with the alignment modules.

APPENDIX D

Appendix of Chapter 5

D1 Medical Coding Background

Medical coding task. Medical coding is the process of translating unstructured clinical documentation into standardized diagnostic and procedural codes, most commonly following the International Classification of Diseases (ICD) standard. Each clinical note may contain multiple diagnoses, procedures, and contextual information that must be mapped to corresponding ICD-10-CM (diagnosis) and ICD-10-PCS (procedure) codes. The task is inherently *multi-label* and *context-dependent*.

Example. Consider the following excerpt from a fabricated discharge summary:

> *"The patient was admitted for acute myocardial infarction and underwent coronary artery bypass graft (CABG). The postoperative course was uncomplicated."*

A correct ICD-10-CM coding outcome for this note includes:

- `I21.3` — ST elevation (STEMI) of unspecified site.

and for ICD-10-PCS (procedure codes):

- `021009W` — Bypass coronary artery, one site, autologous vein, open approach.

Challenges. Accurate code assignment requires multiple reasoning steps:

(1) Identifying relevant clinical entities (e.g., diseases, procedures).

(2) Consulting external resources such as the ICD alphabetic index and tabular list.
(3) Applying coding guidelines (e.g., sequencing, laterality, and exclusion rules).
(4) Cross-checking consistency between diagnoses and procedures.

For example, a note describing "Type 2 diabetes with chronic kidney disease" must yield both the primary diagnosis (`E11.22`) and the complication (`N18.9`) while respecting guideline-specific linkage rules.

Motivation for automation. Given the need for cross-referencing multiple resources and enforcing rule-based logic, medical coding naturally lends itself to *workflow-based* reasoning rather than direct text classification. The MedDCR framework aims to automate this multi-step reasoning process by discovering workflows that can effectively combine tool use, rule application, and reflection.

D2 Data Consent and Usage

All datasets used in this study are derived from publicly available, de-identified clinical corpora with appropriate data use agreements. The MDACE dataset (Cheng et al., 2023a) is constructed from the MIMIC-III database (Johnson et al., 2016), which contains de-identified electronic health records from the Beth Israel Deaconess Medical Center. Access to MIMIC-III requires completion of the PhysioNet credentialing process and acceptance of a data use agreement that prohibits any attempt at patient re-identification. Our use of MDACE fully complies with these requirements.

The ACI-BENCH dataset (Yim et al., 2023) is publicly released under an open-access license and contains synthetic or anonymized clinical notes for benchmarking purposes. No protected health information (PHI) was accessed or processed in any part of this study. All experiments were conducted in accordance with institutional data governance and ethical use policies.

D3 Case Study and Pseudo-Code of the Searched Workflow

Overview. To better illustrate how MedDCR operates in practice, we present the best workflows discovered on the ACI-Benchmark dataset. This case study corresponds to the pointed workflow in Figure 5.2, showing how iterative search progressively refines design choices through reflective feedback and archive-guided learning (Algorithm **??**).

Workflow evolution. In early iterations, the Designer generated linear pipelines focused mainly on entity extraction and description matching. As search progressed, reflective feedback encouraged the integration of verification and reconciliation steps, such as cross-checking alphabetic and tabular index results or re-validating low-confidence predictions. The final workflow incorporates diagnostic reasoning, combining multiple coding tools with validation and guideline enforcement to ensure consistency for the final predictions.

Pseudo-code of the searched workflow. The pipeline balances recall and precision through a two-stage process. In the first stage (Lines 1–7, Algo **??**), it expands recall by generating a large pool of candidate codes from multiple sources including alphabetic index lookup with extracted terms (expanded with synonyms), term-based, and note-based proposals from LLMs, ensuring broad coverage. In the second stage, precision is progressively enforced through validation, evidence linking, and filtering. Invalid codes are removed via rule checks, an evidence-aware judge filters low-confidence candidates based on τ_{keep}, and contrastive screening γ prunes near-duplicates. Finally, reranking integrates judge scores, description similarity, and evidence strength to prioritize high-precision codes while maintaining recall from the initial expansion.

D4 Meta-Prompt and Coding Tools

D4.1 Meta-Prompt for the Designer Agent

The Designer agent is instructed through a meta-prompt that defines its role, input exemplars, design constraints, and expected JSON output format. The prompt dynamically incorporates

the top-k best-performing and n most-recent workflows from the memory archive. Below is the formatted version of the prompt used in this study.

This meta-prompt encourages creativity while constraining the Designer to produce executable, guideline-compliant workflow plans. Tool signatures are embedded to ensure awareness of available operations and their expected parameters.

D4.2 Meta-Prompt for the Coder Agent

The Coder agent translates workflow plans from the Designer into executable Python code that implements the proposed controller. The prompt strictly enforces the plan order, ensures syntactic validity, and outputs JSON-formatted code only. The full formatted prompt is shown below.

The Coder's self-fixing loop detects non-executable code via Python traceback parsing and regenerates corrected versions until a valid workflow is obtained.

D4.3 Meta-Prompt for the Reflector Agent

The Reflector agent evaluates workflow executions, producing both quantitative scores and textual feedback. It receives the workflow plan, predictions, and gold codes, and is responsible for scoring and analysing workflow effectiveness. The structured prompt used is shown below.

The textual feedback is appended to the archive and shown to the Designer in the next iteration, enabling reflective learning.

D4.4 ICD-10 Coding Tool List

The framework provides a curated library of coding tools accessible to both the Designer and Coder agents. These tools are mainly adopted from the simple-icd-10-cm resource for CM codes[1], e.g.,

[1]A full list of tools is available at https://github.com/StefanoTrv/simple_icd_10_CM

- `get_parent` - returns the immediate parent of a code in the ICD-10-CM hierarchy; `prioritize_blocks` disambiguates codes that can denote either a block or a category.
- `get_children` - returns the immediate children of a code; if a code name is ambiguous (both a block and a category), setting `prioritize_blocks=True` treats it as the block (e.g., "B99" example).
- `get_ancestors` - returns all ancestor nodes (e.g., category to block to chapter) of the given code, with optional block prioritization to resolve ambiguity.
- `get_descendants` - returns all descendant nodes of the given code, with optional block prioritization for ambiguous names.
- `is_ancestor` - returns `True` iff `a` is an ancestor of `b`; optional flags control how to interpret ambiguous block/category codes.
- `is_descendant` - returns `True` iff `a` is a descendant of `b`; with the same ambiguity controls.
- `get_nearest_common_ancestor` - returns the nearest common ancestor of `a` and `b` (or empty string if none), with optional block/category disambiguation.
- `is_leaf` - returns `True` iff the code is a leaf (has no children) in the ICD-10-CM hierarchy; supports block prioritization.
- `code_to_text` - returns the textual description associated with a given code.
- `text_to_codes` - maps natural-language text to candidate ICD-10-CM codes (by matching code descriptions).
- `load_taxonomy` - loads the internal ICD-10-CM taxonomy data structure for lookups and hierarchies.
- `is_valid_code` - checks whether the given ICD-10-CM code exists in the taxonomy.

and we additionally create more complex LLM-based tools for both ICD-10-CM (diagnoses) and ICD-10-PCS (procedures):

- `MedicalTermExtraction` - retrieves medical terms from the note using Langextract.

