MIAR: Medical Image Super-Resolution With Autoregressive Modeling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Medical image super-resolution aims to enhance the spatial resolution of low-resolution imaging modalities, thereby improving the visibility of fine anatomical structures without necessitating changes to acquisition hardware. Despite the success of convolutional and transformer-based approaches in natural image domains, existing ISR methods often exhibit limited generalizability and suboptimal trade-offs between fidelity and perceptual realism under constrained computational resources. Recent advances in diffusion models have yielded promising perceptual outcomes, but their high inference latency hinders deployment in clinical settings. In this work, we present **MIAR**, a novel multi-scale autoregressive framework tailored for image super-resolution, which formulates the reconstruction task as a progressive next-scale prediction problem. To enhance the semantic consistency and structural fidelity of the generated outputs, we extend the conventional visual autoregressive backbone by incorporating 2D rotary positional embeddings and a cross-attention fusion strategy. During inference, we replace conventional sampling strategies with beam search, enabling more stable and higher-quality generation paths. Experimental results demonstrate that **MIAR** not only establishes a new SOTA among single-modality methods on standard full-reference metrics such as PSNR and SSIM, but also achieves leading performance on perceptual quality metrics, outperforming both single- and multi-modality baselines. Furthermore, due to its autoregressive nature, **MIAR** offers significant efficiency advantages over traditional diffusion-based approaches, making it a practical solution for clinical and real-time medical imaging applications.

1 Introduction

Medical imaging modalities, such as magnetic resonance imaging (MRI) and cone-beam computed tomography (CBCT), are indispensable tools in clinical diagnosis, disease staging, and preoperative planning [41, 51]. These techniques enable non-invasive acquisition of high-resolution anatomical and functional information. However, physical and physiological constraints often limit the achievable spatial resolution. Specifically, MRI requires prolonged acquisition times, which not only reduce clinical throughput but also increase patient discomfort and susceptibility to motion-induced artifacts [12]. On the other hand, CBCT typically operates under low-dose protocols and hardware limitations, resulting in intrinsically low-resolution (LR) volumes.

To address these limitations, super-resolution reconstruction has emerged as a promising computational technique. Conventional SR methods, such as bicubic interpolation [21], sparse representation [52], and dictionary learning [19], struggle to accurately model the complex nonlinear relationships between low-resolution and high-resolution images, often failing to recover high-frequency anatomical details critical for clinical interpretation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Recent advances in deep learning have led to the development of SR methods, particularly those based on convolutional neural networks (CNNs) [11, 37, 58] and Transformer-based architectures [18, 25, 35, 36], which significantly improve reconstruction performance. Nonetheless, these models may still produce over-smoothed outputs and struggle to recover fine anatomical structures, especially in images with complex textures or pathological variations. Generative models, including generative adversarial networks (GANs) [26, 40], have been employed to improve perceptual realism, but they often suffer from training instability and mode collapse.

Diffusion models (DMs) have recently emerged as a compelling alternative for medical image SR [5, 27, 38, 28, 4, 55], owing to their ability to model complex data distributions via iterative denoising from Gaussian noise. Compared to GANs, DMs offer improved generation stability and fidelity. However, their inference process typically requires hundreds of iterative steps, which imposes high computational costs and limits clinical applicability. Moreover, reducing the number of inference steps to accelerate generation may introduce artificial structures or distortions that deviate from true anatomical representations, thus compromising diagnostic reliability.

Autoregressive models have emerged as a scalable alternative to diffusion models in generative tasks, offering a more efficient framework for image synthesis. Notable implementations include VQGAN [10] and DALL-E [43], which leverage discrete token prediction for image generation. This paradigm enables autoregressive models to produce high-quality images while avoiding the iterative process inherent in diffusion models, thus reducing computational complexity. Several prominent studies adopting this approach have demonstrated significant advancements [2, 3, 31, 30, 14, 34]. Recently, VAR [46] has attracted considerable attention by innovating the quantization of images into scale-wise token maps and generating images through next-scale prediction, achieving impressive results across a wide range of generative tasks. This novel architecture ensures fine-grained detail and fidelity by employing a progressive generation strategy. Additionally, VAR reduces the number of inference steps compared to traditional diffusion models, thereby enhancing computational efficiency.

However, Image Super-Resolution in medical imaging poses several unique challenges: 1) How to effectively condition on LR inputs to reconstruct anatomically accurate and HR outputs; 2) In autoregressive models that linearize image data into one-dimensional token sequences, how to better encode spatial dependencies and positional relationships among visual tokens; 3) Due to the high structural similarity across medical images, autoregressive prediction is prone to cumulative error propagation, which can degrade reconstruction quality over successive steps.

To address these challenges, we propose a novel Next-Scale Prediction Autoregressive Modeling framework for medical ISR. Our approach builds upon the Visual Autoregressive architecture, enhancing it with Cross-Attention Fusion and 2D RoPE to better capture semantic features from LR inputs and maintain spatial coherence within the token sequence. Furthermore, to mitigate error accumulation during autoregressive inference, we introduce a beam search strategy, which broadens the hypothesis space at each step and yields more stable and anatomically consistent reconstructions.

Our main contributions are summarized as follows:

- We propose a novel autoregressive framework, **MIAR**, that incorporates LR priors into VAR pipeline. Additionally, we adopt 2D RoPE to enhance spatial coherence and structural fidelity in medical image super-resolution.
- To mitigate error accumulation during sequential inference, we incorporate a beam search sampling strategy that explores multiple high-probability generation paths. This improves robustness against hierarchical prediction errors.
- Through both quantitative and qualitative analyses, we demonstrate that MIAR exhibits strong performance, achieving high-quality and realistic image generation while also showcasing the promising generation speed.

2 Related Work

2.1 Non-Generative Model-Based SR

Several studies have explored the integration of CNNS and Transformer architectures to improve medical image super-resolution. Feng et al. [11] investigate multi-contrast fusion at different stages of the network to capture the dependencies among fused features, thereby enhancing their

representational capacity. Zhang et al. [58] introduce a squeeze-and-excitation attention mechanism with residual scaling, which not only stabilizes the training process but also enables more accurate reconstruction of fine textures and anatomical details. Li et al. [25] leverage Transformer-based modules to facilitate multi-scale contextual matching, significantly improving feature fusion quality. liang et al.[33] employs a hybrid approach where shallow features are first extracted using CNN layers, followed by hierarchical modeling through Swin Transformer blocks, enabling more effective global feature aggregation and fine-grained detail reconstruction. Furthermore, Lei et al. [24] observe that multi-contrast MR images share consistent information (e.g., anatomical structures and edges) while exhibiting inconsistencies in contrast. Based on this observation, they propose a decomposition-based variational network to disentangle and reconstruct shared and modality-specific components. Georgescu et al. [15] design multi-head spatial attention, where each head captures features at a different spatial scale via distinct receptive fields. Despite the faster generation rates and stability of these methods, they often fall short in terms of the richness of fine details when compared to generative models, which excel in capturing intricate textures and subtle anatomical nuances.

