Exploiting Completeness Perception with Diffusion Transformer for Unified 3D MRI Synthesis

Anonymous authors
Paper under double-blind review

Abstract

Missing data problems, such as missing modalities in multi-modal brain MRI and missing slices in cardiac MRI, pose significant challenges in clinical practice. Existing methods rely on external guidance to supply detailed missing state for instructing generative models to synthesize missing MRIs. However, manual indicators are not always available or reliable in real-world scenarios due to the unpredictable nature of clinical environments. Moreover, these explicit masks are not informative enough to provide guidance for improving semantic consistency. In this work, we argue that generative models should infer and recognize missing states in a self-perceptive manner, enabling them to better capture subtle anatomical and pathological variations. Towards this goal, we propose CoPeDiT, a general-purpose latent diffusion model equipped with **completeness perception** for unified synthesis of 3D MRIs. Specifically, we incorporate dedicated pretext tasks into our tokenizer, CoPeVAE, empowering it to learn completeness-aware discriminative prompts, and design MDiT3D, a specialized diffusion transformer architecture for 3D MRI synthesis that effectively uses the learned prompts as guidance to enhance semantic consistency in 3D space. Comprehensive evaluations on three large-scale MRI datasets demonstrate that CoPeDiT significantly outperforms state-of-the-art methods, achieving superior robustness and yielding high-fidelity, structurally consistent synthesis across diverse missing patterns.

1 Introduction

Magnetic resonance imaging (MRI) provides crucial anatomical and pathological insights, particularly through multi-modal brain and volumetric cardiac scans Dickinson et al. (2013); Lustig et al. (2007); Dayarathna et al. (2024); Manna et al. (2024). However, real-world clinical MRIs frequently suffer from missing data, including absent brain modalities and missing cardiac slices, due to limited scan times, image corruption, or protocol variations Wang et al. (2025b); Paproki et al. (2024); Zhao & Shen (2025).

To address this, generative models have been developed to infer missing data from observed inputs Ibrahim et al. (2025); Ferreira et al. (2024). Existing paradigms rely on auxiliary embeddings, e.g., binary mask codes, as prior knowledge to encode missing patterns (e.g., severity, type, and position) Liu et al. (2023); Hao et al. (2024); Kim & Park (2024); Cho et al. (2024); Meng et al. (2024). Nonetheless, these hand-crafted masks only indicate missing locations, without adequately characterizing the actual incomplete state of the input. This causes three practical limitations. First, because missing patterns vary across hospitals, scanners, and acquisition settings, enumerating them with predefined masks is unrealistic in real deployments Wang et al. (2023); Rui et al. (2025). Second, the resulting condition is insensitive to modality-specific and spatially varying context, making models less robust to unseen incomplete patterns and prone to degraded generalization Lee et al. (2023); Ke et al. (2025); Wenderoth et al. (2025); Hamamci et al. (2024); Azad et al. (2025); Pan et al. (2025). Third, because binary masks carry limited semantics, they provide rigid and insufficiently informative guidance, which can weaken spatial alignment and semantic consistency during synthesis Qiu et al. (2023); Hu et al. (2023); Shin et al. (2025); Wu et al. (2025).

Intuitively, generative models should be capable of inferring and detecting the incomplete state spontaneously, rather than relying on externally provided manual guidance Hu et al. (2023); Graikos et al. (2024). Motivated

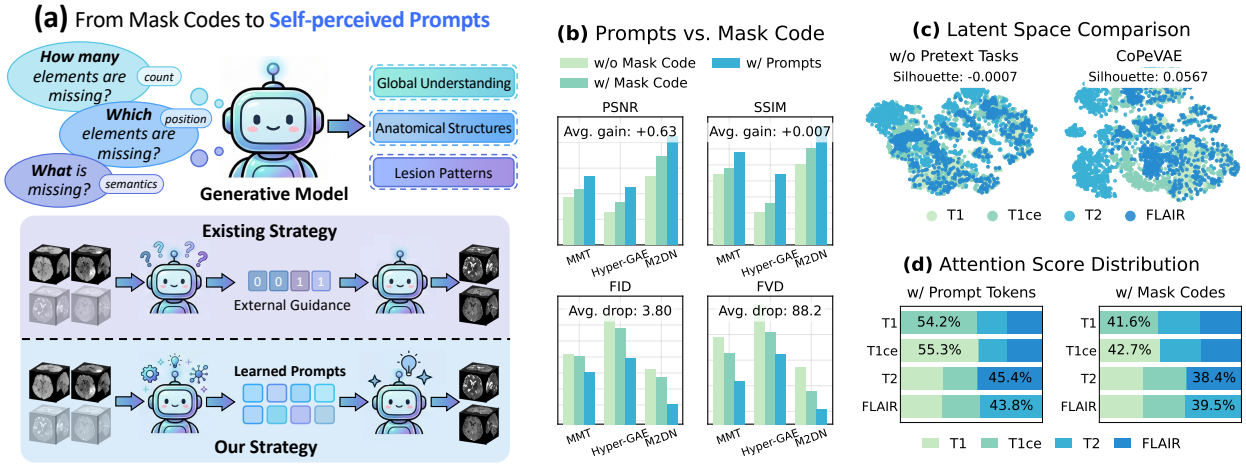

Figure 1: **Overview and advantages of CoPeDiT. (a)** We shift the paradigm from explicit mask-dependent guidance to autonomous completeness perception via prompt generation. This self-perceptive mechanism offers significant advantages: **(b)** quantitative plug-and-play performance gains across existing baselines; **(c)** substantially more discriminative latent representations compared to training without pretext tasks; and **(d)** enhanced semantic attention alignment between correlated modalities (e.g., T1-T1ce, T2-FLAIR) compared to traditional binary mask codes.

by this, we pose a central question: *Can we empower the model with the ability to perceive missing states on its own?* In light of this, we exploit an underexplored property of generative models in medical imaging, i.e., **'*completeness perception*'**, to enhance flexibility and generalizability under arbitrary missing MRI conditions, as illustrated in Fig. 1a. Our fundamental insight is to enable the generative model to recognize the fine-grained incomplete state information in a self-perceptive manner, and to leverage this understanding as internal prompts to guide the generation process. We hypothesize that, for diffusion models, such self-guided prompts may serve as an effective alternative to manually defined masks, and potentially offer even stronger guidance signals (Fig. 1b). The main reason lies in the fact that this self-perceptive strategy encourages the model to learn both global and local anatomical structures and lesion patterns at coarse and fine levels, thus enabling more semantically coherent generation of the missing MRI regions during synthesis (Figs. 1c, 1d) Liang et al. (2022).

Driven by our motivation, we propose CoPeDiT, a 3D latent diffusion model (LDM) framework for unified 3D MRI synthesis. Here, our "unified" denotes a common formulation, training paradigm, and conditioning strategy, with task-specific adaptations for different synthesis settings. Technically, our framework builds on two core components: **(i)** Unlike prior approaches that require explicit missing indicators, a novel tokenizer with a completeness perception function, CoPeVAE, is proposed to autonomously assess the integrity of modalities or volumes through tailored self-supervised pretext tasks. By detecting anatomical structures and variations in lesion patterns, CoPeVAE develops a comprehensive understanding of 3D MRIs. This enables CoPeDiT to eliminate the need for manual intervention with flexible adaptability, enhancing the method's autonomy and improving its feasibility for real-world clinical deployment with diverse missing patterns. **(ii)** A task-specific diffusion transformer instantiation, MDiT3D, is developed as a dependency-aligned conditioning interface for completeness-aware prompts in 3D MRI synthesis. Rather than introducing a fundamentally new DiT paradigm, MDiT3D adapts tokenization, attention, and prompt injection to the long-range, anisotropic, and irregular structural dependencies of volumetric MRIs. This design provides the proper pathway for our learned prompts to influence the generative process, allowing the prompt tokens to propagate observed modality- or slice-specific cues along anatomically meaningful dependencies during diffusion. As a result, MDiT3D enables more reliable synthesis of missing modalities or slices while better preserving structural consistency in high-dimensional 3D data. Incorporating the above two innovations, our architecture not only

enables adaptive self-guidance synthesis, but also demonstrates improved structural coherence and enhanced preservation of fine-grained anatomical details. Our main contributions are summarized as follows:

- We propose a unified formulation with task-specific instantiations, dubbed CoPeDiT, for both 3D brain and cardiac missing MRI synthesis under arbitrary incomplete scenarios, without the need for explicit external indicators as guidance.
- We empower our tokenizer, CoPeVAE, with a strong capacity to perceive completeness by seamlessly integrating carefully designed pretext tasks, enabling the model to recognize missing states and learn informative, self-guided prompts.
- We present MDiT3D, a tailored diffusion transformer for 3D MRI synthesis, designed as a dependency-aligned conditioning interface for our completeness-aware prompts to enable structurally meaningful prompt-guided generation.
- Extensive experiments on three datasets demonstrate that our model surpasses state-of-the-art (SOTA) methods, achieving robustness and real-world clinical applicability.

2 Related Work

Medical Image Generation. Medical image generation has been widely studied for data augmentation Hamamci et al. (2024); Zhao et al. (2025), reconstruction Yu et al. (2025); Xiao et al. (2025); Liu et al. (2024), and image completion Rassmann et al. (2025); Liu et al. (2023); Song et al. (2026) in clinical imaging workflows. Earlier methods based on Generative Adversarial Networks (GANs) improved realism but frequently faced limitations in training stability and mode collapse, restricting their fidelity, especially when scaling to complex, high-dimensional 3D volumetric data Cao et al. (2023); Shao et al. (2025); Weng et al. (2024). To overcome these bottlenecks, diffusion-based models have emerged as a robust alternative. By offering stable training dynamics and superior mode coverage, they have achieved competitive performance in medical image synthesis and are increasingly favored in conditional settings guided by partial observations, anatomical priors, or auxiliary modalities Wang et al. (2025a); Zhao et al. (2025); Zhang et al. (2026); Guo et al. (2025); Yeganeh et al. (2025).