- `TabularIndexSearch` - retrieves hierarchical and related codes from the tabular list.
- `DescMatch` - computes similarity between ICD descriptions and note text.
- `GuidelineValidator` - applies ICD-10 coding rules to remove invalid or conflicting codes.
- `Reconciler` - merges duplicates and resolves inter-tool inconsistencies.
- `EvidenceLinker` - extracts supporting text snippets for each predicted code.
- `Judge` - assigns per-code confidence scores and keep/drop decisions.

Beyond the predefined toolset, the Designer agent is explicitly encouraged to invent new tools using the same LLM interface (`LLMAgentBase.run_text`). When proposing new operations, the Designer must specify their role, purpose, and expected input/output signatures. This design encourages creative exploration of novel reasoning and validation strategies, allowing MedDCR to continuously expand its operational toolkit beyond the initial ICD-10 functions.

- `LLMAgentBase.run_text(prompt, role, loops)` - performs controlled text generation for tasks such as code proposal, evidence linking, or reflection. Explicit roles (e.g., “coder”, “judge”, “reflector”) are specified, and parameters such as `loops` are kept consistent with the workflow plan.

These tools collectively enable workflow designs that emulate professional coding processes, combining retrieval, validation, and reasoning for robust code prediction.

LISTING D.1. Meta-prompt of Designer Agent (Part-1)

```
You are designing a NEW controller for ICD-10 multi-label medical coding.
STAGE A (THIS TURN): PROPOSE A PLAN ONLY - DO NOT WRITE CODE.
Return a compact JSON plan that the system will implement later.

Think creatively while maintaining coding accuracy.

Rules:
- The plan is an ordered list of steps; each step is an object with an
    ↪ operation key ("op")
  and its required parameters.
- Exclude codes related to family history, negative, hypothetical, or
    ↪ ruled-out mentions.
- Include key parameters when relevant (e.g., samples, threshold, k, strategy,
    ↪ by).
- Avoid repeating recent workflow families unless changes are expected to
    ↪ improve F1.
- Keep the workflow concise and executable; the system will validate it.
- You may combine or extend existing operations if logically beneficial.

Available Tool Signatures:
MedicalTermExtraction(note) - Extract medical terms from the medical notes
    ↪ using LangExtract.
TabularIndexSearch(entity)   - Retrive code description from tabular index.
DescMatch(codes, note)       - Compute similarity between code descriptions
    ↪ and note text.
GuidelineValidator(codes, ruleset) - Filter codes violating ICD-10 rules.
Reconciler(codes)                - Merge duplicates or conflicting
    ↪ predictions.
EvidenceLinker(note, codes, window, max_snips) - Extract evidence snippets per
    ↪ code.
Judge(codes, evidence, strategy)  - Assign per-code confidence scores.
...
```

LISTING D.2. Meta-prompt of Designer Agent (Part-2)

```
Exemplars (for diversity & performance):
First are top performers by F1 (TOP_i), followed by the most recent attempts
    ↪ (RECENT_i).
Use these to inspire but not copy - propose an improved or novel workflow.

Top_1:
[Exemplar Workflow i]
...

Recent_1:
[Exemplar Workflow j]
...

Deliverable (Stage A): Return STRICTLY ONE JSON OBJECT (no prose):
{
  "name": "<controller name>",
  "thought": "<why this plan should improve F1>",
  "plan": [ { "op": "...", "...": "..." }, ... ]
}
If you introduce new operations, name them clearly - the system will map or
    ↪ extend the op set.
DO NOT include executable code in this turn.
```

LISTING D.3. Meta-prompt of Coder Agent

```
Stage B: Implement code that EXACTLY follows the accepted plan.

Implement `class Controller` for ICD-10 multi-label coding by following this
    ↪ accepted plan JSON:
<INSERT PLAN JSON HERE>

Constraints:
 - Keep EXACT signature:
       class Controller:
           def forward(self, task)
 - Honour the plan order and parameters (thresholds, temperatures, samples, k,
    ↪ strategies, etc.).
 - Do NOT include any comments in code.
 - Return STRICT JSON ONLY (no prose, no markdown):
   {
     "code": "class Controller:\n    def forward(self, task):\n        ..."
   }

Implementation guidance:
 - For LLM calls, use LLMAgentBase.run_text with explicit roles and keep
    ↪ temperature/loops consistent with the plan.
 - Always de-duplicate outputs.

 === REFERENCE EXAMPLES (for guidance only; DO NOT change your plan) ===
 Reference plan JSON:
   <example plan JSON>

 Reference full implementation (follows the reference plan):
   <example implementation code>

 Additional example snippets:
   <short code fragments for guidance>
```

LISTING D.4. Meta-prompt of Reflector Agent

```
You are the Reflector agent in the MedDCR framework.

Your goal is to evaluate the performance of a medical coding workflow
and provide concise feedback that helps the Designer improve in the next
    ↪ iteration.

Inputs:
 - Workflow plan JSON describing the sequence of operations.
 - Model predictions with supporting evidence spans.
 - Ground-truth ICD-10 codes from the validation set.
 - Quantitative scores (precision, recall, F1) computed for this workflow.

Tasks:
 1. Verify the provided scores for correctness and internal consistency.
 2. Identify the workflow's strengths and weaknesses in terms of tool use,
    reasoning order, and guideline adherence.
 3. Analyse failure cases such as:
      * incorrect entity extraction,
      * guideline violations,
      * missing or redundant validation steps,
      * low-confidence or conflicting predictions.
 4. Provide concise textual feedback that diagnoses the cause of errors
    and suggests actionable refinements (e.g., add validation, change order of
    ↪ tools, adjust threshold).
 5. Maintain objectivity-do not modify the workflow plan or fabricate new
    ↪ scores.

Return STRICT JSON ONLY (no prose, no markdown):
{
  "score": {
      "precision": <float>,
      "recall": <float>,
      "f1": <float>
  },
  "feedback": <str>"
}

Example feedback:
<Example Feedback>
```

Algorithm 2 Evidence-Aware Coding Pipeline

Require: Clinical note N; hyperparameters: T_{terms} (term extractor), S=4, $S_{\max}$=2, W=24, τ_{keep}=0.5, γ=0.15, K=20