2.2 Generative Model-Based SR

Diffusion models (DMs) [17], a class of probabilistic generative models, synthesize data samples from Gaussian noise through a stochastic iterative denoising process. DMs have demonstrated impressive potential in medical image super-resolution, particularly in MRI applications. Li *et al.* [29] introduced the first diffusion-based single-image super-resolution framework and achieved promising results. Mao *et al.* [38] proposed a disentangled conditional diffusion model tailored for multi-contrast MRI brain image super-resolution. Chung *et al.* [5] reduced the computational burden by initializing the reverse diffusion process with a single forward pass, thereby decreasing the number of sampling steps required. Further, Chung *et al.* [6] developed a score-based diffusion model to accelerate MRI reconstruction. Gao *et al.* [13] presented an implicit diffusion model that integrates latent diffusion [44] with a conditional reflection mechanism for high-fidelity and continuous image super-resolution. More recently, Li *et al.* [27] introduced a diffusion-based prior with only four denoising steps and incorporated it into a Transformer-based multi-contrast super-resolution framework for efficient MRI image generation.

2.3 Visual Autoregressive Models

Visual autoregressive models commonly employ VQ-VAE [47] to convert images into discrete token sequences, enabling next-token prediction for image synthesis[23, 45, 49, 16]. While effective, this approach disrupts the inherent spatial structure of 2D images and is often accompanied by slow generation speed. To address the inefficiency,chang et al.[3] propose iterative masked token prediction, which significantly accelerates the image generation process. However, these techniques still fail to fully preserve the spatial continuity and structural integrity of the original image. To overcome these limitations, Tian et al.[46] introduces a paradigm shift from next-token to next-scale prediction. This hierarchical generation strategy substantially improves image fidelity while maintaining high scalability. Built upon this framework, VAR-based models have also demonstrated remarkable performance across a wide range of generative tasks, including text-to-image generation[56], image-to-image generation[54, 32], image restoration[48], and image super-resolution[42].

3 Method

Our method consists of three main components:

Multi-Scale VQ-VAE: This framework decomposes the HR image into a hierarchy of scale-wise latent tokens via a pre-trained Multi-Scale Vector Quantized Variational Autoencoder (MSVQ-VAE). These tokens provide discrete representations that facilitate autoregressive modeling across spatial scales. **MIAR Framework:** The core of our approach, the MIAR framework, follows a Transformer-style decoder architecture enhanced with 2D RoPE to better capture spatial dependencies. During inference, the LR input is first encoded via an MSVQ-VAE-initialized encoder to obtain continuous latent feature vectors. These vectors, together with the discrete scale-wise HR tokens, are fed into the MIAR decoder to reconstruct the HR image. **Beam Search:** To mitigate error accumulation during the Variational Autoregressive generation process, we introduce a beam search mechanism. we select the K vectors and consider both the optimal and suboptimal options for each vector. This results in

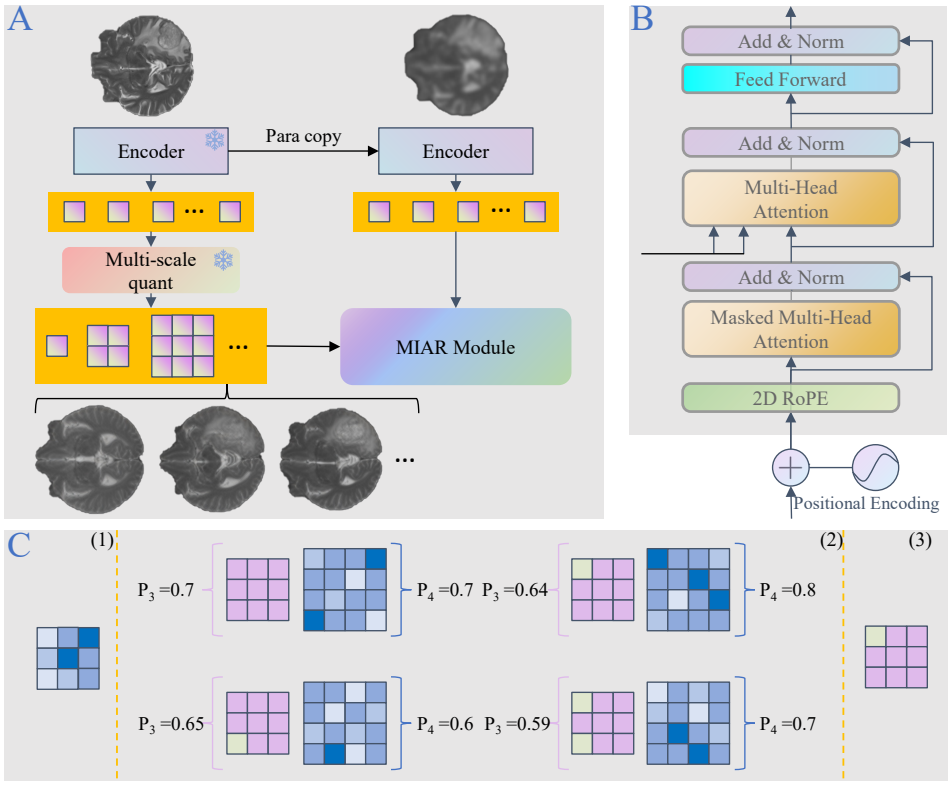


Figure 1: **Overview of the MIAR.** (A) MIAR framework: a HR image is decomposed into multi-scale discrete latent codes via a pre-trained Multi-Scale Vector Quantized VAE (MSVQ-VAE), while the LR input is simultaneously encoded into continuous representations using a parameter-shared encoder. Both representations are fused and processed by the MIAR module. (B) MIAR module: The MIAR module adopts a Transformer architecture equipped with 2D RoPE, enabling spatially-aware and scale-consistent generation. (C) Beam Search: A hierarchical beam search strategy with 2^K candidate paths is introduced to mitigate error accumulation during autoregressive token generation across multiple spatial layers.

2^k possible combinations, which are passed to the next layer. The specific generation method for each layer is then determined based on the outcomes from the subsequent layer.