MRI Synthesis. Recent unified MRI synthesis methods predominantly utilize generative models, ranging from GANs Xia et al. (2021); Zhang et al. (2024; 2019a;b); Yang et al. (2023); Sharma & Hamarneh (2020) and Transformers Liu et al. (2023; 2021; 2025) to advanced diffusion models Qiu et al. (2025); Yeganeh et al. (2025); Meng et al. (2024); Song et al. (2026). While these approaches successfully capture complex inter-modal dependencies to impute missing data, they inherently rely on externally provided masks to explicitly encode missing patterns in randomly incomplete scenarios Liu et al. (2023); Meng et al. (2024); Wang et al. (2023); Azad et al. (2025). This manual guidance is often rigid and lacks informative semantic details about the actual incomplete state of the input. Furthermore, because missing patterns vary widely across different clinical environments, requiring predefined masks limits practical deployment. In contrast to these mask-dependent paradigms, CoPeDiT explores the self-perceptive capability of generative models to autonomously recognize data completeness. By learning to infer missing states internally, our approach eliminates the need for manual intervention, enabling more flexible and high-fidelity MRI synthesis.

Diffusion Models. Diffusion models, particularly LDMs Rombach et al. (2022), have demonstrated remarkable capabilities across various vision tasks Ho et al. (2020); Lu et al. (2024); Ma et al. (2025); Yao et al. (2025). Recently, DiTs Peebles & Xie (2023) have emerged as powerful alternatives to traditional U-Net Ronneberger et al. (2015) backbones, achieving competitive performance in natural image synthesis. However, most existing medical image generation approaches still predominantly rely on U-Net architectures, leaving the potential of transformer-based LDMs largely underexplored in this domain Wang et al. (2025a); Nazir et al. (2025). To bridge this gap, we introduce MDiT3D, which replaces the U-Net backbone with a diffusion transformer. By incorporating task-specific architectural modifications, MDiT3D effectively captures the complex, long-range dependencies inherent in high-dimensional 3D MRIs.

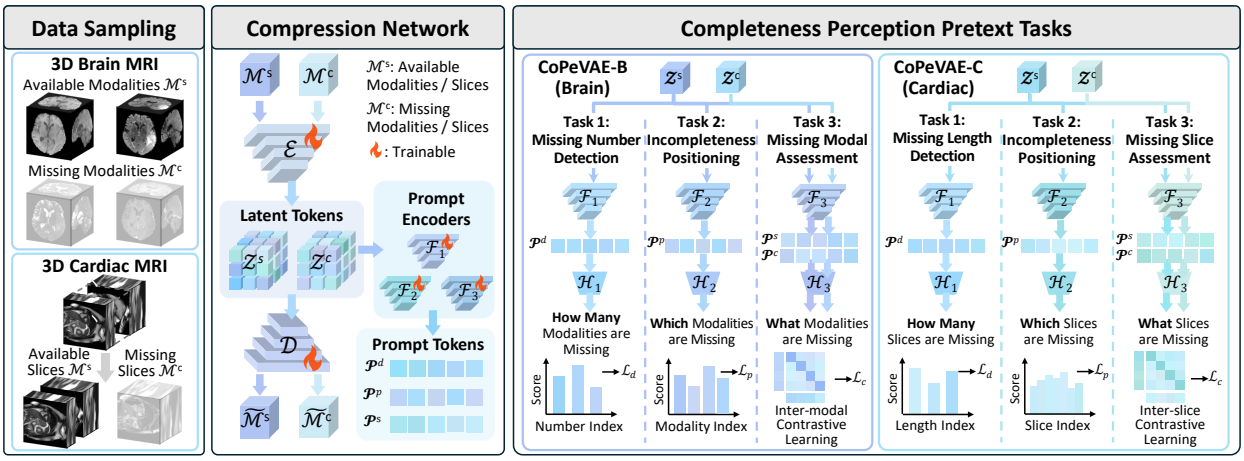

Figure 2: **The overview framework of CoPeVAE.** We implement two variants, CoPeVAE-B and CoPeVAE-C, with slight architectural modifications for brain and cardiac MRI synthesis tasks, respectively.

3 Methodology

3.1 Notations

To unify the formulation of brain missing-modality synthesis and cardiac missing-slice synthesis, let $\mathcal{M} = \{\mathbf{x}^i\}_{i=1}^m$ denote a complete 3D MRI sample comprising m elements (either modalities or slices). We partition $\mathcal{M}$ into an available subset $\mathcal{M}^S = \{\mathbf{x}^{s_i}\}_{i=1}^s$ and a missing subset $\mathcal{M}^C = \{\mathbf{x}^{c_i}\}_{i=1}^c$, where $m = s + c$. The objective is to synthesize $\mathcal{M}^C$ from $\mathcal{M}^S$. Specifically, for **brain MRIs**, $\mathcal{M}^C$ contains randomly missing modalities; for **cardiac MRIs**, $\mathcal{M}^C$ consists of consecutive missing slices. To mimic real-world clinical environments, we evaluate randomly generated incomplete cases with varying missing counts and lengths.

3.2 Stage I: Completeness Perception Tokenizer

The core idea of CoPeVAE (Fig. 2) is that detecting the completeness of high-resolution MRI data enforces the model to perceive both global anatomy and local lesion patterns, thereby producing high-quality prompts as diffusion guidance. Building upon VQGAN van den Oord et al. (2017); Esser et al. (2021), we deploy a 3D autoencoder jointly trained with self-supervised pretext tasks. Each task employs a prompt encoder, denoted as $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$, to transform latent tokens (learned by the encoder $\mathcal{E}$) into low-dimensional prompt tokens. All prompt encoders contain 3D Conv layers followed by spatial average pooling. Afterwards, task-specific projection heads are applied for multi-granular classification and contrastive learning.

Data Sampling. To improve CoPeVAE's adaptability to diverse missing cases, we employ a dual-random sampling strategy, where both the number/length of missing elements and the modality types/slice positions are randomly sampled. Given a complete set $\mathcal{M}$, we first randomly sample a missing count $c \in \{1, \ldots, m-1\}$, then uniformly select c elements to form the missing subset $\mathcal{M}^C$, with the remainder forming the incomplete subset $\mathcal{M}^S$.

Task 1. Missing Number/Length Detection. Equipped with global contextual awareness, the tokenizer is capable of determining how many modalities/slices are missing in the incomplete input. This task aims to enable the tokenizer to identify the severity of incompleteness by perceiving the global context of MRIs, thereby learning modality/spatial attributes in a coarse-grained manner. We formulate this task as an $(m-1)$-class classification task and define the loss using cross-entropy loss as follows:

$$\mathcal{L}_d = \mathcal{L}_{\text{cls}}(\mathcal{H}_1(\mathcal{F}_1(\mathbf{z}^s))_c). \tag{1}$$

where $\mathcal{F}_1$ and $\mathcal{H}_1$ represent the prompt encoder and the projection head for missing detection, respectively. The learned prompt tokens $\mathbf{p}^d = \mathcal{F}_1(\mathbf{z}^s)$ are rich in information about the severity of the missing state.

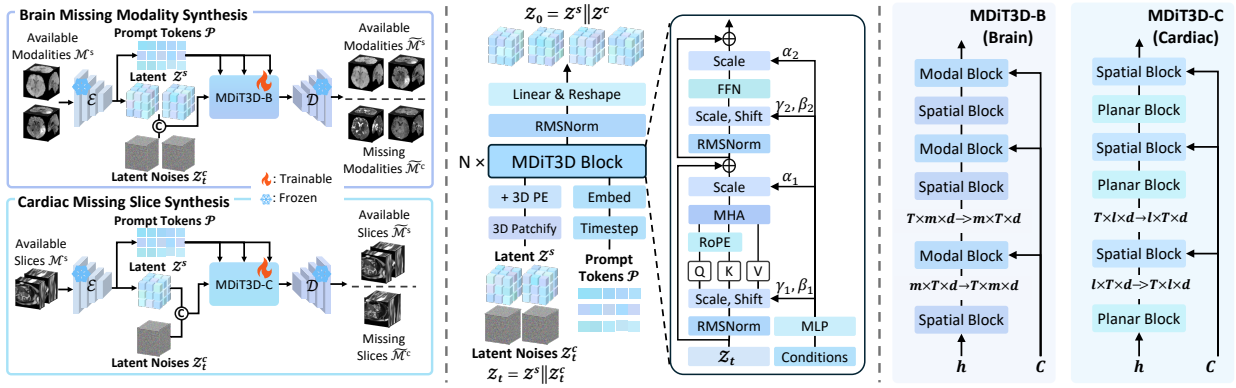

Figure 3: **Architecture of the MDiT3D framework.** We implement two variants with task-specific alternating blocks (Spatial/Modal for MDiT3D-B and Planar/Spatial for MDiT3D-C), using dynamic feature reshaping to accurately model respective anatomical dependencies. Learned prompts are injected via adaLN as conditional guidance. PE: Positional Embeddings; RoPE: Rotary Position Embeddings Su et al. (2024); Yang et al. (2025).

Task 2. Incompleteness Positioning. By identifying which modalities or slices are missing, CoPeVAE yields prompt tokens $\mathbf{p}^p = \mathcal{F}_2(\mathbf{z}^s)$ that capture semantically meaningful local properties. The motivation of this task is to drive the model to develop a finer-grained contextual understanding of subtle anatomical structures and detailed pattern variations. Although the missing position implicitly encodes the count, Task 1 learns a modality/slice-agnostic global magnitude prior that calibrates the conditioning strength, while this task provides discrete, spatially localized cues about the exact missing identity. The incorporation of the two tasks improves robustness to errors from either signal. This task is formulated as an m-class classification problem and is also optimized using the cross-entropy loss, defined as follows:

$$\mathcal{L}_p = \mathcal{L}_{\text{cls}}(\mathcal{H}_2(\mathcal{F}_2(\mathbf{z}^s))_I). \tag{2}$$

where I denotes the index of missing type/position.