Ensure: Ranked list $\mathcal{R}$ of codes with scores and evidence

$E \leftarrow \textsc{Terms}(\textsc{ReasonForVisit}(N); T_{\text{terms}})$ ▷ extract terms/entities w/ reason-for-visit cues
$C_s \leftarrow \textsc{SearchAlphaIndex}(\textsc{SynExp}(E))$
$C_t \leftarrow \textsc{ProposeFromTerms}(E; \text{mode} = \text{plain})$
$C_{tc} \leftarrow \textsc{ProposeFromTermsCoT}(E; \text{ref} = \text{sec}; \text{mode} = \text{CoT})$
$C_n \leftarrow \textsc{ProposeFromNote}(\text{RFV}(N); \text{mode} = \text{plain}; \text{samples} = S)$
$C_{nc} \leftarrow \textsc{ProposeFromSec}(N; \text{mode} = \text{CoT}, \text{samples} = S)$
$\hat{C} \leftarrow \textsc{Merge}(C_s, C_t, C_n, C_{tc}, C_{nc})$ ▷ set union with deduplication
$\hat{C} \leftarrow \textsc{Canonicalize}(\hat{C})$
$\hat{C} \leftarrow \textsc{Validate}(\hat{C})$ ▷ schema/rule compliance; remove invalid codes
$D \leftarrow \textsc{FetchDesc}(\hat{C})$ ▷ D : map code $\mapsto$ description from tabular index
$m_{\text{desc}} \leftarrow \textsc{DescMatch}(\hat{C}, D, N)$ ▷ per-code description$\leftrightarrow$note match score
$Z \leftarrow \textsc{EvidenceLink}(N, \hat{C}; S_{\max}, W)$ ▷ Z : map code $\mapsto$ at most $S_{\max}$ snippets (window W tokens)
$S_{\text{judge}} \leftarrow \textsc{Judge}(\hat{C}, Z; \text{judge\_strategy} = \text{per-code evidence-aware keep/drop}, \text{output} = \text{json\_scores})$ ▷ score in $[0, 1]$ for each code
$\tilde{C} \leftarrow \{\, c \in \hat{C} : S_{\text{judge}}(c) \geq \tau_{\text{keep}} \,\}$ ▷ filter by judge score
$\tilde{C} \leftarrow \textsc{ContrastiveScreen}(\tilde{C}; \text{by} = \text{sibling\_desc}, \text{contrast\_margin} = \gamma, \text{action} = \text{drop\_or\_demote})$ ▷ prune/demote near-duplicate siblings with small description margin
for all $c \in \tilde{C}$ **do**
  $s(c) \leftarrow \textsc{RerankScore}(S_{\text{judge}}(c), m_{\text{desc}}(c), \textsc{EvidenceStrength}(Z(c)))$
end for
$\mathcal{R} \leftarrow \textsc{SortBy}(\tilde{C}, \text{key} = s, \text{desc})$
if $|\mathcal{R}| < K$ **then**
  $\mathcal{B} \leftarrow \textsc{FallbackTopK}(\hat{C} \setminus \tilde{C}, K - |\mathcal{R}|)$
  $\mathcal{R} \leftarrow \textsc{Concat}(\mathcal{R}, \mathcal{B})$
end if
return $\textsc{ReturnRaw}(\mathcal{R}, s, Z, m_{\text{desc}})$

Algorithm 3 Evidence- and Vote-Aware Coding

Require: Clinical note N; S_t=3, S_n=4, $S_{\max}$=2, M=256, θ=0.5, K=20

Ensure: Ranked list $\mathcal{R}$ of codes with scores and evidence

$E \leftarrow \textsc{Terms}(N)$
$C_t^{(1:S_t)} \leftarrow \textsc{ProposeFromTerms}(E;\ \text{mode} = \text{plain},\ \text{samples} = S_t)$
$C_n^{(1:S_n)} \leftarrow \textsc{ProposeFromNote}(N;\ \text{mode} = \text{CoT},\ \text{samples} = S_n)$
$\hat{C} \leftarrow \textsc{Merge}\Big(\bigcup_{s=1}^{S_t} C_t^{(s)},\ \bigcup_{s=1}^{S_n} C_n^{(s)}\Big)$
$\hat{C} \leftarrow \textsc{Canonicalize}(\hat{C})$
$\hat{C} \leftarrow \textsc{Validate}(\hat{C})$ ▷ Self-consistency via normalized vote frequency
$\text{votes}(c) \leftarrow \sum_{s=1}^{S_t} \mathbf{1}\{c \in C_t^{(s)}\} + \sum_{s=1}^{S_n} \mathbf{1}\{c \in C_n^{(s)}\}$ for $c \in \hat{C}$
$\text{vote\_ratio}(c) \leftarrow \dfrac{\text{votes}(c)}{S_t + S_n}$; annotate as "vote_ratio"
$D \leftarrow \textsc{FetchDesc}(\hat{C})$ ▷ D: code $\mapsto$ tabular description
$m_{\text{desc}} \leftarrow \textsc{DescMatch}(\hat{C}, D, N)$ ▷ per-code description$\leftrightarrow$note match score in $[0, 1]$
$Z \leftarrow \textsc{EvidenceExtract}(N, \hat{C};\ \text{strategy} = \text{snippet\_from\_note},\ \text{per\_code\_spans} = S_{\max},\ \text{max\_tokens} = M)$
$\text{evidence\_overlap}(c) \leftarrow \textsc{EvidenceOverlap}(Z(c), N)$; annotate as "evidence_overlap"
$(\text{judge\_keep}(c), \text{judge\_conf}(c)) \leftarrow \textsc{Judge}(c, Z(c), D(c);\ \text{strategy} = \text{evidence\_tabular\_desc})$ for $c \in \hat{C}$
annotate judge_conf as "judge_conf"
$\tilde{C} \leftarrow \{\, c \in \hat{C} : \text{judge\_keep}(c) \vee \text{judge\_conf}(c) \geq \theta \,\}$
$\tilde{C} \leftarrow \textsc{HierPrune}(\tilde{C};\ \text{rules} = \{\text{prefer\_specific\_over\_unspecified}, \text{drop\_duplicate\_laterality}, \text{drop\_mutually\_exclusive\_with\_lower\_conf}\})$

for all $c \in \tilde{C}$ **do**
  $s(c) \leftarrow 0.4 \cdot \text{vote\_ratio}(c) + 0.3 \cdot m_{\text{desc}}(c) + 0.2 \cdot \text{judge\_conf}(c) + 0.1 \cdot \text{evidence\_overlap}(c)$
end for
$\mathcal{R} \leftarrow \textsc{SortBy}(\tilde{C},\ \text{key} = s,\ \text{desc})$
if $|\mathcal{R}| < K$ **then**
  $\mathcal{B} \leftarrow \textsc{FallbackTopK}(\hat{C} \setminus \tilde{C},\ K - |\mathcal{R}|)$
  $\mathcal{R} \leftarrow \textsc{Concat}(\mathcal{R},\ \mathcal{B})$
end if
return $\textsc{ReturnRaw}(\mathcal{R},\ s,\ Z,\ \text{vote\_ratio},\ m_{\text{desc}},\ \text{judge\_conf},\ \text{evidence\_overlap})$

Bibliography

Agostinelli, Andrea et al. (2023). 'Musiclm: Generating music from text'. In: *arXiv preprint arXiv:2301.11325*.

Alayrac, Jean-Baptiste et al. (2022). 'Flamingo: a visual language model for few-shot learning'. In: *Advances in neural information processing systems* 35, pp. 23716–23736.

Alonso, Vera et al. (2020). 'Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions'. In: *Journal of medical systems* 44.3, p. 62.

Ann Barta, MSA (2009). 'ICD-10-CM Official Coding Guidelines'. In: *Journal of AHIMA*.