3.1 Preliminary

Tian et al. [46] argue that the "next-token prediction" used in natural language processing leads to structural degradation and violations of mathematical premises in image generation tasks. Based on this observation, they propose a multi-scale residual VQ-VAE-based autoregressive paradigm, starting with a 1×1 token grid r_1 and autoregressively predicting larger-scale grids $(r_1, r_2, ..., r_S)$. The multi-scale residual VQ-VAE can be expressed as:

$$f_k = f - \sum_{m=1}^{k-1} \text{upsample}\left(\text{lookup}\left(Z, r_m\right)\right) \tag{1}$$

$$q^{(i,j)} = \left(\arg\min_{v \in [V]} \left\| \text{lookup}(Z, v) - f^{(i,j)} \right\|_2 \right) \in [V] \tag{2}$$

Here, $q^{(i,j)}$ represents the token corresponding to the i-th and j-th position, while Z denotes the VQ-VAE codebook. f_k denotes the upsampled feature map at scale k, which is progressively refined by the residual connection.

The autoregressive process is formulated as:

$$p(r_1, r_2, \ldots, r_K) = \prod_{k=1}^{K} p(r_k \mid r_1, r_2, \ldots, r_{k-1}) \tag{3}$$

At the s-th scale, the token mapping $r_s \in [V]^{h_s \times w_s}$ is generated based on the previously predicted tokens $\{r_1, r_2, \ldots, r_{s-1}\}$, where h_s and w_s are the height and width of the token grid at scale s, and V is the VQ-VAE codebook.

For each scale, the model minimizes the negative log-likelihood of the generated tokens using a cross-entropy loss. During generation, tokens are sampled from the conditional probability distribution $p(r_s \mid r_{<s})$. The model can utilize Top-k sampling, where the top k most probable candidates are selected, or Top-p sampling, where the smallest set of candidates whose cumulative probability exceeds p is chosen. In both cases, the selection is weighted by the probability distribution.

The VAR, with next-scale prediction, represents a significant advancement in autoregressive models. For Image Super-Resolution tasks requiring high fidelity and realism, the progressive generation of VAR ensures consistent refinement from coarse to fine, aligning with human perception and the Markovian unidirectional assumption. Therefore, the application of VAR to ISR holds great potential.

3.2 Low-Resolution Control

One of the key challenges in autoregressive super-resolution for medical imaging is the effective incorporation of LR structural priors during the sequential token generation process. While conditioning mechanisms such as context vectors and ControlNet-like structures are effective in diffusion and GAN-based frameworks, they are less compatible with autoregressive transformers. Previous work, such as VARSR [42], explores prefix-based conditioning, but this approach may generate meaningless token sequences during the generation process, leading to performance loss.

To address this, we propose **MIAR**, a multi-scale conditioning framework tailored for medical image autoregressive reconstruction. As shown in Fig. 1A, for an LR image, $\text{img}_{LR} \in \mathbb{R}^{H \times W \times C}$, we use an encoder similar to the MSVQ-VAE structure to extract features. This encoder is initialized with a MSVQ-VAE to obtain a relatively good initial model. After passing through the encoder, the image is encoded as features $f_{LR} \in \mathbb{R}^{C_{lr} \times L_{lr}}$, where C_{lr} is a tunable hyperparameter (typically set to 32). For a HR image, $\text{img}_{HR} \in \mathbb{R}^{H \times W \times C}$, we use MSVQ-VAE to obtain scale-wise tokens, $f_{HR} \in \mathbb{R}^{C_{HR} \times L_{hr}}$, where C_{hr} denotes the aggregated channel width across scales. These representations are then fed into our proposed MIAR module (Fig. 1B), where generation is guided via cross-attention between f_{LR} and the target sequence. To preserve spatial alignment and anatomical fidelity, we incorporate 2D RoPE throughout the decoding pipeline. This design enables effective conditioning at multiple scales while maintaining anatomical consistency during HR image synthesis.

3.3 Beam Search

Due to the structural similarities and local continuity of anatomical features in medical images, the predicted tokens in medical image synthesis often exhibit high numerical proximity (Sec. 4.4.3). This increases the risk of error accumulation in autoregressive generation, especially when employing widely-used sampling strategies such as top-k and top-p [39, 8]. Beam search is a well-established decoding algorithm that mitigates this issue by retaining multiple candidate sequences during inference. It can be viewed as a pruned version of the classic breadth-first search, constrained by a fixed beam width. While standard autoregressive models for text generation predict a single index at each step, next-scale prediction autoregressive models must generate a grid of $h_i \times h_i$ tokens at each scale. To address this problem, we propose a beam search variant specifically designed for next-scale prediction in medical image super-resolution.

Algorithm 1 Beam Search in VAR

Input: choose tokens k, max scale T
$B_1 \leftarrow \{\langle 0, \texttt{ST} \rangle\}$
for $t = 1$ to T **do**
  $B_{t+1} \leftarrow \emptyset$
  $\mathcal{V}_t \leftarrow chooseTokens(B_t, k)$
  $Y \leftarrow getCandidate(\mathcal{V}_t)$
  for each $y_t \in Y$ **do** ▷ 2^k choices
    $s' \leftarrow \texttt{score}(s, y_t)$
    $B_{t+1} \leftarrow B_{t+1} \cup \langle s', y_{1:t-1} \| y_t \rangle$
  end for
  $B_{t+1} \leftarrow \text{Max } B_{t+1}$
end for
Output: $\arg\max B_T$

