DiverseVAR: Balancing Diversity and Quality of Next-Scale Visual Autoregressive Models

Mingue Park*, Prin Phunyaphibarn*, Phillip Y. Lee, and Minhyuk Sung

KAIST, Daejeon, South Korea
{kicikicik, prin10517, phillip0701, mhsung}@kaist.ac.kr

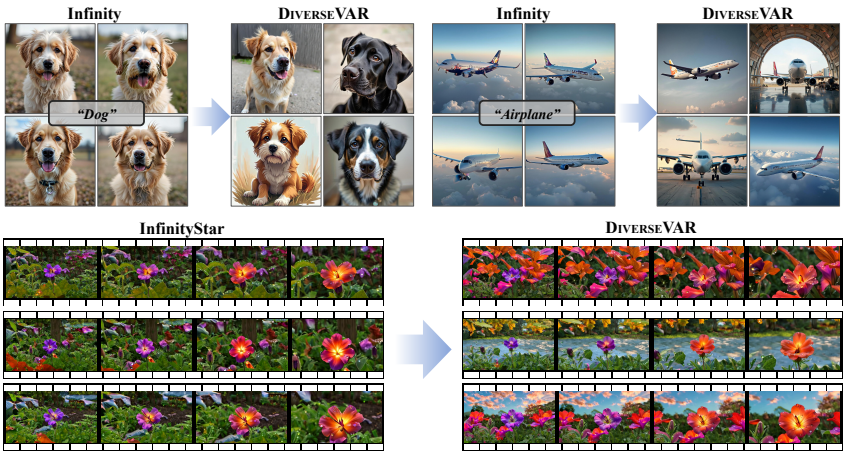

Fig. 1: DiverseVAR. Text-conditioned VAR models (Infinity [21] and InfinityStar [40]) exhibit a clear lack of per-prompt diversity in image and video generation. DiverseVAR combines diversity-enhancing annealing strategies with a novel latent refinement technique to increase the diversity while maintaining visual quality.

Abstract. We introduce DiverseVAR, a test-time framework that enhances the output diversity of text-conditioned visual autoregressive models (VAR) without additional training or substantial computational overhead. While VAR models have recently emerged as strong competitors of diffusion and flow models for image and video generation, they suffer from a critical diversity limitation: even simple prompts often produce nearly identical outputs. This issue has largely gone unnoticed amid the field's predominant focus on image quality. We address this limitation in two stages. First, we systematically explore diversity enhancement techniques by injecting noise into different components of VAR at test time, finding that noise injection into the text embedding yields the best diversity gains. However, this comes at the cost of sharp degradation in image quality. To recover quality, we propose **scale-travel**: a latent refinement

* Equal contribution.

technique that leverages a multi-scale autoencoder to extract coarse-scale tokens to resume generation from intermediate stages. Extensive experiments demonstrate that combining text-embedding noise injection with scale-travel refinement substantially improves diversity while minimizing quality degradation, advancing the diversity-quality Pareto front. Project page: `https://diverse-var.github.io/`

Keywords: Visual Autoregressive Models · Diversity

1 Introduction

Visual autoregressive (VAR) models [21,28,38,64,66,88,91] have recently emerged as strong competitors to diffusion and flow models for image and video generation, achieving comparable output quality. The recent breakthrough in VAR lies in shifting the prediction paradigm from next-token prediction to *next-scale* prediction [66]. VAR represents an image as a pyramid of latent patches, structured from the lowest resolution (a single coarse patch) to progressively higher resolutions; within each level, VAR generates patches jointly, while across levels, finer-scale patches are generated autoregressively, conditioned on the previously generated coarser level. This shift to autoregressive next-scale prediction allowed causal autoregressive models to outperform diffusion models for the first time, while also being significantly faster than both diffusion models and raster-order autoregressive models. Recent works [20, 35] have further improved the speed of VAR, allowing for fast and high-quality image generation.

Recent works have developed large-scale VARs for text-conditioned image and video generation. However, text-conditioned VARs still lack diversity, often generating similar outputs when given the same prompt (Fig. 1). This lack of diversity has also been observed in other generative models such as diffusion models [59, 62] and GANs [41]. While various techniques such as CFG annealing [31] and CADS [59] have been proposed to improve diversity in diffusion models, improving diversity in text-conditioned VAR generation remains an open area of research. In this work, we address this issue in text-conditioned VARs through inference-time modifications, without requiring model retraining, fine-tuning, or substantial additional compute at generation.

Inspired by diversity-enhancement techniques in diffusion models [31, 59], we first explore their impact in the scale-wise autoregressive setting of VAR. These diversity-enhancement methods mainly fall into two categories. The first is CFG weight scheduling [23,31]: similar scheduling can be applied across scales during the autoregressive next-scale generation in VAR. The second technique is noise injection into the text embedding (also known as condition-annealing [59]), which can likewise be applied to VAR models. Both approaches involve a trade-off between diversity and quality; increased diversity typically leads to reduced quality. Empirically, we observe that the second approach (condition-annealing) is more effective for VAR models but results in reduced image quality.

To further mitigate the degradation in output quality, we propose a novel latent refinement framework called **Scale-Travel**. Motivated by image editing [45],

inpainting [42], and reward alignment techniques [44], which perform latent refinement by briefly reverting part of the generative trajectory before resuming generation, we explore a novel mechanism for VAR along the *scale axis*, which is non-trivial in the autoregressive generation process. Our method exploits the latent multi-scale autoencoder to enable an approximate scale rollback and refinement during next-scale autoregressive generation. Specifically, given a partial sequence of token scales, we encode their accumulation into a full sequence of token scales, and then retain only the first few scales. In this way, we can refine the partial sequence while effectively reverting to a previous scale by discarding the newly encoded finer-scale tokens. Incorporating this VAR-based "Scale-Travel" module into the autoregressive generation process significantly enhances realism, helping to overcome the diversity-realism trade-off caused by condition-annealing.

We empirically show that DiverseVAR yields the best diversity-quality tradeoff on both text-conditioned image and video generation and is robust across CFG scales and varying temperatures and nucleus sampling [24] parameters.