Task 3. Missing Modality/Slice Assessment. Motivated by the observation that modalities or slices from the same scan share more similar anatomical and textural context than those from different scans, we adopt an inter-modal/inter-slice contrastive learning scheme Radford et al. (2021) to serve as a missing data estimator. Specifically, we take the incomplete latent tokens $\mathbf{z}^s$ as the anchor, the corresponding missing latent $\mathbf{z}^c$ from the same subject as positives, and latent tokens $\mathbf{z}^c_-$ from different subjects as negatives. The contrastive loss is defined as follows:

$$\mathcal{L}_c = -\log \frac{\varphi\left(\mathcal{H}_3(\mathbf{p}^s), \mathcal{H}_3(\mathbf{p}^c)\right)}{\varphi\left(\mathcal{H}_3(\mathbf{p}^s), \mathcal{H}_3(\mathbf{p}^c)\right) + \sum \varphi\left(\mathcal{H}_3(\mathbf{p}^s), \mathcal{H}_3(\mathbf{p}^c_-)\right)}. \tag{3}$$

where $\mathbf{p}^s = \mathcal{F}_3(\mathbf{z}^s)$, $\mathbf{p}^c = \mathcal{F}_3(\mathbf{z}^c)$, $\varphi(a, b) = \exp(a \cdot b/\tau)$, and τ is the temperature parameter. This contrastive learning scheme encourages the model to focus on inter-modal/slice contextual differences, improving anatomical coherence and fine-grained detail preservation. The overall loss of CoPeVAE is formulated as

$$\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{rec}} + \lambda(\mathcal{L}_d + \mathcal{L}_p + \mathcal{L}_c). \tag{4}$$

where λ is the weighting coefficient, and $\mathcal{L}_{\text{rec}}$ is the reconstruction loss containing a L1 loss, vector quantization loss, adversarial loss, and perceptual loss.

3.3 Stage II: 3D MRI Diffusion Transformer

We propose MDiT3D (see Fig. 3), a task-specific diffusion transformer extending DiT Peebles & Xie (2023) for volumetric MRI synthesis. MDiT3D is conditionally guided by the available latents $\mathbf{z}^s$ alongside the concatenated completeness-aware prompts $\mathbf{p} = \mathbf{p}^d \,\|\, \mathbf{p}^p \,\|\, \mathbf{p}^s$. During inference, these prompts are autonomously

Table 1: **Quantitative results for multi-modal brain MRI synthesis on the BraTS dataset.** The numbers in the first row denote the number of missing modalities.

	1				2				3			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
GAN-based Methods												
MMGAN Sharma & Hamarneh (2020)	24.71 ±1.57	0.817 ±0.027	27.94	489.86	24.38 ±1.74	0.806 ±0.023	32.48	726.93	24.06 ±1.92	0.794 ±0.031	39.37	898.17
MMT Liu et al. (2023)	25.19 ±1.41	0.824 ±0.017	24.53	527.06	24.50 ±1.55	0.811 ±0.020	29.57	686.52	23.92 ±1.53	0.801 ±0.021	39.66	841.46
Hyper-GAE Yang et al. (2023)	24.65 ±1.62	0.813 ±0.022	28.97	609.65	24.42 ±1.74	0.808 ±0.024	33.52	815.92	23.86 ±1.88	0.788 ±0.029	41.79	948.27
Diffusion Model-based Methods												
LDM Rombach et al. (2022)	23.84 ±1.52	0.805 ±0.019	36.47	740.26	23.12 ±1.64	0.798 ±0.021	45.93	897.83	22.65 ±1.77	0.791 ±0.025	52.50	1161.21
ControlNet Zhang et al. (2023)	23.98 ±1.49	0.808 ±0.018	34.09	806.82	23.34 ±1.60	0.801 ±0.020	41.28	986.30	22.85 ±1.68	0.795 ±0.022	48.07	1074.23
M2DN Meng et al. (2024)	26.45 ±1.36	0.830 ±0.016	21.29	376.53	25.87 ±1.48	0.820 ±0.017	27.36	467.66	25.21 ±1.59	0.809 ±0.024	32.40	553.16
DiffM^{4}RI Ye et al. (2026)	26.07 ±1.57	0.824 ±0.024	25.03	449.17	25.49 ±1.41	0.813 ±0.025	32.59	591.35	25.08 ±1.75	0.806 ±0.029	35.91	695.06
CoPeDiT	**28.26** ±1.24	**0.842** ±0.019	**12.67**	**254.71**	**28.13** ±1.49	**0.831** ±0.021	**13.25**	**287.58**	**27.91** ±1.41	**0.822** ±0.023	**14.89**	**323.19**

Table 2: **Quantitative results for multi-modal brain MRI synthesis on the IXI dataset**. The numbers in the first row denote the number of missing modalities.

	1				2			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
GAN-based Methods								
MMGAN Sharma & Hamarneh (2020)	22.29 ±1.35	0.684 ±0.015	70.91	1447.83	21.13 ±1.49	0.668 ±0.019	93.57	1787.39
MMT Liu et al. (2023)	22.64 ±1.49	0.698 ±0.013	53.60	1329.25	21.82 ±1.60	0.687 ±0.019	72.24	1562.44
Hyper-GAE Yang et al. (2023)	22.12 ±1.33	0.682 ±0.017	72.62	1520.49	20.91 ±1.45	0.662 ±0.021	98.79	1712.90
Diffusion Model-based Methods								
LDM Rombach et al. (2022)	21.36 ±1.28	0.679 ±0.016	86.62	1884.64	20.94 ±1.53	0.654 ±0.020	112.57	2314.39
ControlNet Zhang et al. (2023)	21.83 ±1.51	0.681 ±0.015	81.26	1971.40	21.07 ±1.50	0.661 ±0.019	103.77	2386.04
M2DN Meng et al. (2024)	23.47 ±1.43	0.715 ±0.014	42.52	845.29	22.81 ±1.56	0.702 ±0.018	55.64	1078.36
DiffM^{4}RI Ye et al. (2026)	23.71 ±1.26	0.720 ±0.018	39.87	715.63	22.87 ±1.68	0.707 ±0.022	51.93	964.28
CoPeDiT	**24.34** ±1.21	**0.732** ±0.016	**25.84**	**569.22**	**23.92** ±1.53	**0.721** ±0.020	**32.53**	**718.54**

extracted by the frozen CoPeVAE to provide semantic guidance: $\mathbf{p}^d$ calibrates global severity (how many), $\mathbf{p}^p$ localizes missing regions (where), and $\mathbf{p}^s$ supplies fine-grained textural priors (what). To effectively adapt to 3D medical data, we introduce customized operations (using MDiT3D-B as a representative example). Specifically, we apply a 3D patchify operator to project inputs into tokens $\mathbf{h} \in \mathbb{R}^{m \times T \times d}$ Mo et al. (2023), where m is the number of modalities/slices and d is the embedding dimension. Unlike standard 2D implementations Dosovitskiy et al. (2020), we incorporate 3D frequency-based sine-cosine positional embeddings (3D PE) to preserve precise spatial relationships.

Alternating Blocks & Prompt Injection. We design two distinct alternating block architectures tailored to the specific characteristics of brain and cardiac MRIs. For the multi-modal brain task, we alternate between Spatial Blocks (to capture 3D spatial context) and Modal Blocks (to model inter-modal relationships). For the volumetric cardiac task, we alternate between Planar Blocks (for intra-slice features) and Spatial Blocks (for through-plane continuity). To compute attention along the appropriate axes, the latent tokens are dynamically reshaped before entering each block (e.g., aligning the modal dimension to facilitate inter-modal interaction, and then reshaped back for spatial processing). Furthermore, to maximize the efficacy of conditional guidance, prompts are injected via adaptive layer normalization (adaLN) Peebles & Xie (2023) exclusively into the blocks that explicitly model the task's primary dependency. Specifically, prompt tokens are injected only into the Modal Blocks for brain MRI (addressing missing modalities) and the Spatial Blocks for cardiac MRI (addressing missing slices). This targeted injection avoids overwhelming the network and ensures the conditioning signals are both physically meaningful and highly informative.

Joint Reconstruction & Synthesis. During diffusion, rather than filling missing modalities or slices with zeros or learned tokens, we only add noise to the missing sections, keeping the available latents unperturbed to provide rich contextual guidance. Following Meng et al. (2024), MDiT3D is optimized via an $\mathbf{x}_0$-prediction loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t}\left[\|\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t, \mathbf{p})\|_2^2\right], \tag{5}$$

Table 3: **Quantitative results for Cardiac MRI synthesis on the UKBB dataset.** The numbers in the first row denote the length of missing slices.