Antoniou, Antreas, Amos Storkey and Harrison Edwards (2017). 'Data augmentation generative adversarial networks'. In: *arXiv preprint arXiv:1711.04340*.

Bai, Yutong et al. (2024). 'Sequential modeling enables scalable learning for large vision models'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Baksi, Krishanu Das et al. (2025). 'MedCodER: A Generative AI Assistant for Medical Coding'. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 449–459.

Bar, Amir et al. (2022). 'Visual prompting via image inpainting'. In: *Advances in Neural Information Processing Systems* 35, pp. 25005–25017.

Beckham, Christopher and Christopher Pal (2017). 'Unimodal probability distributions for deep ordinal classification'. In: *International Conference on Machine Learning*. PMLR, pp. 411–419.

Blattmann, Andreas et al. (2023). 'Align your latents: High-resolution video synthesis with latent diffusion models'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575.

Boyle, Joseph Spartacus et al. (2023). 'Automated clinical coding using off-the-shelf large language models'. In: *Deep Generative Models for Health Workshop, Advances in neural information processing systems*.

Brown, Tom et al. (2020). 'Language models are few-shot learners'. In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Campbell, Sharon and Katrina Giadresco (2020). 'Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals'. In: *Health Information Management Journal* 49.1, pp. 5–18.

Cao, Pengfei et al. (2020). 'Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 294–301.

Caron, Mathilde et al. (2021). 'Emerging properties in self-supervised vision transformers'. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Chang, Huiwen et al. (2023). 'Muse: Text-to-image generation via masked generative transformers'. In: *arXiv preprint arXiv:2301.00704*.

Chen, Banghao et al. (2023a). 'Unleashing the potential of prompt engineering in large language models: a comprehensive review'. In: *arXiv preprint arXiv:2310.14735*.

Chen, Haoxin et al. (2023b). 'Videocrafter1: Open diffusion models for high-quality video generation'. In: *arXiv preprint arXiv:2310.19512*.

Chen, Honglie et al. (2020a). 'VGGSound: A Large-scale Audio-Visual Dataset'. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Chen, Ricky TQ et al. (2018). 'Isolating sources of disentanglement in variational autoencoders'. In: *Advances in neural information processing systems* 31.

Chen, Ting et al. (2020b). 'A simple framework for contrastive learning of visual representations'. In: *International conference on machine learning*. PmLR, pp. 1597–1607.

Chen, Xi et al. (2016). 'Infogan: Interpretable representation learning by information maximizing generative adversarial nets'. In: *Advances in neural information processing systems* 29.

Chen, Xinlei et al. (2020c). 'Improved baselines with momentum contrastive learning'. In: *arXiv preprint arXiv:2003.04297*.

Cheng, Hua et al. (2023a). 'MDACE: MIMIC Documents Annotated with Code Evidence'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7534–7550.

Cheng, Yi et al. (2023b). 'Robust Image Ordinal Regression with Controllable Image Generation'. In: *arXiv preprint arXiv:2305.04213*.

Deng, Jia et al. (2009). 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Dhuliawala, Shehzaad et al. (2024). 'Chain-of-Verification Reduces Hallucination in Large Language Models'. In: *Findings of the Association for Computational Linguistics ACL 2024*, pp. 3563–3578.

Diaz, Raul and Amit Marathe (2019). 'Soft labels for ordinal regression'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4738–4747.

Dong, Hang et al. (2021). 'Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation'. In: *Journal of biomedical informatics* 116, p. 103728.

Dong, Hang et al. (2022a). 'Automated clinical coding: what, why, and where we are?' In: *NPJ digital medicine* 5.1, p. 159.

Dong, Qingxiu et al. (2022b). 'A survey on in-context learning'. In: *arXiv preprint arXiv:2301.00234*.

Douglas, James C et al. (2025). 'Less is More: Explainable and Efficient ICD Code Prediction with Clinical Entities'. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30835–30847.

Du, Yilun et al. (2023). 'Improving factuality and reasoning in language models through multiagent debate'. In: *Forty-first International Conference on Machine Learning*.

Edin, Joakim et al. (2024). 'An Unsupervised Approach to Achieve Supervised-Level Explainability in Healthcare Records'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4869–4890.

Eidinger, Eran, Roee Enbar and Tal Hassner (2014). 'Age and gender estimation of unfiltered faces'. In: *IEEE Transactions on information forensics and security* 9.12, pp. 2170–2179.

Esser, Patrick, Robin Rombach and Bjorn Ommer (2021). 'Taming transformers for high-resolution image synthesis'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883.

Falis, Matúš et al. (2024). 'Can GPT-3.5 generate and code discharge summaries?' In: *Journal of the American Medical Informatics Association* 31.10, pp. 2284–2293.

Farkas, Richárd and György Szarvas (2008). 'Automatic construction of rule-based ICD-9-CM coding systems'. In: *BMC bioinformatics* 9.Suppl 3, S10.

Fernando, Chrisantha et al. (2024). 'Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution'. In: *Forty-first International Conference on Machine Learning*. URL: https://openreview.net/forum?id=9ZxnPZGmPU.

Gan, Yidong et al. (2025). 'Aligning AI Research with the Needs of Clinical Coding Workflows: Eight Recommendations Based on US Data Analysis and Critical Review'. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 909–922.

Gandelsman, Yossi, Alexei A Efros and Jacob Steinhardt (2024). 'Interpreting CLIP's Image Representation via Text-Based Decomposition'. In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=5Ca9sSzuDp.

Ge, Jiaxin et al. (2023). 'Chain of thought prompt tuning in vision language models'. In: *arXiv preprint arXiv:2304.07919*.

Gero, Zelalem et al. (2023). 'Self-verification improves few-shot clinical information extraction'. In: *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Ghosal, Deepanway et al. (2023). 'Text-to-audio generation using instruction guided latent diffusion model'. In: *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3590–3598.

Girdhar, Rohit et al. (2023). 'ImageBind: One Embedding Space To Bind Them All'. In: *CVPR*.

Glymour, Clark and Kun Zhang (2019). 'Review of causal discovery methods based on graphical models'. In: *Frontiers in genetics* 10, p. 418407.

Goel, Akshay (July 2025). *LangExtract*. Version 1.0.3. DOI: 10.5281/zenodo.17015089. URL: https://github.com/google/langextract.

Goodfellow, Ian et al. (2014). 'Generative adversarial nets'. In: *Advances in neural information processing systems* 27.

Gui, Jie et al. (2024). 'A survey on self-supervised learning: Algorithms, applications, and future trends'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12, pp. 9052–9071.

Guo, Jianyuan et al. (2024). 'Data-efficient Large Vision Models through Sequential Autoregression'. In: *arXiv preprint arXiv:2402.04841*.

Guo, Yuwei et al. (2023). 'Animatediff: Animate your personalized text-to-image diffusion models without specific tuning'. In: *arXiv preprint arXiv:2307.04725*.

Hao, Zhiwei et al. (2024). 'Data-efficient Large Vision Models through Sequential Autoregression'. In: *Forty-first International Conference on Machine Learning*. URL: https://openreview.net/forum?id=KmCoS6WkgG.