2 Related Work

2.1 Autoregressive Visual Generation

Building on the success of transformer-based autoregressive (AR) modeling in LLMs [6,69], researchers have increasingly explored similar techniques for image and video generation. In the image domain, AR methods differ primarily by the order in which they process local patches (or tokens). First, **(1) raster-scan order AR** methods take an analogous approach to language modeling by flattening and quantizing a 2D image into a 1D sequence of discrete tokens, and then training an AR model for next-token prediction [9,16,17,32,49,54–56,60,63,67,75,84]. Early works like PixelCNN [49] perform autoregressive generation in the pixel space, while more recent works, including VQGAN [16] and LlamaGen [63], perform it in the latent space [68]. Follow-up works propose alternative token ordering schemes to address the inflexibility of a fixed raster-scan order [10,50]. In the video domain, transformer-based AR models have also been actively explored, extending next-token prediction to spatiotemporal token sequences [13, 30, 73]. Going further, **(2) masked AR** methods remove the strict unidirectional raster-scan by randomly masking subsets of tokens and predicting them in parallel with full attention over the tokens [3,5,7,8,27,33,37,43,48,76,78,82,86]. Represented by MaskGIT [8], this parallel mask prediction introduces significant inference acceleration and also provides a connection between autoregressive modeling and discrete diffusion [2, 19], as demonstrated by Xie *et al.* [78].

Unlike the AR methods above, which predict individual tokens, visual autoregressive (VAR) model [66] represents an image as a sequence of multi-scale token maps. A transformer is trained to progressively predict each next higher-resolution token map. **(3) VAR** better respects the 2D grid structure of images, scales effectively, and outperforms both raster-scan order AR [16, 32], masked

AR [8], along with diffusion-based methods [15, 52] on ImageNet [14]. Recent VAR-based models, such as Infinity [21] and Switti [70] for images, as well as InfinityStar [40] for text-conditioned video generation, demonstrate that VAR scales effectively to large datasets. Ongoing work continues to introduce VAR variants [11, 20, 25, 28, 57, 83, 87, 88] and their adaptation to downstream applications like editing [71, 74], personalization [12], controllable generation [38, 51, 80], and style transfer [47]. Other works [20, 35, 36] improve the efficiency in VAR-based models, significantly decreasing memory cost and runtime.

2.2 Diversity Enhancement for Visual Generation

Enhancing diversity is a crucial task in visual generation for both diffusion and autoregressive (AR) models. Originally introduced in the context of diffusion models, CFG (Classifier-Free Guidance) sampling [23] is now widely used in both domains. Specifically in AR models, diversity has been promoted by sampling tokens probabilistically [63, 65]. However, the CFG technique tends to narrow the categorical token distribution, improving image quality but simultaneously diminishing output diversity. To mitigate mode collapse, dynamic CFG scheduling [31, 72] has proven effective for diffusion models. Another diversity enhancement technique for diffusion models is CADS [59], which demonstrated that injecting random Gaussian noise into the text embedding allows control over the quality–diversity trade-off during inference. To the best of our knowledge, we are the first to systematically increase the diversity of VAR models. In this work, we extend these approaches to VAR [21, 66] and investigate which components play a key role in enhancing diversity. Additionally, we introduce our novel method, DiverseVAR, a method designed to address the identified challenges.

3 Background: Next-Scale Visual Autoregressive Models

In this section, we provide background on next-scale visual autoregressive (VAR) models, outlining their key principles and architectural components. We then explain how conditional signals are integrated into VAR and discuss a critical drawback, namely the limited diversity of the generated visual content.

VAR: Next-Scale Prediction. Although autoregressive (AR) models extend the language-modeling paradigm to image generation, the standard next-token objective suffers from two major flaws. First, flattening an image or video into a 1D token sequence retains bidirectional dependencies, violating the unidirectional (*i.e.*, causal) assumption in AR sampling. Second, the flattened representation fails to exploit the inherent spatial locality of an image. To resolve these issues, Tian *et al.* [66] introduced VAR—*Visual Autoregressive Model*—which substitutes next-token prediction with a *next-scale* objective. Given an image I, a feature map $\mathbf{Z} \in \mathbb{R}^{H \times W \times D}$ is obtained through an encoder (*e.g.*, VQ-VAE [68]). Rather than patchifying the feature map, VAR employs *multi-scale encoding* in which $\mathbf{Z}$ is decomposed into K **residual token maps** $(\mathbf{r}_1, \cdots, \mathbf{r}_K)$ of sizes $\mathbf{s}_k = (h_k, w_k)$. At stage k the model constructs a **feature map** $\mathbf{Z}_k$ by

upsampling all preceding token maps to the final resolution $\mathbf{s}_K = (H, W)$ and aggregating them:

$$\mathbf{Z}_k = \sum_{i=1}^{k} \texttt{up}(\mathbf{r}_i, \mathbf{s}_K), \tag{1}$$

where $\texttt{up}(\cdot, \cdot)$ denotes bilinear upsampling. The joint probability of the token maps then factorizes over *scales* rather than token orders:

$$p(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_K) = \prod_{k=1}^{K} p(\mathbf{r}_k | \mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{k-1}). \tag{2}$$

with $\mathbf{r}_k \in \mathbb{R}^{h_k \times w_k \times D}$. This next-scale objective preserves spatial locality while preserving the unidirectional assumption for autoregressive models.

Conditional Generation with VAR. The high scalability of VAR has led to its rapid adoption in text-conditioned image generation [21, 64, 91, 92]. Text condition is incorporated into VAR through two complementary mechanisms: **(1) cross-attention with text embeddings**, and **(2) `<SOS>` token initialization**. In the former, the text embedding is first projected to obtain a condition embedding $\mathbf{c}$. $\mathbf{c}$ is then passed through projection networks to obtain key and value matrices which cross-attends with the image tokens, similar to cross-attention in diffusion models like Stable Diffusion [58]. For the latter, the text embedding is projected to produce the `<SOS>`—*Start-of-Sentence*—token $\mathbf{s}$ which initializes the generation as done in typical autoregressive models.

Lack of Diversity in VAR. In this work, we recognize a significant lack of diversity in text-conditioned VAR models. While VAR models make use of basic techniques such as temperature annealing and linear CFG scheduling to increase diversity, we find that these methods provide limited improvements. Motivated by this observation, we explore and propose new diversity-enhancement strategies to boost their diversity while maintaining the image quality and conditional fidelity in VAR models. Beyond generating high-quality images that match a given text prompt, a critical desideratum for text-conditioned image generation is to be able to provide *diverse* outputs from a single condition (*i.e.*, text prompt). This diversity not only offers users a range of options to choose from based on their preferences, but also acts as a prerequisite for test-time search algorithms [29, 77] for downstream applications. Although recent VAR-based models such as Infinity [21] achieve impressive quality, they suffer from severely limited diversity, as demonstrated in Fig. 1. In this work we adopt Sadat *et al.* [59] definition of diversity, which refers to the model's ability to generate varied outputs for a fixed condition $\mathbf{c}$ under different random seeds.