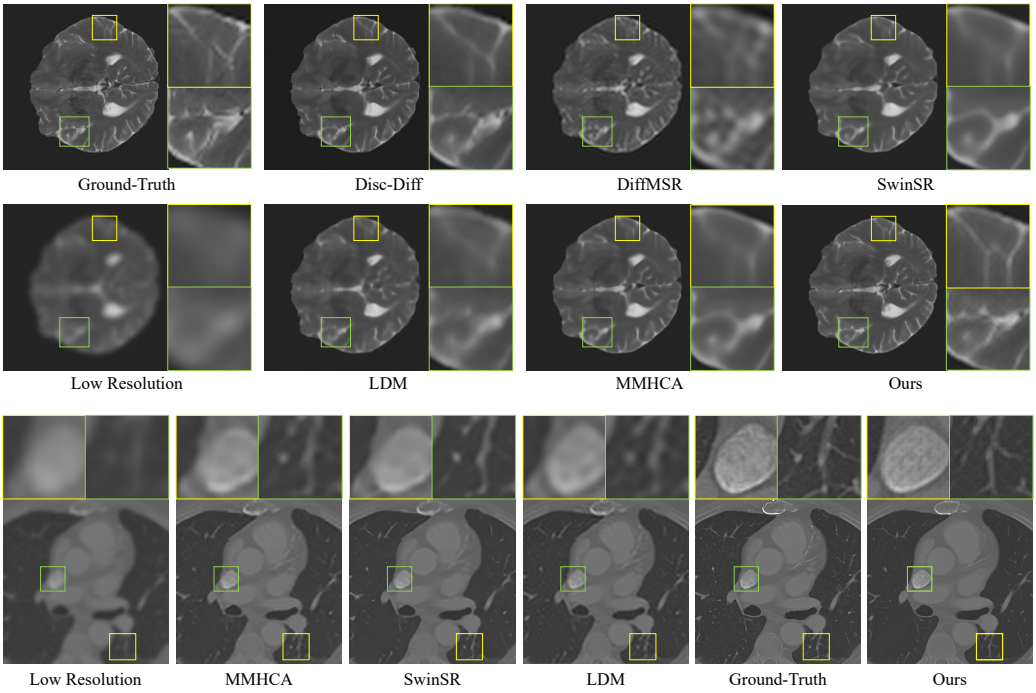


Figure 2: **Visual Experiment** We conduct a comprehensive evaluation of our proposed method, MIAR, against state-of-the-art super-resolution approaches on two medical imaging datasets: *brain* MRI and *lung* CT. For the brain dataset, we include both multimodal (Disc-Diff [38], DiffMSR [27]) and unimodal (SwinIR [33], LDM [44], MMHCA [15]) baselines, while for the lung dataset, we focus on unimodal settings due to the modality constraints. Our approach consistently achieves superior perceptual fidelity as verified through qualitative visual assessment and outperforms all competing methods across standard quantitative metrics. These results underscore the effectiveness of our autoregressive generation strategy and demonstrate its robustness across anatomical domains.

As shown in Alg. 1, the process is as follows: First, the candidate set B_0, initialized with the start symbol, is created. At each generation step t, the model selects k candidate tokens based on a predefined beam width. These candidates are used to construct 2^k possible token configurations, which are evaluated using a scoring function. The top-scoring configuration is propagated to the next step, ensuring both optimality and diversity in the autoregressive decoding process.

Let the token set at scale i be denoted as $\mathcal{V}(i)$. We identify a subset of k tokens, denoted as $\mathcal{V}_{\text{choose}}(i)$, based on the following criterion:

$$\mathcal{V}_{\text{choose}}(i) = \underset{v \in \mathcal{V}(i)}{\text{top-}k} \left(P_{\text{max}}(v) + \alpha (P_{\text{max}}(v) - P_{\text{2nd}}(v)) \right) \tag{4}$$

Here, $P_{\max}(v)$ and $P_{2\text{nd}}(v)$ denote the highest and second-highest predicted probabilities for token v, respectively. The coefficient α is a tunable hyperparameter. This selection prioritizes tokens with low confidence and high uncertainty. For each of the k selected tokens, we consider both optimal and suboptimal candidate values, resulting in 2^k possible token configurations for scale i. These candidate sets are propagated through the autoregressive decoder to predict tokens at scale $i{+}1$. Scoring function is then used to rank all candidates, defined as:

$$\text{Score} = s + \sum \log(P_{\text{max}}(\hat{y})) \tag{5}$$

where s is the accumulated score from previous steps, and $\hat{y}$ represents the predicted tokens. The highest-scoring configuration is retained for subsequent inference.

This method effectively combines the scale-wise progressive generation and scoring mechanism of MIAR model, while reducing computational complexity through selective search. It ensures high-quality generated sequences and alleviates the issue of error accumulation to some extent.

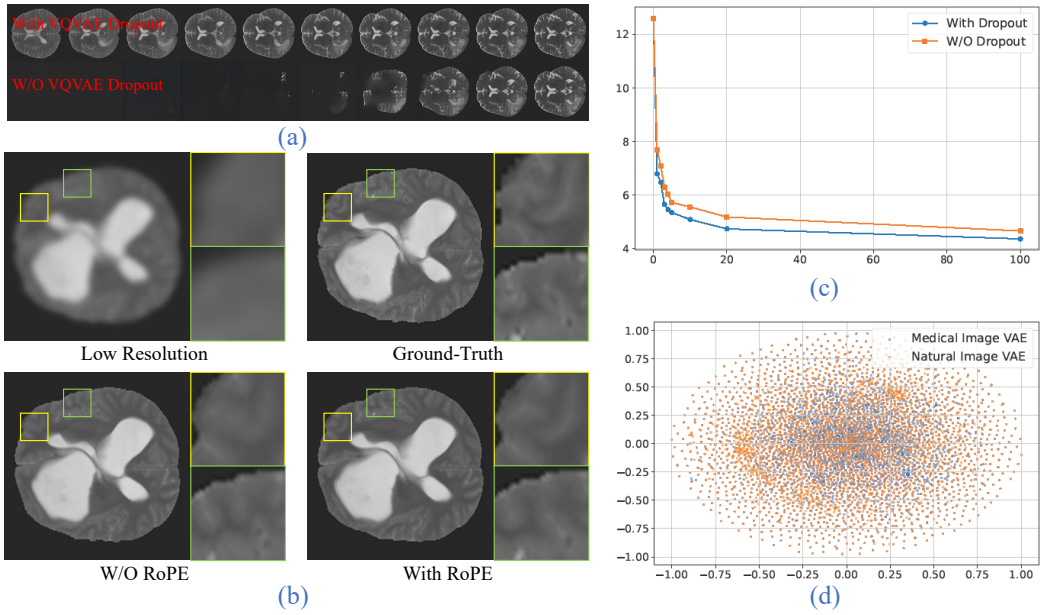


Figure 3: **Ablation studies and visual analyses.** (a) Layer-wise reconstructions from MSVQ-VAE with and without scale-aware dropout. The dropout improves intermediate consistency and structural coherence. (b) Visual comparison of reconstructed slices with and without RoPE, where RoPE improves anatomical fidelity and preserves fine-grained textures in medical images. (c) Training curves show the evolution of mean loss across epochs. The use of scale-aware dropout accelerates convergence and enhances reconstruction quality. (d) t-SNE visualization of latent codebook embeddings from MSVQ-VAEs trained on natural versus medical images. The tighter clustering observed in medical latent codes reflects higher structural similarity.