	8				16				24			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
GAN-based Methods												
MMGAN Sharma & Hamarneh (2020)	25.81 ±0.86	0.815 ±0.012	16.68	392.48	24.35 ±0.84	0.793 ±0.009	27.52	604.37	23.06 ±1.08	0.776 ±0.018	45.88	767.92
MMT Liu et al. (2023)	26.02 ±0.82	0.824 ±0.009	15.90	429.72	24.73 ±0.87	0.809 ±0.010	23.75	570.29	24.12 ±1.06	0.794 ±0.017	37.39	708.68
Hyper-GAE Yang et al. (2023)	25.23 ±0.89	0.810 ±0.012	19.02	517.55	23.66 ±0.94	0.789 ±0.013	31.84	648.35	22.70 ±1.14	0.771 ±0.019	48.27	845.14
Diffusion Model-based Methods												
LDM Rombach et al. (2022)	24.17 ±0.96	0.795 ±0.012	24.61	704.91	23.04 ±1.03	0.778 ±0.010	44.93	830.73	22.19 ±1.21	0.761 ±0.018	60.02	910.15
ControlNet Zhang et al. (2023)	24.62 ±0.91	0.801 ±0.012	22.47	698.63	23.30 ±0.98	0.784 ±0.014	37.51	821.40	22.46 ±1.16	0.765 ±0.020	54.93	922.06
M2DN Meng et al. (2024)	25.48 ±0.79	0.814 ±0.011	17.72	453.03	24.62 ±0.87	0.803 ±0.011	24.15	597.78	24.03 ±0.99	0.780 ±0.021	40.08	823.70
DiffM^{4}RI Ye et al. (2026)	25.19 ±0.89	0.808 ±0.014	19.16	495.23	24.78 ±0.80	0.807 ±0.015	22.96	548.44	24.27 ±1.06	0.791 ±0.018	35.27	674.12
CoPeDiT	**26.42** ±0.81	**0.831** ±0.013	**15.53**	**318.62**	**26.07** ±0.74	**0.826** ±0.013	**18.21**	**382.04**	**25.39** ±0.86	**0.817** ±0.016	**25.84**	**490.57**

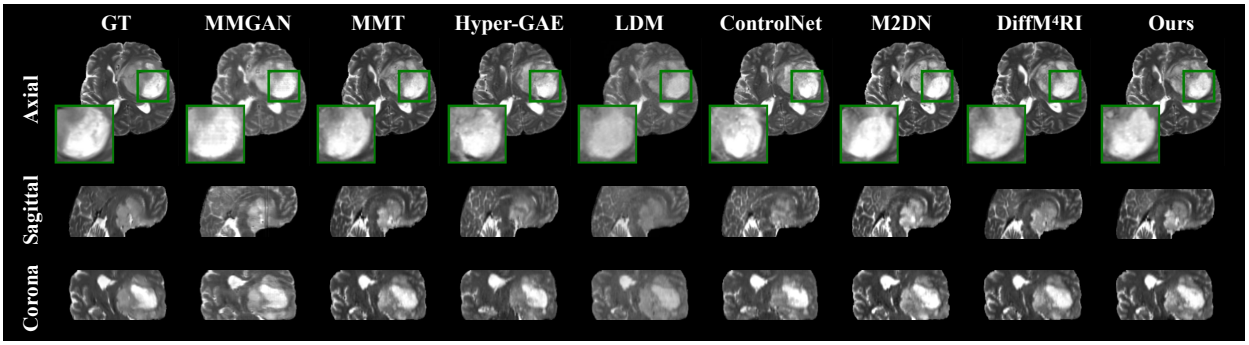

Figure 4: **Qualitative results on the BraTS dataset.** The visual results on the IXI dataset are provided in Appendix C.1.

where $\mathbf{z}_0 = \mathbf{z}^s \,\|\, \mathbf{z}^c$ is the clean target, $\mathbf{z}_t = \mathbf{z}^s \,\|\, \mathbf{z}_t^c$ is the partially noised input at timestep t, $\mathbf{p}$ represents the prompt tokens, and f_θ denotes MDiT3D.

4 Experiments and Results

4.1 Experimental Setup

Datasets. (i) Brain MRI Datasets. We evaluate the effectiveness of CoPeDiT on two public brain MRI datasets: BraTS 2021 Baid et al. (2021) and IXI Brain Development Project (2025). The BraTS 2021 dataset includes 1251 subjects with multi-modal MRI scans across four modalities: T1, T1ce, T2, and FLAIR. The IXI dataset contains 577 subjects with three MRI modalities: T1, T2, and PD. **(ii) Cardiac MRI Datasets.** Missing slice synthesis experiments are conducted on four cardiac MRI datasets: UK Biobank (UKBB) Petersen et al. (2016), MESA Zhang et al. (2018), ACDC Bernard et al. (2018), and MSCMR Zhuang et al. (2022). The model is trained on the combined dataset including all four sources with 32,248 MRI volumes in total, while performance comparisons are conducted on the UKBB dataset. We randomly select 80% of the data for training and use the remaining 20% as the test set. Please refer to Appendix A for more details.

Implementation Details. The compression rate of CoPeVAE is set to $(8, 8, 8)$. For hyperparameters, the dimension of each prompt token is set to 512, τ and λ are set to 0.2 and `1e-2`, respectively. The model is trained with a global batch size of 8 for CoPeVAE-B and 64 for CoPeVAE-C, using a learning rate of `1e-4`. Regarding MDiT3D, we set the number of time steps to 500 with linearly scaled noise scheduling. The model is trained for 100k steps with a global batch size of 32 for MDiT3D-B and 64 for MDiT3D-C, with learning rate of `5e-5`. All training is conducted on four NVIDIA A100 GPUs. More details are provided in Appendix B.

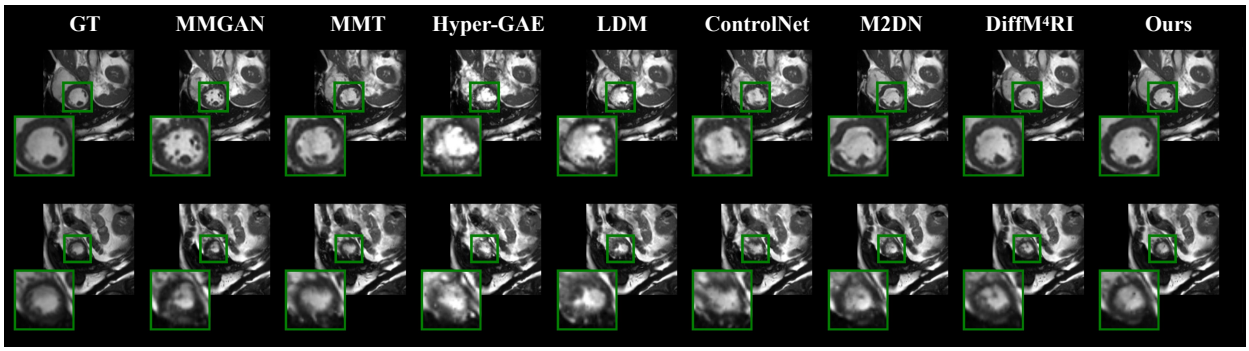

Figure 5: **Qualitative results on the UKBB dataset.** The top and bottom results correspond to the first and last missing slices within a given volume, respectively.

Table 4: Ablation study on the contribution of **completeness-aware prompts**.

	BraTS												UKBB							
	1				2				3				8				24			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
w/o $\mathbf{p}^d$	27.35	0.833	16.04	335.08	27.02	0.824	17.30	475.32	26.73	0.814	19.62	556.79	25.82	0.819	16.34	427.06	24.53	0.803	33.80	610.46
w/o $\mathbf{p}^p$	26.92	0.829	18.46	372.36	26.43	0.819	20.23	460.45	26.06	0.810	25.13	572.03	25.17	0.812	20.08	534.80	24.39	0.798	36.54	697.51
w/o $\mathbf{p}^s$	27.56	0.835	15.26	341.70	27.22	0.827	16.37	395.43	26.90	0.815	17.58	468.23	26.27	0.828	16.79	388.46	25.08	0.813	29.83	578.13
w/o Prompts	25.92	0.823	25.69	418.12	25.06	0.807	32.17	556.92	24.83	0.802	37.97	748.25	24.70	0.797	23.29	721.35	23.56	0.778	42.17	830.72
w/ Mask Codes	27.18	0.831	17.42	356.09	26.82	0.816	20.06	439.13	26.18	0.809	24.59	543.82	26.15	0.823	18.15	409.47	24.65	0.802	35.86	612.73
CoPeDiT	**28.26**	**0.842**	**12.67**	**254.71**	**28.13**	**0.831**	**13.25**	**287.58**	**27.91**	**0.822**	**14.89**	**323.19**	**26.42**	**0.831**	**15.53**	**318.62**	**25.39**	**0.817**	**25.84**	**490.57**

4.2 Performance Comparison

We compare our method against seven SOTA baselines, including three GAN-based approaches Sharma & Hamarneh (2020); Liu et al. (2023); Yang et al. (2023) and four diffusion-based models Rombach et al. (2022); Zhang et al. (2023); Meng et al. (2024); Ye et al. (2026), all reimplemented under identical settings for fair comparison. Performance is evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Fréchet Inception Distance (FID) Heusel et al. (2017), and Fréchet Video Distance (FVD) Unterthiner et al. (2018) for assessing 3D spatial consistency.

Quantitative Results. Quantitative results on the three datasets are presented in Tables 1, 2 and 3, respectively. CoPeDiT consistently outperforms all baselines across all missing configurations in both synthesis tasks. Notably, these performance gains widen in scenarios with a higher number of missing modalities or slices, such as maintaining a high 27.91 PSNR even with three missing modalities. This highlights the robustness of our completeness-aware prompts in complex cases. Moreover, CoPeDiT achieves substantially lower FID and FVD scores (e.g., an FVD of 490.57 for 24 missing cardiac slices). This indicates that the generated MRIs are not only anatomically coherent and texture-preserving, but exhibit superior 3D spatial consistency and structural continuity, enhancing perceptual realism and diagnostic plausibility.

Qualitative Results. As depicted in Figs. 4 and 5, our model generates synthetic MRIs that exhibit the highest visual similarity to the ground truth images, particularly in accurately capturing tumor regions. Our CoPeDiT excels at preserving subtle textural details and modeling complex anatomical structures within brain tissues, justifying our motivation that incorporating completeness perception leads to improved anatomical consistency and realism.