4 DiverseVAR

We present DiverseVAR, a training-free framework for enhancing diversity in text-conditioned VAR models. First, we explore inference-time techniques designed for diversity enhancements and compare their impact on diversity and

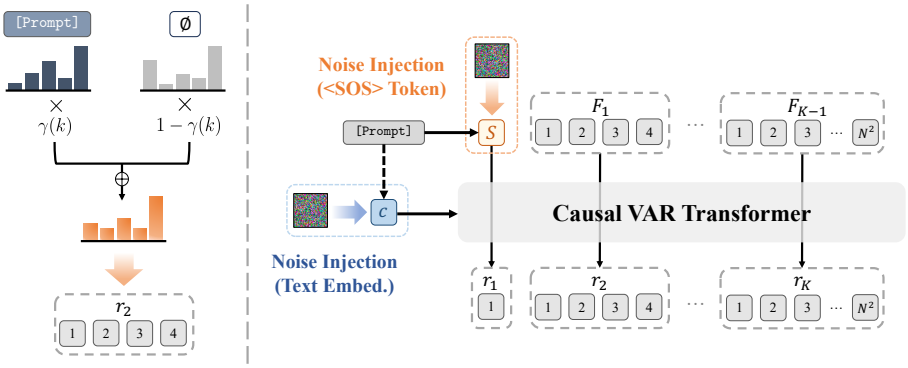

Fig. 2: Diversity-Enhancement Techniques. We explore two main options for diversity enhancement in VAR: (a) CFG-scheduling and (b) condition-annealing. CFG-scheduling modulates the CFG scale over sampling steps to mitigate mode collapse. Condition-annealing injects noise into the text-embedding or the <SOS> token.

quality in VAR (Sec. 4.1). Observing a critical trade-off between these metrics, we then propose a simple latent refinement strategy by taking advantage of the multi-scale structure of VAR (Sec. 4.2). Finally, we combine these two stages into a unified diversify-then-refine pipeline, achieving significant diversity gains with minimal quality degradation.

4.1 Exploring Diversity-Enhancement Techniques

In this section, we search for the best diversity-enhancement strategy for VAR models. Observing that VAR produces nearly identical samples for the same text prompt due to the condition signal dominating the generation process, we first explore two inference-time techniques popularly used to mitigate mode collapse as shown in Fig. 2: (1) custom classifier-free guidance (CFG) schedules [23], and (2) condition-annealing by injecting noise into the condition embedding [59].

Option 1: CFG Scheduling. CFG [23] modulates the influence of the condition in diffusion and VAR [21,66]. Formally, CFG adjusts the model's prediction via

$$\mathbf{Q}_{\text{CFG}} = (1+\omega)\mathbf{Q}_c - \omega\mathbf{Q}_\emptyset, \tag{3}$$

where $\mathbf{Q_c}$ is the conditional prediction (*i.e.*, score estimates in diffusion and logit outputs in VAR) given a condition $\mathbf{c}$, $\mathbf{Q}_\emptyset$ is the unconditional prediction, and ω is the guidance weight. Building on the trend that higher ω leads to higher condition alignment while sacrificing diversity, Kynkäänniemi *et al.* [31] show that carefully scheduling the guidance weight ω throughout generation can boost diversity while retaining image quality. While VAR models have employed

linear CFG scheduling, there has been no systematic study on its impact on diversity nor of the effects of other CFG schedules. We evaluate multiple CFG schedules $\omega(k)$ over the scale index $k \in [1, K]$ for VAR, analogous to timestep $t \in [0, T]$ in diffusion models.

- Piecewise Constant [31]: $\omega(k) = \mathbb{1}[k \in \mathcal{K}]\omega_1 + \mathbb{1}[k \notin \mathcal{K}]\omega_K$
  where $\mathcal{K} := \{k_{\min}, \ldots, k_{\max}\}$, and ω_1 and ω_K are tunable hyperparameters.
- Interpolation [72]: $\omega(k) = (1 - \gamma(k))\omega_1 + \gamma(k)\omega_K$
  Here, $\gamma(1) = 0$ and $\gamma(\ell) = 1$ for all $\ell \geq k_{\max}$, and we test several variants of the scheduling function $\gamma(\cdot)$ suggested by Wang *et al.* [72]. ω_1 and ω_K denote the initial and final guidance weight, respectively. Details on each scheduler and ablations for $k_{\max}$ are provided in the **supplementary**.

Option 2: Condition-Annealing via Noise Injection. Sadat *et al.* [59] introduce condition-annealing diffusion sampler (CADS), a technique that perturbs the condition by injecting noise into the text embedding, starting with full noise at the initial generation steps. This simple modification has proven effective for diversity in diffusion models as it weakens the condition's dominance in the sampling process. We explore condition-annealing in VAR on its two primary condition sources: (1) text embedding and (2) `<SOS>` token.

- Text Embedding: $\hat{\mathbf{c}} = \sqrt{1 - \alpha(k)}\mathbf{c} + \sqrt{\alpha(k)}\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I})$
  Following Sadat *et al.* [59], we add noise ϵ to the text embedding $\mathbf{c}$, modulated by an annealing schedule $\alpha(\cdot)$. We explore different choices for $\alpha(\cdot)$ in the **supplementary**.
- `<SOS>` Token: $\hat{\mathbf{s}} = \sqrt{1 - \beta(k)}\mathbf{s} + \sqrt{\beta(k)}\epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I})$
  Since autoregressive models prepend a special `<SOS>` token $\mathbf{s}$ to initialize the generation process, we also test injecting noise into $\mathbf{s}$ according to schedule $\beta(\cdot)$.

Empirical Analysis. We compare the diversity-quality tradeoff for each of the aforementioned options, and find that injecting noise into the text-embedding yields the best results, as shown in Fig. 6. Although the top-performing configuration does boost diversity, the outputs often exhibit noticeable degradations and visual artifacts (see Fig. 1 and Fig. 5). We hypothesize that these artifacts are an inherent side effect of perturbing the generation process via noise injection. To address this diversity-quality trade-off, in the following Sec. 4.2 we introduce a novel latent refinement method specifically designed for VAR.