He, Kaiming et al. (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Kaiming et al. (2019). 'Momentum Contrast for Unsupervised Visual Representation Learning'. In: *arXiv preprint arXiv:1911.05722*.

He, Yingqing et al. (2024). 'Llms meet multimodal generation and editing: A survey'. In: *arXiv preprint arXiv:2405.19334*.

Higgins, Irina et al. (2017a). 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Sy2fzU9gl.

— (2017b). 'beta-vae: Learning basic visual concepts with a constrained variational framework'. In: *International conference on learning representations*.

Hong, Jie et al. (2024a). 'Goss: Towards generalized open-set semantic segmentation'. In: *The Visual Computer* 40.4, pp. 2391–2404.

Hong, Wenyi et al. (2023). 'CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers'. In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rB6TpjAuSRy.

Hong, Ziming et al. (2024b). 'Improving Non-Transferable Representation Learning by Harnessing Content and Style'. In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=FYKVPOHCpE.

Houlsby, Neil et al. (2019). 'Parameter-efficient transfer learning for NLP'. In: *International conference on machine learning*. PMLR, pp. 2790–2799.

Hu, Edward J et al. (2022). 'Lora: Low-rank adaptation of large language models.' In: *ICLR* 1.2, p. 3.

Hu, Li (2024). 'Animate anyone: Consistent and controllable image-to-video synthesis for character animation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163.

Hu, Shengran, Cong Lu and Jeff Clune (2025). 'Automated Design of Agentic Systems'. In: *The Thirteenth International Conference on Learning Representations*.

Huang, Chao-Wei, Shang-Chi Tsai and Yun-Nung Chen (2022a). 'PLM-ICD: Automatic ICD Coding with Pretrained Language Models'. In: *ClinicalNLP 2022*, p. 10.

Huang, Chin-Wei et al. (2018). 'Neural autoregressive flows'. In: *International Conference on Machine Learning*. PMLR, pp. 2078–2087.

Huang, Guang-Bin, Lei Chen and Chee-Kheong Siew (2024a). 'Universal approximation using incremental constructive feedforward networks with random hidden nodes'. In: *IEEE transactions on neural networks* 17.4, pp. 879–892.

Huang, Jiaxin et al. (2025a). 'MLLM-For3D: Adapting Multimodal Large Language Model for 3D Reasoning Segmentation'. In: *arXiv preprint arXiv:2503.18135*.

Huang, Jiaxin et al. (2025b). 'SURPRISE3D: A Dataset for Spatial Understanding and Reasoning in Complex 3D Scenes'. In: *arXiv preprint arXiv:2507.07781*.

Huang, Rongjie et al. (2023a). 'Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models'. In: *International Conference on Machine Learning*. PMLR, pp. 13916–13932.

Huang, Zhuo et al. (2021). 'Universal semi-supervised learning'. In: *Advances in Neural Information Processing Systems* 34, pp. 26714–26725.

Huang, Zhuo et al. (2022b). 'Harnessing out-of-distribution examples via augmenting content and style'. In: *arXiv preprint arXiv:2207.03162*.

— (2023b). ‘Harnessing Out-Of-Distribution Examples via Augmenting Content and Style’. In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=boNyg20-JDm.

Huang, Zhuo et al. (2024b). ‘Machine Vision Therapy: Multimodal Large Language Models Can Enhance Visual Robustness via Denoising In-Context Learning’. In: *Forty-first International Conference on Machine Learning*. URL: https://openreview.net/forum?id=LwOfVWgEzS.

Huang, Ziyuan et al. (2024c). ‘Accelerating Pre-training of Multimodal LLMs via Chain-of-Sight’. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hurst, Aaron et al. (2024). ‘Gpt-4o system card’. In: *arXiv preprint arXiv:2410.21276*.

Hyvarinen, Aapo and Hiroshi Morioka (2016). ‘Unsupervised feature extraction by time-contrastive learning and nonlinear ica’. In: *Advances in neural information processing systems* 29.

— (2017). ‘Nonlinear ICA of temporally dependent stationary sources’. In: *Artificial Intelligence and Statistics*. PMLR, pp. 460–469.

Iashin, Vladimir and Esa Rahtu (2021). ‘Taming visually guided sound generation’. In: *arXiv preprint arXiv:2110.08791*.

Jaiswal, Ashish et al. (2020). ‘A survey on contrastive self-supervised learning’. In: *Technologies* 9.1, p. 2.

Ji, Shaoxiong et al. (2024). ‘A unified review of deep learning for automated medical coding’. In: *ACM Computing Surveys* 56.12, pp. 1–41.

Johnson, Alistair et al. (2020). ‘Mimic-iv’. In: *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55.

Johnson, Alistair EW et al. (2016). ‘MIMIC-III, a freely accessible critical care database’. In: *Scientific data* 3.1, pp. 1–9.

Joshi, Harshit et al. (2023). ‘Repair is nearly generation: Multilingual program repair with llms’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5131–5140.

Karras, Tero et al. (2020). ‘Analyzing and Improving the Image Quality of StyleGAN’. In: *Proc. CVPR*.

Kavuluru, Ramakanth, Anthony Rios and Yuan Lu (2015). 'An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records'. In: *Artificial intelligence in medicine* 65.2, pp. 155–166.

Ke, Fucai et al. (2025). 'Explain before you answer: A survey on compositional visual reasoning'. In: *arXiv preprint arXiv:2508.17298*.

Khachatryan, Levon et al. (2023). 'Text2video-zero: Text-to-image diffusion models are zero-shot video generators'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964.

Khattab, Omar et al. (2024). 'Dspy: Compiling declarative language model calls into state-of-the-art pipelines'. In: *The Twelfth International Conference on Learning Representations*.

Khemakhem, Ilyes et al. (2020). 'Variational autoencoders and nonlinear ica: A unifying framework'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2207–2217.

Khosla, Prannay et al. (2020). 'Supervised contrastive learning'. In: *Advances in neural information processing systems* 33, pp. 18661–18673.

Kim, Chris Dongjoo et al. (2019). 'AudioCaps: Generating Captions for Audios in The Wild'. In: *NAACL-HLT*.

Kingma, Diederik P and Jimmy Ba (2014). 'Adam: A method for stochastic optimization'. In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). 'Auto-encoding variational bayes'. In: *arXiv preprint arXiv:1312.6114*.

Kojima, Takeshi et al. (2022). 'Large language models are zero-shot reasoners'. In: *Advances in neural information processing systems* 35, pp. 22199–22213.

Kong, Lingjing et al. (2022). 'Partial disentanglement for domain adaptation'. In: *International Conference on Machine Learning*. PMLR, pp. 11455–11472.

Kumar, Abhishek, Prasanna Sattigeri and Avinash Balakrishnan (2017). 'Variational inference of disentangled latent concepts from unlabeled observations'. In: *arXiv preprint arXiv:1711.00848*.

Kwan, Keith (2024). 'Large language models are good medical coders, if provided with tools'. In: *arXiv preprint arXiv:2407.12849*.

Levi, Gil and Tal Hassner (2015). 'Age and gender classification using convolutional neural networks'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–42.