4 Experiments

4.1 Setups

4.1.1 Datasets and Baselines

Dataset. We conduct our experiments on two public datasets: the BraTS2021 brain MRI dataset [7] and the LIDC-IDRI lung CT dataset [1]. The BraTS2021 dataset contains 1,251 multi-contrast MRI volumes with four imaging modalities. The LIDC-IDRI dataset comprises 1,308 scans in thoracic CT. For both datasets, we extract only the central 40 axial slices from each volume to reduce redundancy and focus on diagnostically relevant regions. The datasets are randomly partitioned into training, validation, and testing subsets with a ratio of 7:1:2, respectively. All 2D slices are cropped to 256×256 and a $4\times$ super-resolution factor is applied. To better emulate real-world acquisition artifacts, Gaussian blur (kernel radius = 1) is applied to the LR image.

Baselines. To comprehensively evaluate the effectiveness of our proposed method, we compare it with several state-of-the-art super-resolution models, including both multimodal baselines (Disc-Diff [38], DiffMSR [27]) and unimodal baselines (SwinIR [33], LDM [44], and MMHCA [15]). For BraTS2021, we assess both multimodal (using T1 to reconstruct T2) and unimodal setups. For LIDC-IDRI, only unimodal CT-based reconstruction is performed.

4.1.2 Implementation Details

We implement our proposed framework using PyTorch and conduct all experiments on four NVIDIA RTX 4090 GPUs. The network is optimized using the Adam optimizer [22] 100 epochs, divided into two sequential training stages. In the first stage, we initialize the variational autoencoder (VAE) and the discriminator with a batch size of 24, using learning rates of 2×10^{-4} and 1×10^{-4}, respectively. In the second stage, we train the full autoregressive model with a batch size of 32, using a fixed learning rate of 1×10^{-4} across all components.

Table 1: **Quantitative results on Brats and Lung dataset.**

Dataset	Method	Type	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑
Brats	Disc-Diff[38]	MC	33.57	0.9241	0.1002	0.1497	39.34	0.2511
	DiffMSR[27]	MC	34.49	0.9318	0.1991	0.2145	27.35	0.2439
	Bicubic	SC	29.11	0.8382	0.4470	0.3185	18.01	0.1438
	SwinIR[33]	SC	33.33	0.9161	0.1302	0.1674	34.75	0.2756
	MMHCA[15]	SC	33.48	0.9230	0.1316	0.1693	35.84	0.2888
	LDM[44]	SC	32.43	0.9156	0.1544	0.1762	34.14	0.2689
	MIAR(Ours)	SC	34.22	0.9367	0.0556	0.0773	47.87	0.3431
Lung	Bicubic	SC	33.07	0.8426	0.4481	0.2965	21.49	0.1457
	MMHCA[15]	SC	35.77	0.9072	0.1470	0.1602	33.93	0.2487
	SwinIR[33]	SC	35.13	0.8926	0.1526	0.1736	32.55	0.2423
	LDM[44]	SC	34.80	0.8713	0.1997	0.1885	26.10	0.2204
	MIAR(Ours)	SC	36.39	0.9171	0.0779	0.0840	40.87	0.2674

4.1.3 Metric

We assess reconstruction quality using both reference-based and no-reference metrics. PSNR and SSIM [50] are standard full-reference metrics evaluating pixel-level fidelity and structural similarity, respectively. LPIPS [57] and DISTS [9] measure deep feature differences using pretrained networks, offering perceptual insights. For no-reference evaluation, MUSIQ [20] and MANIQA [53] capture multi-scale and attention-based perceptual quality.

4.2 Qualitative Analysis

We present a qualitative comparison of various super-resolution methods on two benchmark medical datasets, namely BraTS (top row) and LIDC-IDRI (bottom row). Highlighted regions in yellow and green boxes provide magnified views for detailed visual inspection. As shown in Fig. 2, **DisC-Diff** produces perceptually pleasing reconstructions but fails to preserve structural fidelity, introducing hallucinated anatomical details that deviate from the ground truth. In contrast, **DiffMSR** retains more fine-grained anatomical features, albeit at the cost of reduced perceptual sharpness. **SwinIR** and **MMHCA** strike a balance between perceptual quality and detail retention; however, their outputs are often overly smoothed, leading to loss of subtle textures and edge information. **LDM** generates relatively blurred reconstructions with limited structural detail preservation. In comparison, **our proposed method**, leveraging controllable autoregressive generation, achieves the best perceptual realism while maintaining high anatomical accuracy. The generated images exhibit comparable clarity to the ground truth, with minor discrepancies.

4.3 Quantitative Analysis

As shown in Tab. 1, we conduct a comprehensive quantitative comparison between our proposed method and three state-of-the-art or widely-used single-modality super-resolution (SCSR) methods, as well as two leading multi-modality super-resolution (MCSR) methods on the BraTS MRI and LIDC-IDRI CT datasets. Our method consistently outperforms all SCSR baselines across most evaluation metrics, establishing a new performance benchmark within the single-modality domain. Although slightly inferior to MCSR approaches in terms of PSNR, our method achieves the best performance on perceptual metrics, highlighting its superior visual fidelity. Specifically, **DisC-Diff**, based on a diffusion model, demonstrates strong perceptual quality and favorable no-reference image quality, yet suffers from relatively low PSNR due to the generation of semantically plausible but structurally inaccurate content. **DiffMSR**, leveraging a four-step denoising process to generate prior guidance, achieves the highest PSNR, but exhibits poor perceptual quality and visual realism. **SwinIR** and **MMHCA**, as non-generative SISR baselines, exhibit a compromise between fidelity and sharpness but fail to deliver compelling results in either aspect. **LDM**, which lacks adaptation to the medical imaging domain, underperforms in structural fidelity but still provides reasonable visual quality, showcasing the generalization capability of diffusion-based frameworks. Our proposed method, benefiting from controllable autoregressive generation, ranks second in PSNR among all models, trailing only the MCSR-based DiffMSR, while achieving state-of-the-art results in both perceptual similarity and no-reference quality assessment. These findings suggest that our approach

offers a compelling balance between accuracy and perceptual realism, making it a promising direction for future medical image super resolution research.

4.4 Ablation Study

4.4.1 ROPE

As shown in Tab. 1, the incorporation of RoPE significantly enhances the performance on reference-based metrics. This improvement indicates that RoPE enables the model to better preserve fine-grained structural details and semantic fidelity in the reconstructed HR images. Moreover, visual comparisons in Fig. 3b further validate that RoPE facilitates superior retention of anatomical features, which is critical for clinical applications requiring high-precision image interpretation.