4.2 Scale-Travel: Latent Refinement for VAR

In this section, we introduce a simple but effective refinement method designed for VAR. While condition-annealing (Sec. 4.1) significantly increases the diversity of VAR, it also introduces significant visual artifacts (second row of Fig. 5). Inspired by diffusion model refinement techniques [4, 42, 45, 81, 85], we introduce

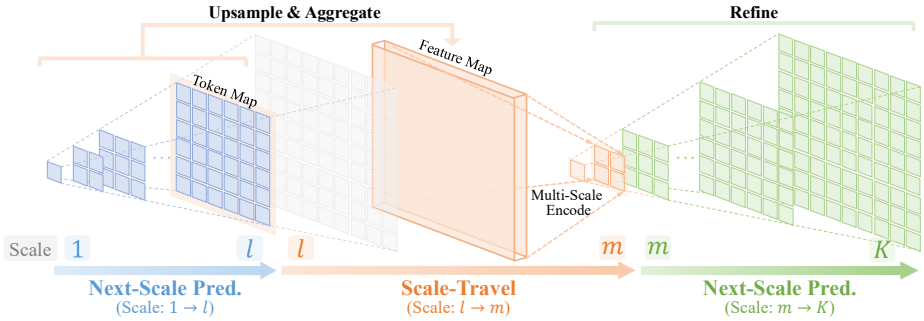

Fig. 3: Latent Refinement via Scale-Travel. We introduce scale-travel, a novel refinement strategy that leverages the multi-scale structure of VAR models. By reverting intermediate representations to a coarser scale and running generation without noise injection, clean and coherent details can be reconstructed. This process corrects visual artifacts and degradations while preserving the overall coarse structures.

Scale-Travel, a playback strategy for repairing corrupted latents in VAR based on its *multi-scale* nature.

Recall from Eq. 1 that in each stage l the model forms a feature map $\mathbf{Z}_l$ by upsampling and summing the token grids $\{\mathbf{r}_k\}_{k=1}^{l}$. After the noise injection of Sec. 4.1, $\mathbf{Z}_l$ may deviate from the model's image manifold. Scale-travel *rewinds* the feature map $\mathbf{Z}_l$ to an earlier (*i.e.*, coarser) scale $m < l$ and then resumes generation. To rewind the current feature map to an earlier scale, we utilize the *multi-scale encoder* of VAR. Recall that during VAR training, an image I is first encoded with a VQ-VAE [68] to obtain a dense feature map $\mathbf{Z}$, which is passed through a multi-scale encoder to be decomposed into scale-wise token grids

$$\mathbf{r}_{1:K} = \texttt{MultiScaleEnc}(\mathbf{Z}, \mathbf{s}_{1:K}) \tag{4}$$

where $\mathbf{s}_k$ is the grid size at stage k, as shown in Alg. 1. Here, we highlight a key property of VAR: early maps $\mathbf{Z}_l$ already contain the coarse layout and semantics of the final image (details included in the **supplementary**). While a similar observation has been presented in previous works [28,74], we further employ this finding to apply multi-scale encoding to $\mathbf{Z}_l$—instead of ground-truth feature $\mathbf{Z}$—to reconstruct the lower-resolution grids $\tilde{\mathbf{r}}_{1:m}$. We denote this technique as $\tilde{\mathbf{r}}_{1:m} = \texttt{ScaleTravel}(\mathbf{Z}_l, \mathbf{s}_{1:m}, m)$ and compare the exact pseudocode to that of multi-scale encoding in Fig. 4, showing that only a minor change in algorithm enables scale-travel. Once replayed to stage m, VAR's autoregressive generation proceeds from $k = m{+}1$ to K, effectively fixing the visual artifacts. We refer readers to the **supplementary** for more details and visualization on the trajectory of scale-travel. The full latent refinement technique is shown in Fig. 3 and can be written as

$$\texttt{Scale-Travel:} \quad \tilde{\mathbf{r}}_{1:m} = \texttt{ScaleTravel}(\mathbf{Z}_l, \mathbf{s}_{1:m}, m) \tag{5}$$

$$\texttt{Refine:} \quad \hat{\mathbf{r}}_{m+1:K} \sim \prod_{k=m+1}^{K} p(\mathbf{r}_k | \tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_m, \mathbf{r}_{m+1}, \ldots, \mathbf{r}_{k-1}) \tag{6}$$

Algorithm 1: Multi-Scale Encode

Params : Latent $\mathbf{Z}$, Scales $\mathbf{s}_{1:K}$

```
def MultiScaleEnc(Z, s_{1:K})
    R ← [];
    for k = 1, 2, ..., K do
        r_k ← Q(down(Z − Z_{k−1}, s_k));
        R.append(r_k);
        Z_k ← Σ_{i=1}^{k} up(r_i, s_K);
    return R = r_{1:K}
```

Algorithm 2: Scale-Travel

Params : Latent $\mathbf{Z}_l$, Scales $\mathbf{s}_{1:m}$, Target m

```
def ScaleTravel(Z_l, s_{1:m}, m)
    R ← [];
    for k = 1, 2, ..., m do
        r_k ← Q(down(Z_l − Z_{k−1}, s_k));
        R.append(r_k);
        Z_k ← Σ_{i=1}^{k} up(r_i, s_K);
    return R = r_{1:m}
```

Fig. 4: Scale-Travel Pseudocode. Side-by-side pseudocode for **(left)** multi-scale encoding process in VAR [66] and **(right)** our proposed scale-travel step in Sec. 4.2. $\texttt{up}(\cdot,\cdot)$ and $\texttt{down}(\cdot,\cdot)$ denote bilinear upsampling and downsampling, respectively.

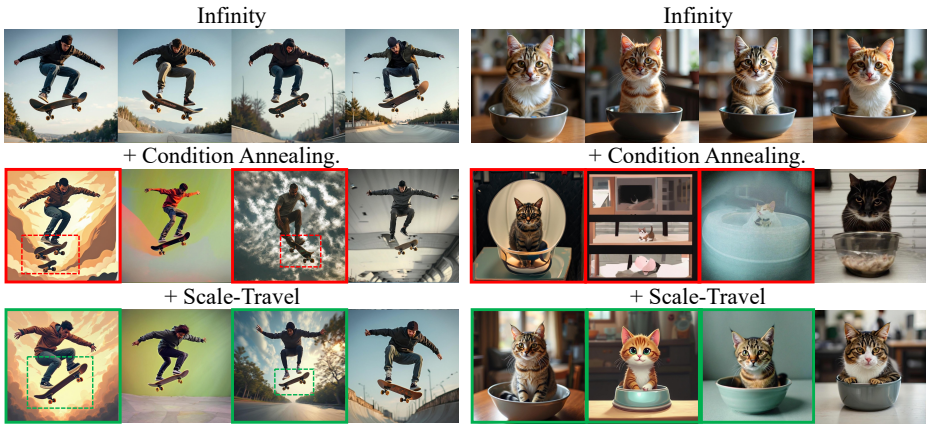

Fig. 5: Qualitative Comparisons on Image Generation. Infinity produces images with little variation and diversity (row 1). Injecting noise into the text-embedding increases diversity (row 2) but results in visual artifacts (red box). Applying our **scale-travel** refinement technique (row 3) fixes these visual artifacts (green box) while retaining diversity.

5 Results

In this section, we present experimental results of DiverseVAR, comparing diversity and quality with existing VAR diversity enhancement techniques.