Li, Fei and Hong Yu (2020). 'ICD coding from clinical text using multi-filter residual convolutional neural network'. In: *proceedings of the AAAI conference on artificial intelligence*, pp. 8180–8187.

Li, Feng et al. (2024a). 'Visual in-context prompting'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12861–12871.

Li, Guohao et al. (2023a). 'Camel: Communicative agents for" mind" exploration of large language model society'. In: *Advances in Neural Information Processing Systems* 36, pp. 51991–52008.

Li, Junnan et al. (2023b). 'Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models'. In: *International conference on machine learning*. PMLR, pp. 19730–19742.

Li, Junyou et al. (2024b). 'More Agents Is All You Need'. In: *Transactions on Machine Learning Research*. ISSN: 2835-8856. URL: https://openreview.net/forum?id=bgzUSZ8aeg.

Li, Muyang et al. (2023c). 'InstanT: Semi-supervised Learning with Instance-dependent Thresholds'. In: *Thirty-seventh Conference on Neural Information Processing Systems*.

Li, Rumeng, Xun Wang and Hong Yu (2024c). 'Exploring llm multi-agents for icd coding'. In: *arXiv preprint arXiv:2406.15363*.

Li, Shikun et al. (2022a). 'Selective-supervised contrastive learning with noisy labels'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325.

Li, Wanhua et al. (2019). 'Bridgenet: A continuity-aware probabilistic network for age estimation'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1145–1154.

Li, Wanhua et al. (2021). 'Learning probabilistic ordinal embeddings for uncertainty-aware regression'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13896–13905.

Li, Yewen et al. (2022b). 'Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae'. In: *Advances in Neural Information Processing Systems* 35, pp. 7383–7396.

Liang, Tian et al. (2024). 'Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate'. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904.

Lin, Runqi et al. (2025). 'Understanding and Enhancing the Transferability of Jailbreaking Attacks'. In: *arXiv preprint arXiv:2502.03052*.

Lin, Tsung-Yi et al. (2014). 'Microsoft coco: Common objects in context'. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, pp. 740–755.

Lin, Yexiong et al. (2023). 'CS-Isolate: Extracting Hard Confident Examples by Content and Style Isolation'. In: *Thirty-seventh Conference on Neural Information Processing Systems*.

Lin, Zinan et al. (2019). 'Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers'. In: *arXiv preprint arXiv:1906.06034*, p. 60.

Liu, Haohe et al. (2023a). 'AudioLDM: Text-to-Audio Generation with Latent Diffusion Models'. In: *International Conference on Machine Learning*. PMLR, pp. 21450–21474.

Liu, Haohe et al. (2024a). 'Audioldm 2: Learning holistic audio generation with self-supervised pretraining'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Liu, Haotian et al. (2023b). *Visual Instruction Tuning*.

Liu, Kai et al. (2025). 'JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization'. In: *arxiv*.

Liu, Ruhan et al. (2022). 'DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge'. In: *Patterns*, p. 100512. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2022.100512. URL: https://www.sciencedirect.com/science/article/pii/S2666389922001040.

Liu, Xiaofeng et al. (2018a). 'Ordinal regression with neuron stick-breaking for medical diagnosis'. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0.

Liu, Yanzhu, Adams Wai Kin Kong and Chi Keong Goh (2018b). 'A constrained deep neural network for ordinal regression'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 831–839.

Liu, Yanzhu, Adams Wai-Kin Kong and Chi Keong Goh (2017). 'Deep ordinal regression based on data relationship for small datasets.' In: *IJCAI*, pp. 2372–2378.

Liu, Yanzhu, Fan Wang and Adams Wai Kin Kong (2019). 'Probabilistic deep ordinal regression based on gaussian processes'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5301–5309.

Liu, Zuyan et al. (2024b). 'Chain-of-Spot: Interactive Reasoning Improves Large Vision-Language Models'. In: *arXiv preprint arXiv:2403.12966*.

Locatello, Francesco et al. (2020). 'Weakly-supervised disentanglement without compromises'. In: *International Conference on Machine Learning*. PMLR, pp. 6348–6359.

Lu, Pan et al. (2022). 'Learn to explain: Multimodal reasoning via thought chains for science question answering'. In: *Advances in Neural Information Processing Systems* 35, pp. 2507–2521.

Luo, Junyu et al. (2025). 'Large language model agent: A survey on methodology, applications and challenges'. In: *arXiv preprint arXiv:2503.21460*.

Luo, Run et al. (2024). 'Deem: Diffusion models serve as the eyes of large language models for image perception'. In: *arXiv preprint arXiv:2405.15232*.

Luo, Simian et al. (2023). 'Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models'. In: *Advances in Neural Information Processing Systems* 36, pp. 48855–48876.

Madaan, Aman et al. (2023). 'Self-refine: Iterative refinement with self-feedback'. In: *Advances in Neural Information Processing Systems* 36, pp. 46534–46594.

Majumder, Navonil et al. (2024). 'Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization'. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572.

Mao, Yuxin et al. (2024). 'TAVGBench: Benchmarking text to audible-video generation'. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6607–6616.

Mihail, Radu Paul et al. (2016). 'Sky Segmentation in the Wild: An Empirical Study'. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–6. DOI: 10.1109/WACV.2016.7477637.

Mirza, Mehdi and Simon Osindero (2014). 'Conditional generative adversarial nets'. In: *arXiv preprint arXiv:1411.1784*.

Mitra, Chancharik et al. (2023). 'Compositional chain-of-thought prompting for large multimodal models'. In: *arXiv preprint arXiv:2311.17076*.

Mokady, Ron, Amir Hertz and Amit H Bermano (2021). 'Clipcap: Clip prefix for image captioning'. In: *arXiv preprint arXiv:2111.09734*.

Motzfeldt, Andreas et al. (2025). 'Code Like Humans: A Multi-Agent Solution for Medical Coding'. In: *arXiv preprint arXiv:2509.05378*.

Mullenbach, James et al. (2018). 'Explainable Prediction of Medical Codes from Clinical Text'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111.

Mustafa, Akram, Usman Naseem and Mostafa Rahimi Azghadi (2025). 'Evaluating Hierarchical Clinical Document Classification Using Reasoning-Based LLMs'. In: *arXiv preprint arXiv:2507.03001*.

Nakano, Reiichiro et al. (2021). 'Webgpt: Browser-assisted question-answering with human feedback'. In: *arXiv preprint arXiv:2112.09332*.

Niu, Zhenxing et al. (2016). 'Ordinal regression with multiple output cnn for age estimation'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4920–4928.

Olausson, Theo X. et al. (2024). 'Is Self-Repair a Silver Bullet for Code Generation?' In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=y0GJXRungR.

Oord, Aaron van den, Yazhe Li and Oriol Vinyals (2018). 'Representation learning with contrastive predictive coding'. In: *arXiv preprint arXiv:1807.03748*.

OpenAI (2025). URL: https://cdn.openai.com/gpt-5-system-card.pdf.

Peters, Jonas, Dominik Janzing and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pryzant, Reid et al. (2023). 'Automatic Prompt Optimization with "Gradient Descent" and Beam Search'. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*. URL: https://openreview.net/forum?id=WRYhaSrThy.