4.3 Ablation Study

Effect of Prompt Tokens. Table 4 compares learned prompt tokens against one-hot mask codes. Under identical injection strategies, all individual prompts and their combinations consistently outperform binary masks, strongly validating our completeness perception approach. Specifically, the positioning prompt (p^p) proves most effective; its removal triggers the sharpest performance drop, likely because explicitly locating

Table 5: Quantitative results by **incorporating our learned prompt tokens into baselines** instead of mask codes.

	PSNR↑	SSIM↑	FID↓	FVD↓
MMT Liu et al. (2023)	25.19	0.824	24.53	527.06
+ Prompts (ours)	25.68 (+0.49)	0.826 (+0.002)	22.07 (–2.46)	416.34 (–110.72)
Hyper-GAE Yang et al. (2023)	24.65	0.813	28.97	609.65
+ Prompts (ours)	25.26 (+0.61)	0.822 (+0.009)	24.26 (–4.71)	523.75 (–85.90)
M2DN Meng et al. (2024)	26.45	0.830	21.29	376.53
+ Prompts (ours)	27.23 (+0.78)	0.837 (+0.007)	17.06 (–4.23)	308.68 (–67.85)

Table 6: Ablation study of each pretext task's contribution to **CoPeVAE's reconstruction capacity**. 'Cardiac' refers to the combined dataset.

	BraTS		**IXI**		**Cardiac**	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o Task 1	34.38	0.926	30.92	0.914	32.34	0.921
w/o Task 2	33.62	0.918	30.35	0.908	31.59	0.914
w/o Task 3	34.69	0.929	31.13	0.917	32.62	0.925
CoPeVAE	**35.05**	**0.935**	**31.28**	**0.921**	**33.34**	**0.931**

Table 7: Ablation study on the choice of **prompt injection positions within the MDiT3D blocks.** The labels denote the target blocks for brain and cardiac tasks, respectively (e.g., "Modal / Spatial" means injecting prompts exclusively into the modal blocks for brain MRIs and spatial blocks for cardiac MRIs). "Both" indicates injection into both types of blocks for the respective tasks.

	BraTS												**UKBB**							
	1				2				3				8				24			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
Spatial / Planar	27.47	0.834	15.41	317.62	27.20	0.825	16.83	484.39	27.03	0.817	17.45	592.17	26.29	0.827	16.82	405.72	25.22	0.814	29.24	604.32
Both	27.76	0.836	13.94	294.37	27.54	0.827	15.48	367.20	27.29	0.818	16.77	434.83	26.33	0.827	16.39	414.37	25.28	0.815	28.45	587.13
Modal / Spatial	**28.26**	**0.842**	**12.67**	**254.71**	**28.13**	**0.831**	**13.25**	**287.58**	**27.91**	**0.822**	**14.89**	**323.19**	**26.42**	**0.831**	**15.53**	**318.62**	**25.39**	**0.817**	**25.84**	**490.57**

missing sections heightens sensitivity to subtle structural variations. To further explore the potential of our learned prompts, we incorporate them into prior baselines as replacements for mask codes. As illustrated in Table 5, this substitution consistently boosts their performance, such as yielding a 0.78 dB PSNR gain for M2DN, demonstrating the strong plug-and-play utility and generalizability of our method.

Effect of Pretext Tasks. As Table 6 displays, CoPeVAE preserves strong reconstruction capability while benefiting from the incorporation of pretext tasks. Each task contributes positively, and their combination leads to further improvements. This improvement can be attributed to the fact that pretext tasks promote the tokenizer to capture anatomical structure variances in both coarse and fine-grained manner, learning highly discriminative features.

Choice of Prompt Injection Position. We aim to justify our choice of injecting prompts exclusively into the modal and spatial blocks for brain and cardiac tasks, respectively. As shown in Table 7, injecting prompts into the modal/spatial block yields the best performance, whereas injecting into spatial/planar blocks or both blocks results in inferior outcomes. This can be attributed to the fact that aligning the prompt with the block that explicitly models the task's primary dependency, namely modality fusion for brain and through-plane continuity for cardiac, maximizes the efficacy of conditional signals.

Impact of 3D MRI Diffusion Transformer. The evaluation of MDiT3D and existing diffusers Rombach et al. (2022); Peebles & Xie (2023); Mo et al. (2023) is presented in Table 8. MDiT3D achieves superior performance, notably outperforming vanilla 3D DiTs by over 1.4 dB in PSNR on the BraTS dataset. This confirms that our task-driven design, which couples alternating blocks with targeted prompt injection to explicitly model inter-modal relationships and through-plane continuity, effectively improves synthesis.

4.4 Tumor Segmentation

To evaluate clinical utility, we conduct downstream tumor segmentation on the BraTS dataset. Following Liu et al. (2023); Meng et al. (2024), we train a multi-modal U-Net Ronneberger et al. (2015) using three available modalities alongside one modality synthesized by CoPeDiT or baselines. We also evaluate a "Missing" baseline trained solely on the three available modalities. As shown in Table 9, all synthesis methods improve upon the "Missing" baseline in terms of average Dice scores for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). Notably, CoPeDiT consistently outperforms competing methods across all subregions,

Table 8: Ablation study on the contribution of **3D MRI Diffusion Transformer.**

	BraTS												UKBB							
	1				2				3				8				24			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
UNet Rombach et al. (2022)	25.61	0.825	23.56	433.55	25.20	0.816	28.37	545.89	24.78	0.805	36.70	707.83	24.46	0.800	22.34	689.53	23.01	0.778	41.59	871.64
DiT Peebles & Xie (2023)	26.78	0.835	19.72	392.09	26.14	0.823	21.84	498.13	25.74	0.808	24.47	611.26	25.89	0.825	16.28	403.83	24.78	0.799	36.10	592.72
DiT-3D Mo et al. (2023)	26.84	0.835	19.23	359.83	26.23	0.822	21.95	487.12	25.90	0.810	26.08	593.42	26.08	0.825	16.37	397.71	24.83	0.806	33.73	610.89
MDiT3D	**28.26**	**0.842**	**12.67**	**254.71**	**28.13**	**0.831**	**13.25**	**287.58**	**27.91**	**0.822**	**14.89**	**323.19**	**26.42**	**0.831**	**15.53**	**318.62**	**25.39**	**0.817**	**25.84**	**490.57**

Table 9: Results of **tumor segmentation** experiments on the BraTS dataset.

	Dice Score (%)↑			
	WT	TC	ET	AVG
Missing	86.08	84.67	81.59	84.11
GAN-based Methods				
MMGAN Sharma & Hamarneh (2020)	89.35	88.14	87.73	88.41
MMT Liu et al. (2023)	90.43	88.37	86.92	88.57
Hyper-GAE Yang et al. (2023)	88.72	86.54	85.37	86.88
Diffusion Model-based Methods				
LDM Rombach et al. (2022)	87.86	85.91	84.19	85.99
ControlNet Zhang et al. (2023)	88.27	87.05	85.23	86.85
M2DN Meng et al. (2024)	91.28	90.09	88.20	89.86
DiffM^{4}RI Ye et al. (2026)	90.04	89.23	87.68	88.72
CoPeDiT	**91.35**	**90.41**	**88.94**	**90.23**

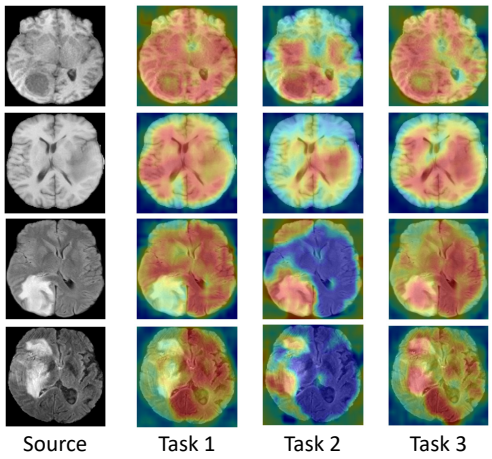

Figure 6: Visualization of **salient regions** on the BraTS dataset.

achieving the highest average Dice of 90.23%. These results confirm that our synthesized MRIs provide highly informative and clinically valuable inputs for downstream tasks.

4.5 Computational Efficiency Analysis

Training Compute and Wall-Clock Cost. As shown in Table 10, our CoPeVAE adds minimal overhead in parameters, FLOPs, and training time compared to variants. This modest cost yields superior synthesis quality and spatial consistency, proving a highly efficient complexity-gain trade-off for our pretext tasks.

Inference Cost. To assess practical applicability, Table 11 reports end-to-end latency and VRAM usage per volume under standardized settings (batch size = 1, single A100 GPU, DDIM Song et al. (2021) 200 steps, mixed precision). CoPeDiT remains efficient, with resource demands comparable to standard models and markedly lower than heavier baselines, supporting its suitability for clinical deployment.

4.6 Visualization Analysis

Salient Regions. To understand the learning procedure of pretext tasks, we visualize their activation maps using GradCAM Selvaraju et al. (2017). As shown in Fig. 6, the examples reveal strong correlations between salient regions and modality-discriminative features. These patterns mirror each task's objective: Task 1 and Task 3 must assess global consistency, so they rely on coarse anatomical layout (gray and white matter) that summarize the volume. In contrast, Task 2 needs to pinpoint where/which is missing, so it keys on high-frequency, modality-specific cues (tumors, lesions, and white matter hyperintensities in FLAIR). The results emphasize that our pretext tasks capture modality-specific properties and improve the model's understanding of MRI context and inter-modality relationships.

Prompt Distribution. We implement the t-SNE visualization of learned prompts on BraTS and color them by ground-truth incompleteness labels. As depicted in Figs. 7a and 7b, the count-focused prompt $\mathbf{p}^d$ forms compact, well-separated clusters aligned with the missing number classes corresponding to each missing state

Table 10: The **computational cost and wall-clock time comparison** of CoPeVAE on the BraTS dataset.