4.4.2 Scale-Aware Dropout

As shown in Tab. 1, applying dropout in the MSVQ-VAE bottleneck notably improves reference-based metrics such as PSNR and SSIM, indicating that regularized latent representations facilitate higher-fidelity reconstructions. From Fig. 3a, we observe that models trained with scale-aware dropout tend to incrementally enrich structural and textural details across successive scales. In contrast, when dropout is disabled, the model defers most of the information reconstruction to the final few scales, leading to an imbalanced distribution of semantic content throughout the multi-scale hierarchy.

Moreover, as shown in Fig. 3c, removing scale-aware dropout significantly slows convergence and degrades final performance. This likely results from the elevated difficulty of learning meaningful representations early on, when information is sparse. Due to the model's autoregressive nature, early-stage errors propagate, harming overall generation quality.

4.4.3 Beam Search

As shown in Tab. 1 and Tab. 2, incorporating beam search with a beam width of $k = 1$ (equivalent to selecting two candidates per scale) yields substantial improvements in reconstruction quality. Although further increasing k can still lead to marginal gains, the performance benefit quickly diminishes relative to the associated computational overhead during inference.

Table 2: **Ablation on Beam Search**

Type	PSNR↑	SSIM↑	Speed↑
Disc-Diff[38]	33.57	0.9241	1455ms
LDM[44]	32.43	0.9156	832ms
w/o Beam Search	33.49	0.9287	257ms
Beam Search K=1	34.22	0.9367	432ms
Beam Search K=2	34.25	0.9369	705ms
Beam Search K=3	34.27	0.9369	1372ms

To better understand the effectiveness of beam search, we conduct an empirical analysis of the learned codebook distributions. As illustrated in Fig. 3d, we apply T-SNE to visualize the latent embeddings from the MSVQ-VAE trained on natural images (VAR) and our domain-specific model. The codebook learned by VAR exhibits a more dispersed and uniform distribution, reflecting the high visual diversity in natural images. In contrast, medical images tend to be more structurally consistent, resulting in a more compact and clustered latent space. This discrepancy causes VAR to be more prone to sampling errors during autoregressive decoding. Beam search proves particularly beneficial in this context, as it mitigates error accumulation by chooseing multiple high-probability candidate paths during generation.

5 Conclusion and Disscusion

We propose **MIAR**, a multi-scale autoregressive framework for medical image super-resolution. By enhancing the VAR backbone with *Cross-Attention Fusion* and *Rotary Positional Encoding*, and introducing a beam search decoding strategy, our model improves semantic consistency, spatial fidelity, and inference stability. **MIAR** achieves superior performance on perceptual quality metrics and competitive results on pixel-wise fidelity measures, while significantly outperforming diffusion-based models in inference speed. This highlights its effectiveness and practicality for high-resolution medical image super-resolution.

Despite its effectiveness, our method has limitations. Its performance heavily relies on encoder tokenization quality. Additionally, VQ-VAE quantization introduces information loss, potentially impairing super-resolution. Future work may explore quantization-free tokenization schemes.

References

[1] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[2] Y. Bai, X. Geng, K. Mangalam, A. Bar, A. L. Yuille, T. Darrell, J. Malik, and A. A. Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024.

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

[4] H. Chung, E. S. Lee, and J. C. Ye. Mr image denoising and super-resolution using regularized reverse diffusion. *IEEE transactions on medical imaging*, 42(4):922–934, 2022.

[5] H. Chung, B. Sim, and J. C. Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12413–12422, 2022.

[6] H. Chung and J. C. Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

[7] Z. Cui, Y. Fang, L. Mei, B. Zhang, B. Yu, J. Liu, C. Jiang, Y. Sun, L. Ma, J. Huang, et al. A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. *Nature Communications*, 13(1):1–11, 2022.

[8] N. Deutschmann, M. Alberts, and M. R. Martínez. Conformal autoregressive generation: Beam search with coverage guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11775–11783, 2024.

[9] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

[10] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[11] C.-M. Feng, H. Fu, S. Yuan, and Y. Xu. Multi-contrast mri super-resolution via a multi-stage integration network. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 140–149. Springer, 2021.

[12] C.-M. Feng, Y. Yan, G. Chen, Y. Xu, Y. Hu, L. Shao, and H. Fu. Multimodal transformer for accelerated mr imaging. *IEEE Transactions on Medical Imaging*, 42(10):2804–2816, 2022.

[13] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[14] Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

[15] M.-I. Georgescu, R. T. Ionescu, A.-I. Miron, O. Savencu, N.-C. Ristea, N. Verga, and F. S. Khan. Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2195–2205, 2023.

[16] B. Guo, X. Zhang, H. Wu, Y. Wang, Y. Zhang, and Y.-F. Wang. Lar-sr: A local autoregressive model for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1909–1918, 2022.

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

[18] S. Huang, J. Li, L. Mei, T. Zhang, Z. Chen, Y. Dong, L. Dong, S. Liu, and M. Lyu. Accurate multi-contrast mri super-resolution via a dual cross-attention transformer network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 313–322. Springer, 2023.

[19] C. Jiang, Q. Zhang, R. Fan, and Z. Hu. Super-resolution ct image reconstruction based on dictionary learning and sparse representation. *Scientific reports*, 8(1):8799, 2018.

[20] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[21] D. Khaledyan, A. Amirany, K. Jafari, M. H. Moaiyeri, A. Z. Khuzani, and N. Mashhadi. Low-cost implementation of bilinear and bicubic image interpolation for real-time image super-resolution. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–5. IEEE, 2020.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[24] P. Lei, F. Fang, G. Zhang, and T. Zeng. Decomposition-based variational network for multi-contrast mri super-resolution and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21296–21306, 2023.

[25] G. Li, J. Lv, Y. Tian, Q. Dou, C. Wang, C. Xu, and J. Qin. Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20636–20645, 2022.

[26] G. Li, J. Lv, X. Tong, C. Wang, and G. Yang. High-resolution pelvic mri reconstruction using a generative adversarial network with attention and cyclic loss. *IEEE Access*, 9:105951–105964, 2021.

[27] G. Li, C. Rao, J. Mo, Z. Zhang, W. Xing, and L. Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11365–11374, 2024.

[28] G. Li, L. Zhao, J. Sun, Z. Lan, Z. Zhang, J. Chen, Z. Lin, H. Lin, and W. Xing. Rethinking multi-contrast mri super-resolution: Rectangle-window cross-attention transformer and arbitrary-scale upsampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21230–21240, 2023.