Qin, Xuebin et al. (2020). 'U2-Net: Going deeper with nested U-structure for salient object detection'. In: *Pattern recognition* 106, p. 107404.

Qin, Yujia et al. (2024). 'ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs'. In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=dHng2O0Jjr.

Radford, Alec et al. (2021). 'Learning transferable visual models from natural language supervision'. In: *International conference on machine learning*. PmLR, pp. 8748–8763.

Rose, Daniel et al. (2023). 'Visual chain of thought: Bridging logical gaps with multimodal infillings'. In: *arXiv preprint arXiv:2305.02317*.

Rothe, Rasmus, Radu Timofte and Luc Van Gool (2018). 'Deep expectation of real and apparent age from a single image without facial landmarks'. In: *International Journal of Computer Vision* 126.2-4, pp. 144–157.

Ruan, Ludan et al. (2023). 'Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228.

Schick, Timo et al. (2023). 'Toolformer: Language models can teach themselves to use tools'. In: *Advances in Neural Information Processing Systems* 36, pp. 68539–68551.

Schuhmann, Christoph et al. (2022). 'Laion-5b: An open large-scale dataset for training next generation image-text models'. In: *Advances in neural information processing systems* 35, pp. 25278–25294.

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

Shafiq, Muhammad and Zhaoquan Gu (2022). 'Deep residual learning for image recognition: A survey'. In: *Applied sciences* 12.18, p. 8972.

Shanahan, Murray, Kyle McDonell and Laria Reynolds (2023). 'Role play with large language models'. In: *Nature* 623.7987, pp. 493–498.

Shen, Wei et al. (2018). 'Deep regression forests for age estimation'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2304–2313.

Shin, Nyeong-Ho, Seon-Ho Lee and Chang-Su Kim (2022). 'Moving window regression: A novel approach to ordinal regression'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18760–18769.

Shinn, Noah et al. (2023). 'Reflexion: Language agents with verbal reinforcement learning'. In: *Advances in Neural Information Processing Systems* 36, pp. 8634–8652.

Simonyan, Karen and Andrew Zisserman (2014). 'Very deep convolutional networks for large-scale image recognition'. In: *arXiv preprint arXiv:1409.1556*.

Singer, Uriel et al. (2023). 'Make-A-Video: Text-to-Video Generation without Text-Video Data'. In: *The Eleventh International Conference on Learning Representations*.

Sohl-Dickstein, Jascha et al. (2015). 'Deep unsupervised learning using nonequilibrium thermodynamics'. In: *International conference on machine learning*. PMLR, pp. 2256–2265.

Sun, Jiankai et al. (2025). 'A survey of reasoning with foundation models: Concepts, methodologies, and outlook'. In: *ACM Computing Surveys* 57.11, pp. 1–43.

Sun, Quan et al. (2024). 'Emu: Generative Pretraining in Multimodality'. In: *The Twelfth International Conference on Learning Representations*.

Tan, Xu et al. (2024). 'Naturalspeech: End-to-end text-to-speech synthesis with human-level quality'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.6, pp. 4234–4245.

Tang, Zineng et al. (2023). 'Any-to-any generation via composable diffusion'. In: *Advances in Neural Information Processing Systems* 36, pp. 16083–16099.

Team, Gemini et al. (2024). 'Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context'. In: *arXiv preprint arXiv:2403.05530*.

Touvron, Hugo et al. (2023a). ‘Llama 2: Open foundation and fine-tuned chat models’. In: *arXiv preprint arXiv:2307.09288*.

Touvron, Hugo et al. (2023b). ‘Llama: Open and efficient foundation language models’. In: *arXiv preprint arXiv:2302.13971*.

Trabucco, Brandon et al. (2023). ‘Effective data augmentation with diffusion models’. In: *arXiv preprint arXiv:2302.07944*.

Tran, Toan et al. (2017). ‘A bayesian data augmentation approach for learning deep models’. In: *Advances in neural information processing systems* 30.

Tu, Weijie et al. (2024). ‘Ranked from within: Ranking large multimodal models for visual question answering without labels’. In: *arXiv preprint arXiv:2412.06461*.

Unterthiner, Thomas et al. (2019). ‘FVD: A new metric for video generation’. In.

Van Den Oord, Aaron, Oriol Vinyals et al. (2017). ‘Neural discrete representation learning’. In: *Advances in neural information processing systems* 30.

Vaswani, Ashish et al. (2017). ‘Attention is all you need’. In: *Advances in neural information processing systems* 30.

Von Kügelgen, Julius et al. (2021). ‘Self-supervised learning with data augmentations provably isolates content from style’. In: *Advances in neural information processing systems* 34, pp. 16451–16467.

Vu, Thanh, Dat Quoc Nguyen and Anthony Nguyen (2021). ‘A label attention model for ICD coding from clinical text’. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3335–3341.

Wang, Haoyu et al. (2024a). ‘NoiseGPT: Label Noise Detection and Rectification through Probability Curvature’. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, Heng et al. (2024b). ‘V2A-Mapper: A Lightweight Solution for Vision-to-Audio Generation by Connecting Foundation Models’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, Jinhong et al. (2025a). ‘A Survey on Ordinal Regression: Applications, Advances and Prospects’. In: *arXiv preprint arXiv:2503.00952*.

Wang, Wenxiao, Priyatham Kattakinda and Soheil Feizi (2025b). 'Maestro: Joint Graph & Config Optimization for Reliable AI Agents'. In: *arXiv preprint arXiv:2509.04642*.

Wang, Xin et al. (2022a). 'Disentangled representation learning'. In: *arXiv preprint arXiv:2211.11695*.

— (2024c). 'Disentangled representation learning'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12, pp. 9677–9696.

Wang, Xinlong et al. (2023a). 'Images speak in images: A generalist painter for in-context visual learning'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839.

Wang, Xinlong et al. (2023b). 'Seggpt: Segmenting everything in context'. In: *arXiv preprint arXiv:2304.03284*.

Wang, Xuezhi et al. (2022b). 'Self-consistency improves chain of thought reasoning in language models'. In: *arXiv preprint arXiv:2203.11171*.

Wang, Xuezhi et al. (2023c). 'Self-Consistency Improves Chain of Thought Reasoning in Language Models'. In: *The Eleventh International Conference on Learning Representations*.

— (2023d). 'Self-Consistency Improves Chain of Thought Reasoning in Language Models'. In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=1PL1NIMMrw.

Wei, Jason et al. (2022). 'Chain-of-thought prompting elicits reasoning in large language models'. In: *Advances in neural information processing systems* 35, pp. 24824–24837.

Wei, Yuxiang et al. (2021). 'Orthogonal jacobian regularization for unsupervised disentanglement in image generation'. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6721–6730.

Wertheimer, Max and Kurt Riezler (1944). 'Gestalt theory'. In: *Social Research*, pp. 78–99.

Wu, Jay Zhangjie et al. (2023a). 'Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7623–7633.

Wu, Qingyun et al. (2024a). 'Autogen: Enabling next-gen LLM applications via multi-agent conversations'. In: *First Conference on Language Modeling*.