	Param. (M)	Flops (G)	Epoch Time (s)	Total Time (GPU hours)	BraTS (Avg)	
					PSNR↑	FVD↓
w/o Task 1	136.77	21167.61	331.3	561.2	27.03	455.73
w/o Task 2	136.55	21171.83	332.9	563.9	26.47	468.13
w/o Task 3	136.74	21171.83	330.7	560.1	27.23	401.79
CoPeVAE	137.63	21184.57	336.7	570.3	**28.10**	**288.49**

Table 11: The **inference cost** comparison of CoPeDiT on the BraTS dataset.

	Inference Latency (s)	VRAM (G)
LDM Rombach et al. (2022)	1.38	0.429
M2DN Meng et al. (2024)	4.67	0.819
DiffM^{4}RI Ye et al. (2026)	2.16	0.584
DiT-3D Mo et al. (2023)	1.65	0.614
CoPeDiT	1.73	0.626

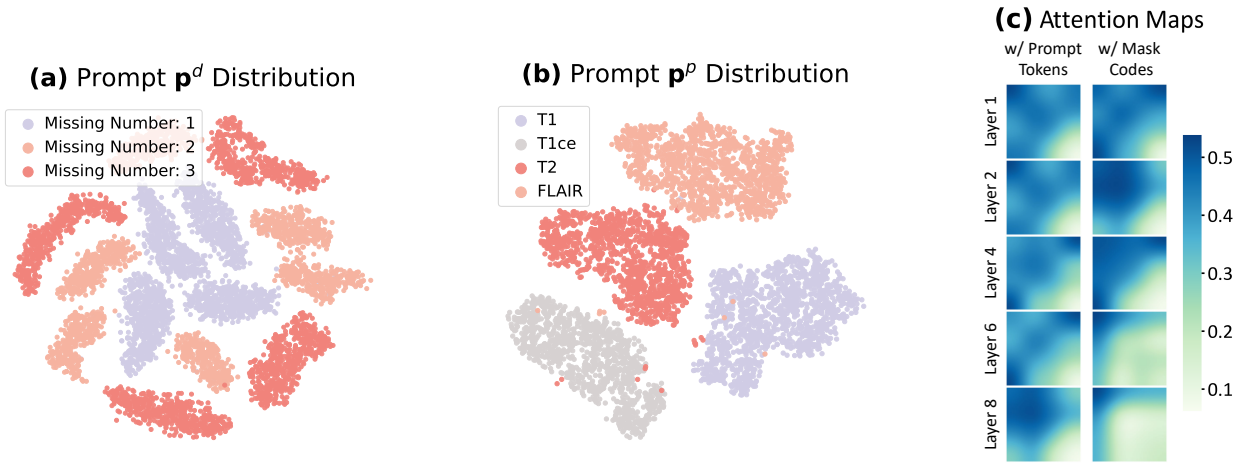

Figure 7: **Qualitative analysis of the learned prompts** on the BraTS dataset. **(a)** t-SNE projection of the count-focused prompt $\mathbf{p}^d$, showing clear clustering by the number of missing modalities. **(b)** t-SNE projection of the identity-focused prompt $\mathbf{p}^p$, revealing distinct separation by missing modality types. **(c)** Visualization of modal block attention maps under a missing configuration of [1, 0, 0, 0].

(1/2/3 absent modalities). Meanwhile, the identity-focused prompt $\mathbf{p}^p$ produces modality-specific clusters (T1, T1ce, T2, FLAIR) with distinct boundaries. This clear separation visually confirms that our pretext tasks effectively drive the tokenizer to learn highly discriminative features. Together, they demonstrate that our learned prompt tokens capture explicitly decoupled and complementary semantic priors. By avoiding feature entanglement, these prompts successfully provide the network with nuanced, multi-granular guidance for reliable synthesis.

Attention Maps. We visualize the attention maps of each modal block under a specific missing configuration (e.g., [1, 0, 0, 0]). As shown in Fig. 7c, our learned prompts actively guide the attention mechanism to progressively focus on the actual missing elements (the first row and column) as the block depth increases. This concentrated focus facilitates precise, layer-wise feature aggregation for the absent modality. In comparison, explicit mask codes lack sufficient informativeness, yielding diffuse attention weights that fail to provide the nuanced, dynamic guidance necessary to align with the true missing state.

4.7 Analyzing Completeness-Aware Prompts

To quantitatively validate the effectiveness and rationale of our learned prompts, we conduct comprehensive sensitivity and intervention analyses.

Prompt Accuracy Sensitivity. To assess CoPeDiT's reliance on prompt accuracy, we randomly replace the degree ($\mathbf{p}^d$), position ($\mathbf{p}^p$), semantic ($\mathbf{p}^s$), or "All" prompts with incorrect ones for an r-fraction ($r \in [0, 1.0]$) of BraTS validation samples. Fig. 8a demonstrates that synthesis quality degrades consistently as the

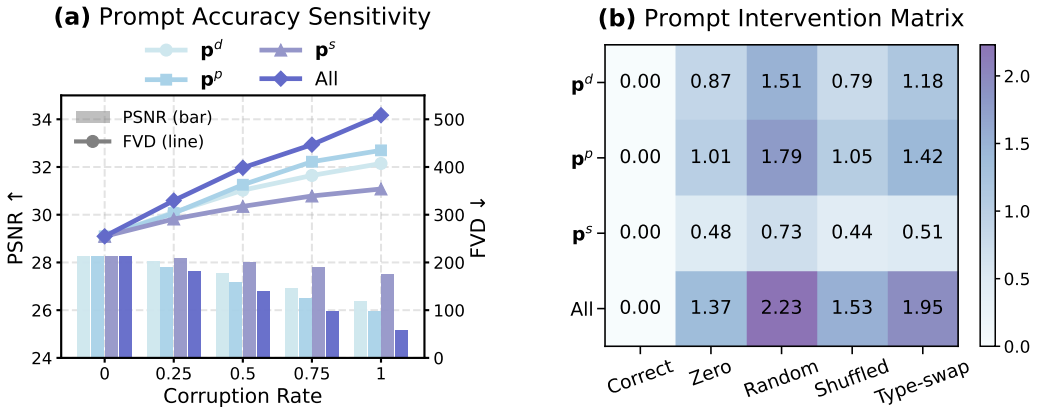

Figure 8: **Quantitative analysis of completeness-aware prompts** on the BraTS dataset. **(a)** Sensitivity of CoPeDiT to increasing prompt corruption rates. Synthesis quality degrades monotonically, particularly when perturbing the position prompt $\mathbf{p}^p$. **(b)** Prompt intervention matrix illustrating the PSNR drop under structured perturbations. Misleading signals (e.g., random, type-swap) and the simultaneous perturbation of all prompts cause the most severe performance degradation.

corruption rate r increases. Corrupting all prompts simultaneously yields the worst outcomes, confirming their complementary and essential roles.

Prompt Intervention. We further apply structured interventions by replacing prompt components with zeroed, random, shuffled, or type-swapped variants. As Fig. 8b illustrates, misleading signals (random/type-swapped) degrade performance more severely than missing or mildly perturbed guidance (zeroed/shuffled). Most notably, across both sensitivity and intervention experiments, perturbing $\mathbf{p}^p$ consistently triggers the sharpest performance drops. This combined evidence robustly demonstrates that precise missing region localization ($\mathbf{p}^p$) is the most crucial prompt for guiding high-fidelity 3D MRI synthesis.

5 Conclusion

This work presents CoPeDiT, a unified model for 3D brain and cardiac MRI synthesis that explores completeness perception. We demonstrate that enabling the model to autonomously infer the missing state, rather than relying on externally pre-defined masks, can provide more discriminative and informative guidance. To this end, we equip our tokenizer with completeness perception capability through carefully designed pretext tasks. MDiT3D is then developed to utilize the learned prompt tokens as guidance for 3D MRI generation. Extensive evaluations validate CoPeDiT's remarkable accuracy and robustness across diverse scenarios, highlighting its potential for practical clinical deployment.

Limitations. While highly effective, CoPeDiT requires a fixed number of modalities during training and may lose some fine high-frequency details due to latent space compression. Future work will focus on modality-agnostic tokenizers and exploring pixel-space diffusion refinement.

References

Reza Azad, Mohammad Dehghanmanshadi, Nika Khosravi, Julien Cohen-Adad, and Dorit Merhof. Addressing missing modality challenges in mri images: A comprehensive review. *Computational Visual Media*, 11 (2):241–268, 2025.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021

benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

Imperial College London Brain Development Project. Ixi dataset (information extraction from images), 2025. Accessed 2025-07-28.

Bing Cao, Zhiwei Bi, Qinghua Hu, Han Zhang, Nannan Wang, Xinbo Gao, and Dinggang Shen. Autoencoder-driven multimodal collaborative learning for medical image synthesis. *International Journal of Computer Vision*, 131(8):1995–2014, 2023.

Jihoon Cho, Jonghye Woo, and Jinah Park. A unified framework for synthesizing multisequence brain mri via hybrid fusion. *arXiv preprint arXiv:2406.14954*, 2024.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical Image Analysis*, 92:103046, 2024.

Louise Dickinson, Hashim U Ahmed, Clare Allen, Jelle O Barentsz, Brendan Carey, Jurgen J Futterer, Stijn W Heijmink, Peter Hoskin, Alex P Kirkham, Anwar R Padhani, et al. Clinical applications of multiparametric mri within the prostate cancer diagnostic pathway. *Urologic oncology*, 31(3):281, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

André Ferreira, Jianning Li, Kelsey L. Pomykala, Jens Kleesiek, Victor Alves, and Jan Egger. Gan-based generation of realistic 3d volumetric data: A systematic review and taxonomy. *Medical Image Analysis*, 93:103100, 2024.

Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8532–8542, 2024.

Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro RAS Bassi, Zongwei Zhou, Benjamin D Simon, Stephanie Anne Harmon, et al. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025.

Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *ECCV*, pp. 126–143, 2024.

Huaibo Hao, Jie Xue, Pu Huang, Liwen Ren, and Dengwang Li. Qgformer: Queries-guided transformer for flexible medical image synthesis with domain missing. *Expert Systems with Applications*, 247:123318, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Vincent Tao Hu, David W. Zhang, Yuki M. Asano, Gertjan J. Burghouts, and Cees G. M. Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18413–18422, 2023.

Mahmoud Ibrahim, Yasmina Al Khalil, Sina Amirrajab, Chang Sun, Marcel Breeuwer, Josien Pluim, Bart Elen, Gökhan Ertaylan, and Michel Dumontier. Generative ai for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, 189:109834, 2025.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873, 2025.

Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *WACV*, pp. 7604–7613, 2024.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023.

Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *AAAI*, volume 36, pp. 1564–1573, 2022.

Jiang Liu, Srivathsa Pasumarthi, Ben Duffy, Enhao Gong, Keshav Datta, and Greg Zaharchuk. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE Transactions on Medical Imaging*, 42(9):2577–2591, 2023.

Junkai Liu, Nay Aung, Theodoros N Arvanitis, Stefan K Piechnik, Joao AC Lima, Steffen E Petersen, and Le Zhang. Sagcnet: Spatial-aware graph completion network for missing slice imputation in population cmr imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 457–466. Springer, 2025.

Xuhui Liu, Zhi Qiao, Runkun Liu, Hong Li, Juan Zhang, Xiantong Zhen, Zhen Qian, and Baochang Zhang. Diffux2ct: Diffusion learning to reconstruct ct images from biplanar x-rays. In *European conference on computer vision*, pp. 458–476, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. VDT: General-purpose video diffusion transformers via mask modeling. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael Lustig, David Donoho, and John M. Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Siladittya Manna, Saumik Bhattacharya, and Umapada Pal. Self-supervised visual representation learning for medical image analysis: A comprehensive survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Survey Certification.

Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 43(7):2587–2598, 2024.

Shentong Mo, Enze Xie, Ruihang Chu, Lanqing Hong, Matthias Niessner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation. In *Neural Discrete Representation Learning*, volume 36, pp. 67960–67971, 2023.

Maham Nazir, Muhammad Aqeel, and Francesco Setti. Diffusion-based data augmentation for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1330–1339, 2025.

Tan Pan, Zhaorui Tan, Kaiyu Guo, Dongli Xu, Weidi Xu, Chen Jiang, Xin Guo, Yuan Qi, and Yuan Cheng. Structure-aware semantic discrepancy and consistency for 3d medical image self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20257–20267, 2025.

Anthony Paproki, Olivier Salvado, and Clinton Fookes. Synthetic data for deep learning in computer vision & medical imaging: A means to reduce data bias. *ACM Computing Surveys*, 56(11), 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Steffen E. Petersen, Paul M. Matthews, Jane M. Francis, Matthew D. Robson, Filip Zemrak, Redha Boubertakh, Alistair A. Young, Sarah Hudson, Peter Weale, Steve Garratt, Rory Collins, Stefan Piechnik, and Stefan Neubauer. Uk biobank's cardiovascular magnetic resonance protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, 2016.

Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15672–15681, 2025.

Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3228–3239, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pp. 8748–8763, 2021.

Sebastian Rassmann, David Kügler, Christian Ewert, and Martin Reuter. Regression is all you need for medical image translation. *IEEE Transactions on Medical Imaging*, pp. 1–1, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5164–5174, 2025.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

Minye Shao, Xingyu Miao, Haoran Duan, Zeyu Wang, Jingkun Chen, Yawen Huang, Xian Wu, Jingjing Deng, Yang Long, and Yefeng Zheng. Trace: Temporally reliable anatomically-conditioned 3d ct generation with enhanced efficiency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 627–637, 2025.

Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE Transactions on Medical Imaging*, 39(4):1170–1183, 2020.

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, and Dosik Hwang. Anatomical consistency and adaptive prior-informed transformation for multi-contrast mr image synthesis via diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30918–30927, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Tao Song, Yicheng Wu, Minhao Hu, Xiangde Luo, Linda Wei, Guotai Wang, Yi Guo, Feng Xu, and Shaoting Zhang. Learning modality-aware representations: Adaptive group-wise interaction network for multimodal mri synthesis. *IEEE Transactions on Medical Imaging*, pp. 1–1, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 2025a.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023.

Yulin Wang, Honglin Xiong, Kaicong Sun, Shuwei Bai, Ling Dai, Zhongxiang Ding, Jiameng Liu, Qian Wang, Qian Liu, and Dinggang Shen. Toward general text-guided multimodal brain mri synthesis for diagnosis and medical image analysis. *Cell Reports Medicine*, 2025b.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In *AAAI*, volume 39, pp. 21501–21509, 2025.

Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pp. 338–357, 2024.

Haoning Wu, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Mrgen: Segmentation data engine for underrepresented mri modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19903–19913, 2025.

Yan Xia, Le Zhang, Nishant Ravikumar, Rahman Attar, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, and Alejandro F. Frangi. Recovering from missing data in population imaging – cardiac mr image imputation via conditional generative adversarial nets. *Medical Image Analysis*, 67:101812, 2021.

Jia Xiao, Wen Zheng, Wenji Wang, Qing Xia, Zhennan Yan, Qian Guo, Xiao Wang, Shaoping Nie, and Shaoting Zhang. Slice2mesh: 3d surface reconstruction from sparse slices of images for the left ventricle. *IEEE Transactions on Medical Imaging*, 44(3):1541–1555, 2025.

Heran Yang, Jian Sun, and Zongben Xu. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 42(12):3678–3689, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025.

Wen Ye, Zhetao Guo, Yuxiang Ren, Yi Tian, Yushi Shen, Zan Chen, Junjun He, Jing Ke, and Yiqing Shen. Diffm4ri: A latent diffusion model with modality inpainting for synthesizing missing modalities in mri analysis. *IEEE Journal of Biomedical and Health Informatics*, 30(2):1006–1018, 2026.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7685–7695, 2025.

Weihao Yu, Yuanhao Cai, Ruyi Zha, Zhiwen Fan, Chenxin Li, and Yixuan Yuan. X2-gaussian: 4d radiative gaussian splatting for continuous-time tomographic reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24728–24738, 2025.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

Le Zhang, Marco Pereañez, Christopher Bowles, Stefan Piechnik, Stefan Neubauer, Steffen Petersen, and Alejandro Frangi. Missing slice imputation in population cmr imaging via conditional generative adversarial nets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 651–659. Springer, 2019a.

Le Zhang, Marco Pereañez, Christopher Bowles, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. Unsupervised standard plane synthesis in population cine mri via cycle-consistent adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 660–668. Springer, 2019b.

Le Zhang, Kevin Bronik, Stefan K Piechnik, Joao AC Lima, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. Automatic plane pose estimation for cardiac left ventricle coverage estimation via deep adversarial regression network. *IEEE Transactions on Artificial Intelligence*, 5(9):4738–4752, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Ruiheng Zhang, Jingfeng Yao, Huangxuan Zhao, Hao Yan, Xiao He, Lei Chen, Zhou Wei, Yong Luo, Zengmao Wang, Lefei Zhang, et al. Unix: Unifying autoregression and diffusion for chest x-ray understanding and generation. *arXiv preprint arXiv:2601.11522*, 2026.

Can Zhao, Pengfei Guo, Dong Yang, Yucheng Tang, Yufan He, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. *arXiv preprint arXiv:2508.05772*, 2025.

Chenhui Zhao and Liyue Shen. Part-aware prompted segment anything model for adaptive segmentation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, Cheng Ouyang, Daniel Rueckert, Victor M. Campello, Karim Lekadir, Sulaiman Vesal, Nishant RaviKumar, Yashu Liu, Gongning Luo, Jingkun Chen, Hongwei Li, Buntheng Ly, Maxime Sermesant, Holger Roth, Wentao Zhu, Jiexiang Wang, Xinghao Ding, Xinyue Wang, Sen Yang, and Lei Li. Cardiac segmentation on late gadolinium enhancement mri: A benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis*, 81:102528, 2022.

Appendix

A Datasets

The details of the brain and cardiac MRI datasets used in our experiment are summarized in Table 12. Notably, we train our CoPeVAE and MDiT3D models on the brain MRI synthesis task on BraTS and IXI datasets separately, due to differences in the number and types of modalities. The evaluation and results are also reported for the two datasets separately. For Cardiac MRI synthesis, we leverage a combination of all four datasets to train both two-stage models.

Table 12: Details of brain and cardiac MRI datasets.

Datasets	**Modality**	**Cases**	**Train**	**Test**
Brain MRI				
BraTS Baid et al. (2021)	T1, T1ce, T2, FLAIR	1251	1000	251
IXI Brain Development Project (2025)	T1, T2, PD	577	462	115
Cardiac MRI				
UKBB Petersen et al. (2016)	-	31350	25080	6270
MESA Zhang et al. (2018)	-	298	238	60
ACDC Bernard et al. (2018)	-	300	240	60
MSCMR Zhuang et al. (2022)	-	300	240	60
Combined	-	32248	25798	6450

B More Implementation Details

B.1 Data Preprocessing

Brain MRI data. Following Liu et al. (2023); Meng et al. (2024), we use 90 and 80 middle axial slices for BraTS and IXI datasets, respectively. These slices are further cropped to a size of 192×192 from the central region. Ultimately, all volumes are resized to a fixed size of $192 \times 192 \times 64$ to serve as model input.