5.1 Evaluation Setup

VAR Models. For text-conditioned image generation VAR models, we use Infinity [21] and Switti [70]. Infinity trains a bitwise transformer for scaling the vocabulary to infinite size, and Switti improves the efficiency and scalability of VAR [66] by addressing training instability and using non-causal self-attention

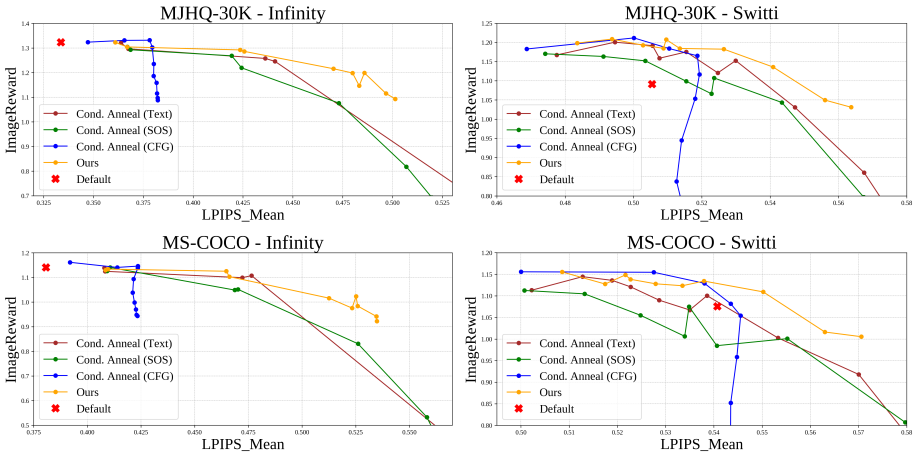

Fig. 6: Pareto Fronts for Diversity-Quality Trade-Off on Infinity [21] and Switti [70]. Each curve is obtained by varying hyperparameters for each method: noise scales for SOS and CADS, guidance schedules for CFG, and target scales for DiverseVAR (Ours). The default setting of each model is included as a reference. On both MJHQ-30K [34] and MS-COCO [39], DiverseVAR consistently sets the Pareto frontier, demonstrating the best balance between diversity and quality in VAR.

for faster sampling. For text-conditioned video generation with VAR models, we use InfinityStar-8B [40], which generates 480p videos with 81 frames. Since InfinityStar follows a two-stage pipeline, first generating keyframes and then performing temporal generation, we apply DiverseVAR to the keyframe generation stage.

Benchmarks and Evaluation Metrics. For text-conditioned image generation, we evaluate on MS-COCO [39] and MJHQ-30K [34] datasets, each of which consists of image-caption pairs. Note that MS-COCO consists of realistic photos, while MJHQ-30K is constructed with images generated by Midjourney [46]. For text-conditioned video generation, we evaluate all methods on VBench [26, 90] which evaluates video generative models across multiple dimensions. We generate videos based on VBench-2.0 [90] prompts. We use multiple metrics to measure image/video quality and diversity as follows:

- Image Diversity: **LPIPS-MPD** [1] computes the mean pairwise LPIPS [89] distance between images to measure per-prompt diversity.
- Image Diversity: **Vendi-Score** [18] computes the effective number of "dissimilar" elements in a sample of images. Following CADS [59], we use SSCD [53] to measure pairwise dissimilarities.
- Image Quality: **ImageReward** [79] evaluates image quality and alignment based on human preference.
- Image Quality & Diversity: **FID** [22] simultaneously captures both diversity and quality with respect to a reference set of images.

– Video Diversity: **VBench-2.0 Diversity** [90] computes video diversity using inter-sample style and content variation with a pretrained VGG-19 [61].
– Video Quality: **VBench Quality Score** [26] is used to evaluate video generation quality based on frame-wise metrics and temporal metrics. Results for each VBench category are provided in the **supplementary** material.

5.2 Pareto Fronts on Diversity-Quality Trade-Off

We assess DiverseVAR's diversity–quality balance for Infinity [21] and Switti [70] by plotting Pareto fronts on the metrics above.

Settings. We compare DiverseVAR against the three diversity enhancement techniques from Sec. 4.1: CFG scheduling and condition-annealing (*i.e.*, noise injection) with text embeddings and `<SOS>` tokens. For each dataset [34, 39], we sample 100 prompts and generate 10 images per prompt to trace each Pareto curve in Fig. 6. We use 1000 prompts and generate 10 images per prompt for Tab. 1. Our DiverseVAR pairs scale-travel with text-embedding annealing, which we observed to be the most effective option. Specifically, we fix $k_{\min} = 1$ and vary $k_{\max} \in \{2, 3, 4, 5, 6\}$ with the maximum noise scales $\alpha(1), \beta(1) \in \{0.5, 1.0\}$ for condition-annealing on both text embedding and `<SOS>`. For CFG scheduling, we set $k_{\max} \in \{1, 2, \ldots, 10\}$ and use each model's default guidance scale as w_K. For DiverseVAR, scale-travel starts at scale $l = 8$, where we observed VAR forms an image's global structure and semantics. We use a constant scheduler for CFG scheduling and a cosine scheduler for condition-annealing. More details and ablation studies on the scheduler, including specific hyperparameters used, are provided in the **supplementary**.

Results. Fig. 6 shows the Pareto fronts for all methods. Across both benchmarks and different metrics, DiverseVAR consistently defines the frontier, underscoring that our scale-travel effectively refines the output image quality while retaining the diversity introduced by condition-annealing. CFG scheduling increases the diversity but also significantly degrades the image quality. Injecting noise into the text embedding or `<SOS>` token enhances diversity beyond what CFG scheduling can offer, by preventing the dominance of the conditioning signal during early generation steps. Among these condition-annealing techniques, noise injection to the text embedding yields the best diversity-quality trade-off, yet results in clear quality degradation—as shown in the sharp drop in the Pareto front as diversity increases. Our DiverseVAR overcomes this drawback by applying scale-travel to recover image quality while preserving the added diversity, thereby achieving the Pareto frontier.

5.3 Quantitative and Qualitative Comparisons

We additionally report quantitative and qualitative evaluations of DiverseVAR. For this, we use the optimal hyperparameter settings for each method, selected based on a toy experiment, which we report in detail in the **supplementary**. Since we find that injecting noise at every scale results in severe quality degradation, we adopt a scheduling strategy with a threshold $k_{\max} < K$ [59].