Wu, Shengqiong et al. (2024b). 'NExT-GPT: Any-to-Any Multimodal LLM'. In: *Forty-first International Conference on Machine Learning*. URL: https://openreview.net/forum?id=NZQkumsNlf.

Wu, Yusong et al. (2023b). 'Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation'. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.

Xia, Xiaobo et al. (2022). 'Pluralistic image completion with gaussian mixture models'. In: *Advances in Neural Information Processing Systems* 35, pp. 24087–24100.

Xiao, Tete et al. (2021). 'What Should Not Be Contrastive in Contrastive Learning'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=CZ8Y3NzuVzO.

Xie, Shaoan et al. (2022). 'Multi-domain image generation and translation with identifiability guarantees'. In: *The Eleventh International Conference on Learning Representations*.

— (2023). 'Multi-domain image generation and translation with identifiability guarantees'. In: *The Eleventh International Conference on Learning Representations*.

Xie, Shaoan et al. (2025). 'SmartCLIP: Modular Vision-language Alignment with Identification Guarantees'. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29780–29790.

Xing, Yazhou et al. (2024). 'Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161.

Yang, Chengrun et al. (2023a). 'Large language models as optimizers'. In: *The Twelfth International Conference on Learning Representations*.

Yang, Zhichao et al. (2023b). 'Multi-label few-shot icd coding as autoregressive generation with prompt'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5366–5374.

Yang, Zhichao et al. (2023c). 'Surpassing GPT-4 medical coding with a two-stage approach'. In: *arXiv preprint arXiv:2311.13735*.

Yang, Zhuoyi et al. (2024). 'CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer'. In: *arXiv preprint arXiv:2408.06072*.

Yao, Yu et al. (2021). ‘Instance-dependent label-noise learning under a structural causal model’. In: *Advances in Neural Information Processing Systems* 34, pp. 4409–4420.

Yao, Yu et al. (2023). ‘Which is better for learning with noisy labels: the semi-supervised method or modeling label noise?’ In: *International Conference on Machine Learning*. PMLR, pp. 39660–39673.

Yim, Wen-wai et al. (2023). ‘ACI-BENCH: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation’. In: *Nature Scientific Data*.

Yu, Keunwoo et al. (Nov. 2024). ‘Eliciting In-Context Learning in Vision-Language Models for Videos Through Curated Data Distributional Properties’. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 20416–20431. URL: https://aclanthology.org/2024.emnlp-main.1137.

Yuan, Zheng, Chuanqi Tan and Songfang Huang (2022). ‘Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding’. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 808–814.

Zakari, Rufai Yusuf et al. (2022). ‘Vqa and visual reasoning: An overview of recent datasets, methods and challenges’. In: *arXiv preprint arXiv:2212.13296*.

Żelaszczyk, Maciej and Jacek Mańdziuk (2024). ‘Text-to-image cross-modal generation: A systematic review’. In: *arXiv preprint arXiv:2401.11631*.

Zha, Kaiwen et al. (2022). ‘Supervised Contrastive Regression’. In: *arXiv preprint arXiv:2210.01189*.

Zhan, Huangying et al. (2022). ‘Activermap: Radiance field for active mapping and planning’. In: *arXiv preprint arXiv:2211.12656*.

Zhang, Guibin et al. (2025a). ‘Multi-agent Architecture Search via Agentic Supernet’. In: *Forty-second International Conference on Machine Learning*. URL: https://openreview.net/forum?id=imcyVlzpXh.

Zhang, Jialu et al. (2024a). ‘Pydex: Repairing bugs in introductory python assignments using llms’. In: *Proceedings of the ACM on Programming Languages* 8.OOPSLA1, pp. 1100–1124.

Zhang, Jiayi et al. (2025b). 'AFlow: Automating Agentic Workflow Generation'. In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=z5uVAKwmjf.

Zhang, Lei et al. (2024b). 'Hierarchical Context Pruning: Optimizing Real-World Code Completion with Repository-Level Pretrained Code LLMs'. In: *arXiv preprint arXiv:2406.18294*.

Zhang, Shaokun et al. (2023a). 'Ideal: Influence-driven selective annotations empower in-context learners in large language models'. In: *arXiv preprint arXiv:2310.10873*.

Zhang, Yabo et al. (2023b). 'Controlvideo: Training-free controllable text-to-video generation'. In: *arXiv preprint arXiv:2305.13077*.

Zhang, Yuanhan, Kaiyang Zhou and Ziwei Liu (2022a). *Neural Prompt Search*. arXiv: 2206.04673 [cs.CV].

— (2023c). 'What makes good examples for visual in-context learning?' In: *Advances in Neural Information Processing Systems* 36.

Zhang, Yuanhan et al. (Apr. 2024c). *LLaVA-NeXT: A Strong Zero-shot Video Understanding Model*. URL: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Zhang, Zhuosheng et al. (2022b). 'Automatic chain of thought prompting in large language models'. In: *arXiv preprint arXiv:2210.03493*.

— (2023d). 'Automatic Chain of Thought Prompting in Large Language Models'. In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=5NTt8GFjUHkr.

Zhang, Zhuosheng et al. (2023e). 'Multimodal chain-of-thought reasoning in language models'. In: *arXiv preprint arXiv:2302.00923*.

Zheng, Ge et al. (2023a). 'Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models'. In: *Advances in Neural Information Processing Systems* 36, pp. 5168–5191.

Zheng, Jiyang et al. (2022). 'Towards open-set object detection and discovery'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3961–3970.

Zheng, Jiyang et al. (2024). 'Enhancing contrastive learning for ordinal regression via ordinal content preserved data augmentation'. In: *The Twelfth International Conference on Learning Representations*.

Zheng, Jiyang et al. (2025). 'Chain-of-Focus Prompting: Leveraging Sequential Visual Cues to Prompt Large Autoregressive Vision Models'. In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=noidywkBba.

Zheng, Lianmin et al. (2023b). 'Judging llm-as-a-judge with mt-bench and chatbot arena'. In: *Advances in neural information processing systems* 36, pp. 46595–46623.

Zhong, Wanjun et al. (2024). 'Memorybank: Enhancing large language models with long-term memory'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 19724–19731.

Zhou, Guanglin et al. (2023). 'On the opportunity of causal deep generative models: A survey and future directions'. In: *arXiv preprint arXiv* 2301.

Zhou, Han et al. (2025). 'Multi-agent design: Optimizing agents with better prompts and topologies'. In: *arXiv preprint arXiv:2502.02533*.

Zhou, Yiwei et al. (2024). 'Few-Shot Adversarial Prompt Learning on Vision-Language Models'. In: *arXiv preprint arXiv:2403.14774*.

Zhu, Xinqi, Chang Xu and Dacheng Tao (2021). 'Where and what? examining interpretable disentangled representations'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5861–5870.

Zhu, Xizhou et al. (2023). 'Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory'. In: *arXiv preprint arXiv:2305.17144*.

Zimmermann, Roland S et al. (2021). 'Contrastive learning inverts the data generating process'. In: *International Conference on Machine Learning*. PMLR, pp. 12979–12990.