Cardiac MRI data. All slices within each cardiac MRI volume are used and cropped to 192×192 from the central region. Each volume is then resized to a fixed size of $192 \times 192 \times 32$ for training and inference.

For all datasets, we apply intensity normalization by linearly scaling voxel intensities between the 0.5th and 99.5th percentiles to the range $[0, 1]$. The data augmentations we employ include random spatial cropping, rotation, flipping, scaling, and shifting.

B.2 Architecture of Prompt Encoders and Projection Heads

We devise three lightweight prompt encoders to generate completeness-aware prompt tokens in our CoPeVAE, followed by three projection heads for each pretext task. The detailed architecture of each prompt encoder and projection head is illustrated in Table 13.

Notably, in our framework, binary mask codes are only used in the pretraining of CoPeVAE, where we synthetically remove modalities/slices and supervise the pretext tasks with known missing patterns. Once CoPeVAE is trained, we freeze its parameters and use it as a completeness-aware tokenizer: given any incomplete MRI with an arbitrary missing pattern, CoPeVAE directly infers the corresponding completeness prompts $\mathbf{p} = \mathbf{p}^d \,\|\, \mathbf{p}^p \,\|\, \mathbf{p}^s$ from the observed data, without requiring explicit mask codes as input. During both diffusion model training and inference, the diffusion backbone receives only the latent representations and these learned prompts. The original binary masks that were used to generate synthetic missingness are not provided to the diffusion model. In this sense, CoPeDiT no longer depends on hand-crafted or externally supplied mask codes at the generation stage, but instead relies entirely on the learned completeness prompts produced by the frozen CoPeVAE.

Table 13: Detailed architecture of each prompt encoder and projection head in the pretext task.

Prompt Encoder	$\mathcal{F}_1$	$\mathcal{F}_2$	$\mathcal{F}_3$
Architecture	3D Conv (in 8, out 256) 3D BatchNorm ReLU 3D Conv (in 256, out 512) 3D BatchNorm ReLU 3D Adaptive Avg Pool Linear (512, 1024) ReLU Linear (1024, 512)	3D Conv (in 8, out 256) 3D BatchNorm ReLU 3D Conv (in 256, out 512) 3D BatchNorm ReLU 3D Adaptive Avg Pool Linear (512, 1024) ReLU Linear (1024, 512)	3D Conv (in 8, out 256) 3D BatchNorm ReLU 3D Conv (in 256, out 512) 3D BatchNorm ReLU 3D Adaptive Avg Pool Linear (512, 1024) ReLU Linear (1024, 512)
Projection Head	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
Architecture	SiLU Linear (512, $m-1$)	SiLU Linear (512, m)	SiLU Linear (512, 128)

Table 14: Hyperparameter setup of CoPeVAE.

	CoPeVAE-B	**CoPeVAE-C**
Architecture		
Input dim.	$m \times 192 \times 192 \times 64$	$192 \times 192 \times 32$
Num. codebook	8192	8192
Latent dim.	8	8
Channels	(256, 384, 512)	(32, 64, 128)
Compression ratio	(8, 8, 8)	(8, 8, 8)
Prompt dim.	512	512
τ	0.2	0.2
λ	0.01	0.01
Optimization		
Batch size	8	64
Learning rate	1e-4	1e-4
Optimizer	Adam	Adam
(β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)
LR schedule	Warmup cosine	Warmup cosine
Training steps	400k	100k

B.3 Hyperparameter Setups

CoPeVAE. The detailed hyperparameter setup of CoPeVAE is provided in Table 14. Built upon VQ-VAE van den Oord et al. (2017) and VQGAN Esser et al. (2021), our model employs a codebook containing 8192 codes with the latent dimensionality of 8. The values of τ and λ are empirically set, as they have only a slight impact on model performance. The Adam optimizer is applied with a warmup cosine learning rate schedule. The training steps of CoPeVAE-B and CoPeVAE-C are 400k and 100k, respectively.

MDiT3D. For MDiT3D, we carefully design the hyperparameters to balance dataset size and model capacity, as summarized in Table 15. Following DiT Peebles & Xie (2023), we report the experimental results using the exponential moving average (EMA) with a decay rate of 0.9999. During training, we set the timestep to 500 with linearly scaled noise levels ranging from 0.0015 to 0.0195. MDiT3D is trained for 100k iterations using the AdamW optimizer and a warmup cosine learning rate schedule. During inference, the DDIM

Table 15: Hyperparameter setup of MDiT3D.

	MDiT3D-B	**MDiT3D-C**
Architecture		
Input dim.	$m \times 8 \times 24 \times 24 \times 8$	$4 \times 8 \times 24 \times 24$
Hidden dim.	768	576
Num. blocks	16	12
Num. heads	12	12
Patch size	2	1
Params (M)	173.3	33.0
Flops (G)	555.1 (BraTS) 424.5 (IXI)	104.0
Optimization		
Batch size	32	64
Learning rate	5e-5	5e-5
Optimizer	AdamW	AdamW
(β_1, β_2)	(0.9, 0.999)	(0.9, 0.999)
LR schedule	Warmup cosine	Warmup cosine
Training steps	100k	100k
EMA decay	0.9999	0.9999
Interpolants		
Training objective	$\mathbf{x}_0$-prediction	$\mathbf{x}_0$-prediction
Noise schedule	scaled-linear	scaled-linear
Timesteps	500	500
Sampler	DDIM	DDIM
Sampling steps	200	250

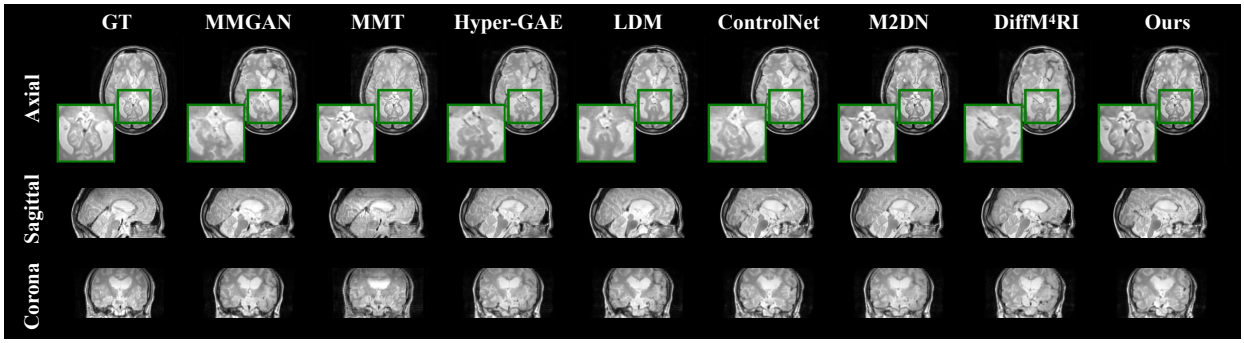

Figure 9: **Qualitative results on the IXI dataset.**

sampler Song et al. (2021) is applied with sample steps of 200 and 250 for MDiT3D-B and MDiT3D-C, respectively.

In addition, we use mixed-precision training with a gradient clipping to accelerate training and save computational resources throughout all two-stage experiments.

C Additional Experimental Results

C.1 Qualitative Results

Fig. 9 provides the qualitative evaluation results of brain MRI synthesis on the IXI dataset. As shown, we can also observe that our CoPeDiT outperforms other baselines in preserving the intricate structures and texture information in the synthesized MRIs.

Table 16: Ablation study on the contribution of completeness perception prompt tokens on the IXI dataset.

	1				2			
	PSNR↑	SSIM↑	FID↓	FVD↓	PSNR↑	SSIM↑	FID↓	FVD↓
w/o $\mathbf{p}^d$	23.56	0.720	35.93	724.18	23.04	0.708	48.71	917.76
w/o $\mathbf{p}^p$	22.81	0.697	50.70	1267.70	21.99	0.692	63.16	1281.31
w/o $\mathbf{p}^s$	24.17	0.724	31.89	685.13	23.56	0.710	39.87	876.94
w/o Prompts	22.48	0.696	57.64	1315.59	21.67	0.683	81.49	1542.24
w/ Mask Codes	23.35	0.711	44.03	783.46	22.84	0.702	56.72	1067.16
CoPeDiT	**24.34**	**0.732**	**25.84**	**569.22**	**23.92**	**0.721**	**32.53**	**718.54**

Table 17: Quantitative results on the IXI dataset by incorporating our completeness perception prompt tokens into baseline methods instead of mask codes.

	PSNR↑	SSIM↑	FID↓	FVD↓
MMT Liu et al. (2023)	22.64	0.698	53.60	1329.25
+ Prompts (ours)	23.19 (+0.55)	0.707 (+0.009)	46.13 (–7.47)	1096.06 (–233.19)
Hyper-GAE Yang et al. (2023)	22.12	0.682	72.62	1520.49
+ Prompts (ours)	22.46 (+0.34)	0.694 (+0.012)	59.43 (–13.19)	1261.53 (–258.96)
M2DN Meng et al. (2024)	23.47	0.715	42.52	845.29
+ Prompts (ours)	23.84 (+0.37)	0.726 (+0.011)	33.79 (–8.73)	684.62 (–160.67)

C.2 Ablation Study

We further evaluate the effectiveness of our prompt token design on the IXI dataset. As shown in Table 16, the complete set of prompt tokens yields the best performance, outperforming both conventional mask codes and all individual prompt tokens. Furthermore, we apply our learned prompt tokens to other baseline models originally using mask codes. As illustrated in Table 17, our prompts also lead to consistent performance gains across all baselines. In summary, the effectiveness of our prompt learning scheme is validated on the IXI dataset through the additional evaluations presented above.