Method	MJHQ-30K [34]				MS-COCO [39]			
	FID↓	IR↑	MPD_L ↑	Vendi ↑	FID↓	IR↑	MPD_L ↑	Vendi ↑
Infinity [21]	19.16	1.22	0.33	4.10	37.37	1.16	0.37	4.41
+ Cond. Anneal.	15.53	0.49	0.56	7.68	22.51	0.39	0.58	8.13
+ Scale-Travel	15.28	1.08	0.48	6.07	28.97	1.04	0.52	6.68
Switti [70]	16.18	1.11	0.44	4.79	25.92	1.18	0.46	5.32
+ Cond. Anneal.	21.34	0.63	0.58	7.34	25.26	0.78	0.59	7.74
+ Scale-Travel	17.30	0.93	0.56	6.68	23.47	1.05	0.57	7.22

Table 1: Quantitative Comparisons on Image Generation. We evaluate four metrics: ImageReward (IR) [79] for quality, FID [22] for both quality and diversity, and LPIPS-MPD (MPD_L) [89], Vendi [18] for diversity. For both Switti [70] and Infinity [21], condition annealing enhances diversity but noticeably degrades quality. In contrast, DiverseVAR's scale-travel refinement finds a more favorable trade-off, balancing quality and diversity.

Image Generation. As shown in Tab. 1, our method generalizes to different text-to-image VARs. Starting from the original model, condition annealing significantly enhances diversity. Notably, adding noise to the text embedding (Cond. Anneal.) yields a substantial diversity gain in LPIPS-MPD [1] on the MJHQ-30K [34] dataset, +69.7% improvement over the baseline Infinity. However, this comes with a clear trade-off in terms of image quality. Our scale-travel method significantly enhances image quality from noisy samples, achieving a +120.41% improvement on ImageReward [79], while incurring only a modest reduction in diversity (−14.29% in LPIPS-MPD). Compared to the base model, DiverseVAR yields a +45.45% improvement in LPIPS-MPD with minimal quality degradation for ImageReward (−11.48%). A similar trend is also observed for Switti. This underscores the effectiveness of our refinement technique: starting from diverse but low-quality samples, scale-travel improves image quality while retaining a substantial diversity gain over standard text-conditioned VARs, with only a minor decrease in quality.

Moreover, in Fig. 5, we present qualitative results from each stage of our DiverseVAR pipeline. The outputs from the default model (row 1) exhibit a severe lack of diversity across the same text condition. After injecting noise into the text conditions, we observe a substantial increase in diversity compared to the previous stage (row 2). However, this also leads to noticeable quality degradation, often failing to generate clear objects or preserve fine details. The last row shows the final output of DiverseVAR, demonstrating that our method not only corrects the corrupted details introduced by condition-annealing but also preserves the high-level semantics. More qualitative results are included in the **supplementary**.

Video Generation. As shown in Tab. 2, DiverseVAR improves diversity while largely preserving overall VBench performance, yielding a more favorable trade-off than condition annealing. Compared to the baseline In-

Method	Diversity	Quality Score
InfinityStar [40]	0.270	0.806
+ Cond. Anneal.	0.627	0.794
+ Scale-Travel	0.487	0.804

Table 2: Quantitative results on InfinityStar [40] using VBench [26, 90].

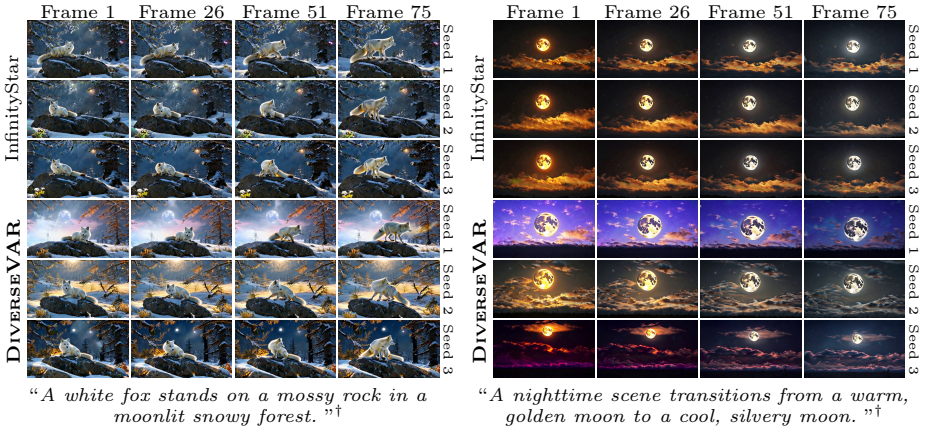

Fig. 7: Qualitative Comparisons on Video Generation. InfinityStar [40] produces videos with little variation and diversity (row 1-3). Injecting noise into the text-embedding and applying our **scale-travel** refinement technique (row 4-6) can significantly increase per-prompt video diversity. † denotes using augmented prompt for generation.

finityStar, condition annealing delivers a larger diversity increase (+132.2%) but reduces the quality score. In contrast, DiverseVAR raises diversity from 0.270 to 0.487 (+80.4%) with a negligible change in quality (−0.002 points).

In addition, Fig. 7 provides qualitative comparisons between the original InfinityStar [40] and our DiverseVAR pipeline across multiple random seeds. For a fixed prompt, the original model (rows 1–3) exhibits limited diversity, producing nearly identical videos across different seeds. In contrast, DiverseVAR (rows 4–6) yields substantially more varied outputs over the entire sequence. The differences are particularly evident in color attributes and scene composition: InfinityStar [40] tends to reproduce the same background and object layout with minimal variation, while our method generates diverse videos and maintains overall coherence. More qualitative results are included in the **supplementary**.

5.4 Comparison with Alternative Diversity Controls

CFG Scale, Temperature, and Nucleus Sampling. We also provide additional results under varying CFG weights, temperatures, and nucleus sampling parameters [24]. As shown in Tab. 3, DiverseVAR consistently improves the diversity-quality tradeoff across all CFG weights. As shown in Tab. 4 and Tab. 5, DiverseVAR also yields a better diversity-quality tradeoff than simply varying the temperature (Tab. 4) or using nucleus sampling (Tab. 5).

Prompt Rewriting. Prompt rewriting uses an LLM to modify the input prompt to improve the generation quality and diversity. While this can introduce additional variation without much quality degradation, prompt rewriting changes the condition itself, whereas DiverseVAR aims to increase diversity under the

ω	Method	MPD ↑	Vendi ↑	IR ↑	ω	Method	MPD ↑	Vendi ↑	IR ↑
2.0†	Infinity	0.334	4.031	1.323	2.0	Switti	0.465	5.262	1.100
	DiverseVAR	0.486	6.087	1.199		DiverseVAR	0.568	7.034	0.919
4.0	Infinity	0.324	3.862	1.367	4.0	Switti	0.448	4.921	1.165
	DiverseVAR	0.486	6.127	1.215		DiverseVAR	0.565	6.705	1.035
6.5	Infinity	0.320	3.804	1.384	6.0†	Switti	0.444	4.744	1.210
	DiverseVAR	0.481	6.101	1.234		DiverseVAR	0.564	6.609	1.031
9.0	Infinity	0.320	3.770	1.406	9.0	Switti	0.438	4.668	1.225
	DiverseVAR	0.476	6.019	1.239		DiverseVAR	0.562	6.557	0.974

Table 3: DiverseVAR with Varying CFG Weights. We vary the CFG guidance weights ω for the VAR models and DiverseVAR. DiverseVAR consistently improves the diversity-quality tradeoff across all CFG weights.