[29] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[30] T. Li, Q. Sun, L. Fan, and K. He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025.

[31] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.

[32] X. Li, K. Qiu, H. Chen, J. Kuen, Z. Lin, R. Singh, and B. Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024.

[33] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[34] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[35] J. Lyu, G. Li, C. Wang, Q. Cai, Q. Dou, D. Zhang, and J. Qin. Multicontrast mri super-resolution via transformer-empowered multiscale contextual matching and aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[36] J. Lyu, G. Li, C. Wang, C. Qin, S. Wang, Q. Dou, and J. Qin. Region-focused multi-view transformer-based generative adversarial network for cardiac cine mri reconstruction. *Medical Image Analysis*, 85:102760, 2023.

[37] Q. Lyu, H. Shan, C. Steber, C. Helis, C. Whitlow, M. Chan, and G. Wang. Multi-contrast super-resolution mri through a progressive network. *IEEE transactions on medical imaging*, 39(9):2738–2749, 2020.

[38] Y. Mao, L. Jiang, X. Chen, and C. Li. Disc-diff: Disentangled conditional diffusion model for multi-contrast mri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–397. Springer, 2023.

[39] C. Meister, T. Vieira, and R. Cotterell. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809, 2020.

[40] B. Murugesan, S. Vijaya Raghavan, K. Sarveswaran, K. Ram, and M. Sivaprakasam. Recon-glgan: a global-local context based generative adversarial network for mri reconstruction. In *Machine Learning for Medical Image Reconstruction: Second International Workshop, MLMIR 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 2*, pages 3–15. Springer, 2019.

[41] E. Plenge, D. H. Poot, M. Bernsen, G. Kotek, G. Houston, P. Wielopolski, L. van der Weerd, W. J. Niessen, and E. Meijering. Super-resolution methods in mri: can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time? *Magnetic resonance in medicine*, 68(6):1983–1993, 2012.

[42] Y. Qu, K. Yuan, J. Hao, K. Zhao, Q. Xie, M. Sun, and C. Zhou. Visual autoregressive modeling for image super-resolution. *arXiv preprint arXiv:2501.18993*, 2025.

[43] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

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

[45] P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[46] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[47] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[48] S. Wang and F. Zhao. Varformer: Adapting var's generative prior for image restoration. *arXiv preprint arXiv:2412.21063*, 2024.

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

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[51] Y. Wei, C. Li, and S. J. Price. Quantifying structural connectivity in brain tumor patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–529. Springer, 2021.

[52] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[53] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[54] Z. Yao, J. Li, Y. Zhou, Y. Liu, X. Jiang, C. Wang, F. Zheng, Y. Zou, and L. Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024.

[55] F. Zhang, G. Feng, X. Gao, and S. Niu. Efficient large-scale pre-trained model guided mr imaging super-resolution. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2852–2857. IEEE, 2024.

[56] Q. Zhang, X. Dai, N. Yang, X. An, Z. Feng, and X. Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024.

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[58] Y. Zhang, K. Li, K. Li, and Y. Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13425–13434, 2021.

A Implementation Details

Table 3: **Quantitative results on Brats and Lung dataset.**

Dataset	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑
Lung	38.15	0.9330	0.06478	0.0678	42.64	0.2723
Brats	36.34	0.9594	0.0407	0.0618	47.71	0.3371

VQVAE Tokenizer. We follow the basic setup of VAR. The VQ-VAE uses a downsampling factor of 8, with a codebook size of |V| = 4096 and a codebook feature dimension of 32. The results of the VQ-VAE are shown in Tab. 3. Comparing the data here with our super-resolution data, we can even find that on the Brats dataset, our super-resolution images even exceed VQVAE in image quality (MUSIQ, MANIQA). This also shows that VQVAE is only a compressed representation of the modeling pixel space, not the upper limit of the generation capability.

MIAR module. We design an autoregressive Transformer tailored for image synthesis, based on a modular GPT-style architecture. The model consists of 4 residual blocks, each comprising RMSNorm, a fused MLP, and attention layers. Each block integrates Scale-aligned 2D Rotary Position Embedding (RoPE), where spatially-aware sinusoidal embeddings are precomputed for each scale and applied to queries and keys to maintain spatial coherence across resolutions. Each block includes a self-attention layer followed by a cross-attention layer, where the LR image embeddings $\mathbf{f}_{\mathrm{LR}}$ serve as conditions. To maintain autoregressive constraints during training, we employ a scale-wise causal attention mask that ensures each scale F_k can only attend to its preceding context $\langle \mathrm{SOS} \rangle, F_1, \ldots, F_{k-1}$. At inference time, we leverage KV caching without mask to avoid redundant computation and accelerate generation.

B More Experiments

Zero shot

To evaluate the cross-domain generalization capability of our model, we conduct a zero-shot inference experiment. The model is trained exclusively on the Lung dataset and directly tested on the BraTS dataset without any fine-tuning or adaptation. This experiment simulates realistic clinical scenarios where access to annotated target-domain data is limited or unavailable.

As shown in Table 4, the quantitative results demonstrate that our model maintains reasonable generative performance despite the domain shift. The synthesized images preserve global anatomical structures; however, we observe partial degradation in localized structures, likely due to domain-specific appearance variations. Qualitative visualizations in Figure **??** provide further evidence of the model's ability to generate plausible structures under distributional shift, while also highlighting areas where fine-grained accuracy may be affected.

Table 4: **Quantitative results on Brats and Lung dataset.**

train Dataset	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	MANIQA↑
Lung	31.04	0.8990	0.1025	0.1542	43.97	0.3121
Brats	34.22	0.9367	0.0556	0.0773	47.87	0.3431