Infinity [21]	MPD ↑	Vendi ↑	IR ↑	**Switti** [70]	MPD ↑	Vendi ↑	IR ↑
$\tau = 0.5$	0.284	3.500	1.305	$\tau = 0.5$	0.437	4.483	1.204
$\tau = 1^{\dagger}$	0.334	4.031	1.323	$\tau = 1^{\dagger}$	0.444	4.744	1.210
$\tau = 2$	0.375	4.577	1.331	$\tau = 2$	0.453	5.014	1.095
$\tau = 5$	0.414	4.813	1.108	$\tau = 5$	0.477	5.613	0.813
$\tau = 10$	0.459	4.954	-0.430	$\tau = 10$	0.487	5.897	0.619
DiverseVAR	0.470	5.833	1.216	DiverseVAR	0.563	6.609	1.031

Table 4: VAR models with varying temperatures. While increasing τ leads to larger diversity, the image quality (IR) is significantly degraded. DiverseVAR achieves the highest diversity while at the same time maintaining image quality.

Infinity [21]	MPD ↑	Vendi ↑	IR ↑	**Switti** [70]	MPD ↑	Vendi ↑	IR ↑
$p = 0.9$	0.323	3.934	1.320	$p = 0.9$	0.441	4.714	1.206
$p = 0.95$	0.331	4.016	1.329	$p = 0.95^{\dagger}$	0.444	4.744	1.210
$p = 0.97^{\dagger}$	0.334	4.031	1.323	$p = 0.97$	0.444	4.763	1.205
$p = 1.0$	0.335	4.071	1.330	$p = 1.0$	0.445	4.771	1.200
DiverseVAR	0.470	5.833	1.216	DiverseVAR	0.563	6.609	1.031

Table 5: VAR models with varying nucleus sampling parameters [24]. Varying p does not significantly impact the diversity. DiverseVAR achieves significantly higher diversity without significantly degrading the image quality.

same fixed prompt. However, in practical settings where prompts are highly detailed, prompt rewriting itself has less room to introduce meaningful variation, leading to only limited changes in the generated outputs as shown in Fig. 8. Moreover, prompt rewriting is also complementary to DiverseVAR, as it helps preserve image quality when combined with our method, while DiverseVAR still provides substantial diversity gains. Additional discussion and quantitative results comparing DiverseVAR with prompt rewriting are included in the **supplementary**.

Supplementary. Further details and extended results are provided in the supplementary material, including analyses of scale-travel, runtime overhead, additional metrics and comparisons, and extended qualitative and ablation results.

† default hyperparameter used in the original paper.

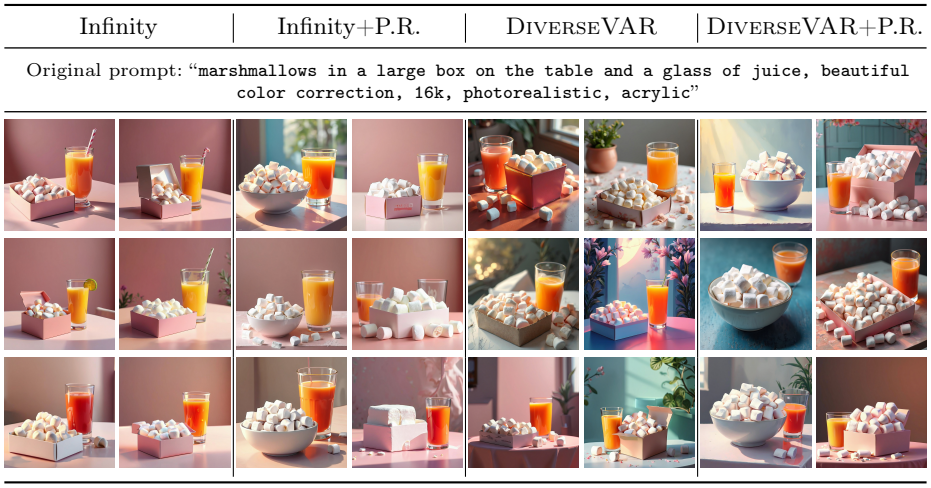

Fig. 8: Qualitative results for prompt rewriting. We compare Infinity, Infinity with prompt rewriting (Infinity+P.R.), DiverseVAR, and DiverseVAR with prompt rewriting (DiverseVAR+P.R.) under the same original prompt.

6 Conclusion

In this work, we first report a significant lack of diversity in text-conditioned VAR models, an issue that has not been addressed in prior work. Inspired by diversity enhancement techniques in diffusion models, we explore these methods and find that injection into the text embedding yields the most diverse results. However, this diversity enhancement comes at the cost of significant degradation in image quality. To address this, we propose a novel **scale-travel** latent refinement method tailored to the multi-scale nature of VAR, aiming to improve image quality while preserving diversity. Experimental results show that our two-stage method, DiverseVAR, achieves the best diversity–quality Pareto frontier.

Limitations. Although our scale-travel refinement technique improves the Pareto frontier, the inference cost is slightly higher than using noise injection. We also observe that the most diverse results still suffer from visual artifacts, occurring in approximately 6% of the generated samples.

Acknowledgements.

This work was supported by an NRF grant (RS-2026-25486000); IITP grants (RS-2024-00399817, RS-2025-25441313, RS-2025-25443318, RS-2026-25526850) funded by MSIT, Korea; the Industrial Technology Innovation Program grant (RS-2025-02317326) funded by MOTIE, Korea; the InnoCORE program of MSIT (AI Meta-Scientist, N10260110); the National Supercomputing Center with supercomputing resources and technical support (KSC-2025-CRE-0475); the Advanced GPU Utilization Support Program; and the DRB-KAIST SketchTheFuture Research Center.