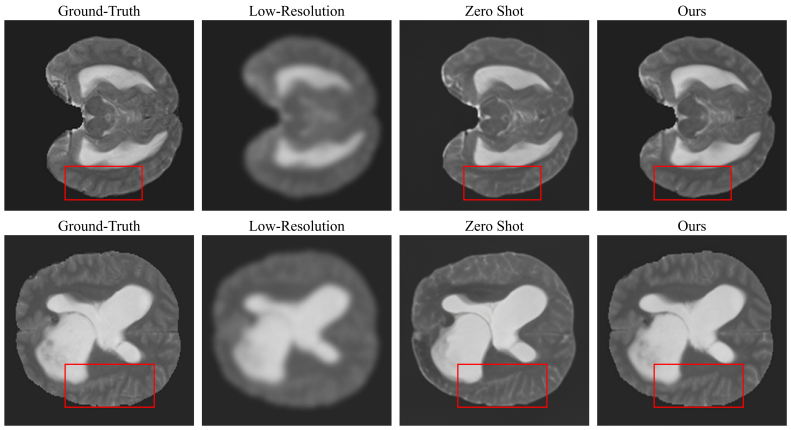
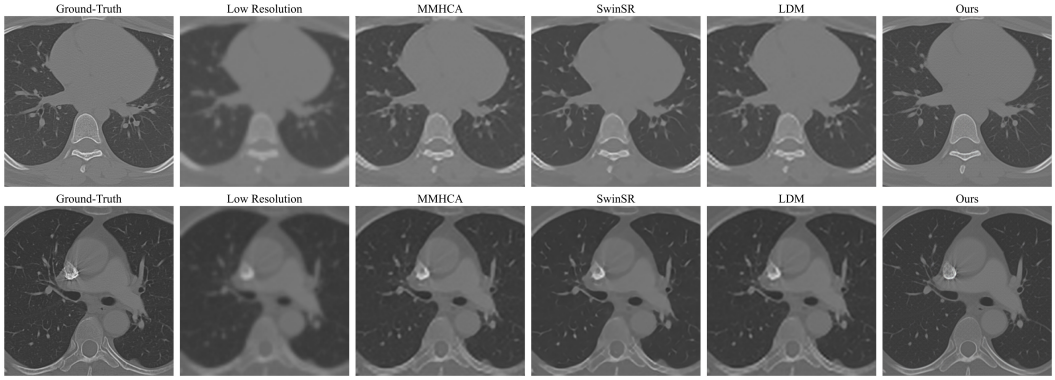
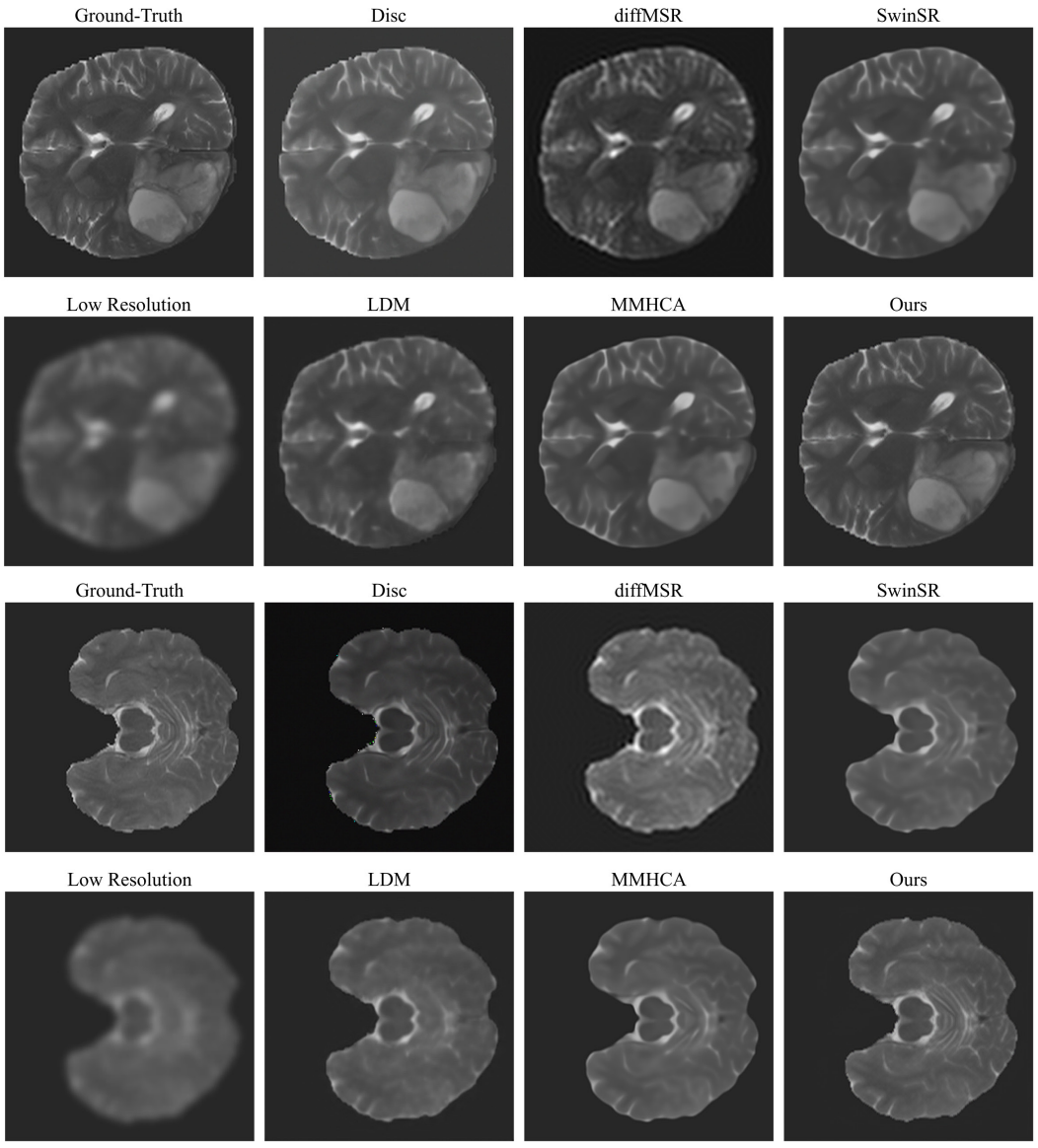


Figure 4: **Ablation studies and visual analyses.**

C More visualization results





NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, I well reflected the paper's contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discusses limitations in Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, the paper provide the full set of assumptions and a complete (and correct) proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we will provide the code and data to ensure the paper is reproducible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, I will publish the process and code as much as possible so that the paper can be reproduced.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper contains almost all the important hyperparameters and settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: No, the resources and cost required to perform a significance analysis are too great, so I did not perform a replication test. However, I did repeat the experiment for some results that were clearly worse or better.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, I mentioned the GPU type and the number of training epochs required in Sec. 4.1.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, I abide by the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: While medical image super-resolution can improve diagnostic clarity and reduce the need for high-dose acquisitions, it also carries risks of introducing spurious anatomical details or masking pathology, potentially leading to misdiagnosis. Mitigation strategies include clinician oversight and deployment of authentication.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers and sources for BraTS2021 and LIDC-IDRI datasets, and respect their usage terms. All external code repositories are credited with proper citations and license information.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our super-resolution code in a public repository, accompanied by a detailed README.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: I just used LLM to polish my paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.