References

1. Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H.H., Jin, K.H., Kim, S.: Self-rectifying diffusion sampling with perturbed-attention guidance. In: ECCV (2024)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS (2021)
3. Bai, J., Ye, T., Chow, W., Song, E., Chen, Q.G., Li, X., Dong, Z., Zhu, L., YAN, S.: Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In: ICLR (2025)
4. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: CVPR (2023)
5. Besnier, V., Chen, M., Hurych, D., Valle, E., Cord, M.: Halton scheduler for masked generative image transformer. In: ICLR (2025)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
7. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. In: ICML (2023)
8. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: CVPR (2022)
9. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML (2020)
10. Chen, X., Mishra, N., Rohaninejad, M., Abbeel, P.: Pixelsnail: An improved autoregressive generative model. In: ICML (2018)
11. Chen, Z., Ma, X., Fang, G., Wang, X.: Collaborative decoding makes visual autoregressive modeling efficient. arXiv preprint arXiv:2411.17787 (2024)
12. Chung, J., Hyun, S., Kim, H., Koh, E., Lee, M., Heo, J.P.: Fine-tuning visual autoregressive models for subject-driven generation. arXiv preprint arXiv:2504.02612 (2025)
13. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. In: ICLR (2025)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
17. Fan, L., Li, T., Qin, S., Li, Y., Sun, C., Rubinstein, M., Sun, D., He, K., Tian, Y.: Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. arXiv preprint arXiv:2410.13863 (2024)
18. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. arXiv preprint arXiv:2210.02410 (2022)
19. Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R.T., Synnaeve, G., Adi, Y., Lipman, Y.: Discrete flow matching. In: NeurIPS (2024)
20. Guo, H., Li, Y., Zhang, T., Wang, J., Dai, T., Xia, S.T., Benini, L.: Fastvar: Linear visual autoregressive modeling via cached token pruning. arXiv preprint arXiv:2503.23367 (2025)

21. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: CVPR (2025)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
23. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
24. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: ICLR (2020)
25. Huang, Z., Qiu, X., Ma, Y., Zhou, Y., Zhang, C., Li, X.: Nfig: Autoregressive image generation with next-frequency prediction. arXiv preprint arXiv:2503.07076 (2025)
26. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
27. Hur, J., Lee, D., Han, G., Choi, J., Jeon, Y., Kim, J.: Unlocking the capabilities of masked generative models for image synthesis via self-guidance. In: NeurIPS (2024)
28. Jiao, S., Zhang, G., Qian, Y., Huang, J., Zhao, Y., Shi, H., Ma, L., Wei, Y., Jie, Z.: Flexvar: Flexible visual autoregressive modeling without residual prediction. arXiv preprint arXiv:2502.20313 (2025)
29. Kim, J., Yoon, T., Hwang, J., Sung, M.: Inference-time scaling for flow models via stochastic generation and rollover budget forcing. arXiv preprint arXiv:2503.19385 (2025)
30. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: a large language model for zero-shot video generation. In: ICML (2024)
31. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In: arXiv preprint arXiv:2404.07724 (2024)
32. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: CVPR (2022)
33. Lezama, J., Chang, H., Jiang, L., Essa, I.: Improved masked image generation with token-critic. In: ECCV (2022)
34. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024)
35. Li, J., Ma, Y., Zhang, X., Wei, Q., Liu, S., Zhang, L.: Skipvar: Accelerating visual autoregressive modeling via adaptive frequency-aware skipping. arXiv preprint arXiv:2506.08908 (2025)
36. Li, K., Chen, Z., Yang, C.Y., Hwang, J.N.: Memory-efficient visual autoregressive modeling with scale-aware kv cache compression. In: NeurIPS (2025)
37. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: NeurIPS (2024)
38. Li, X., Qiu, K., Chen, H., Kuen, J., Lin, Z., Singh, R., Raj, B.: Controlvar: Exploring controllable visual autoregressive modeling. arXiv preprint arXiv:2406.09750 (2024)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

40. Liu, J., Han, J., Yan, B., Zhu, F., Wang, X., Jiang, Y., PENG, B., Yuan, Z., et al.: Infinitystar: Unified spacetime autoregressive modeling for visual generation. In: NeurIPS (2025)
41. Liu, S., Wang, T., Bau, D., Zhu, J.Y., Torralba, A.: Diverse image generation via self-conditioned gans. In: CVPR (2020)
42. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022)
43. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410 (2024)
44. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 (2025)
45. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
46. Midjourney Team: Midjourney. `https://www.midjourney.com/` (2022), accessed: 2025-05-16
47. Nguyen, Q.B., Luu, M., Nguyen, Q., Tran, A., Nguyen, K.: CSD-VAR: Content-Style Decomposition in Visual Autoregressive Models. In: ICCV (2025)
48. Ni, Z., Wang, Y., Zhou, R., Guo, J., Hu, J., Liu, Z., Song, S., Yao, Y., Huang, G.: Revisiting non-autoregressive transformers for efficient image synthesis. In: CVPR (2024)
49. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: NeurIPS (2016)
50. Pang, Z., Zhang, T., Luan, F., Man, Y., Tan, H., Zhang, K., Freeman, W.T., Wang, Y.X.: Randar: Decoder-only autoregressive visual generation in random orders. arXiv preprint arXiv:2412.01827 (2024)
51. Park, J., Gim, J., Lee, K., Oh, M., Choi, M., Kim, J., Park, W.C., Im, S.: A training-free style-aligned image generation with scale-wise autoregressive model. arXiv preprint arXiv:2504.06144 (2025)
52. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
53. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: CVPR (2022)
54. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
55. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: NeurIPS (2019)
56. Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Chen, Y., Belov, D., Freitas, N.: Parallel multiscale autoregressive density estimation. In: ICML (2017)
57. Ren, S., Yu, Y., Ruiz, N., Wang, F., Yuille, A., Xie, C.: M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. arXiv preprint arXiv:2411.10433 (2024)
58. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
59. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In: ICLR (2024)

60. Salimans, T., Karpathy, A., Chen, X., Kingma, D.P.: Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In: ICLR (2017)
61. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
62. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. In: NeurIPS (2023)
63. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
64. Tang, H., Wu, Y., Yang, S., Xie, E., Chen, J., Chen, J., Zhang, Z., Cai, H., Lu, Y., Han, S.: HART: Efficient visual generation with hybrid autoregressive transformer. In: ICLR (2025)
65. Teng, Y., Shi, H., Liu, X., Ning, X., Dai, G., Wang, Y., Li, Z., Liu, X.: Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. In: ICLR (2025)
66. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NeurIPS (2024)
67. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: ICML (2016)
68. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017)
69. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
70. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. arXiv preprint arXiv:2412.01819 (2024)
71. Wang, J., Chen, Y., Yu, J., Lu, G., Pei, W.: Editinfinity: Image editing with binary-quantized generative models. In: NeurIPS (2025)
72. Wang, X., Dufour, N., Andreou, N., CANI, M.P., Abrevaya, V.F., Picard, D., Kalogeiton, V.: Analysis of classifier-free guidance weight schedulers. TMLR (2024)
73. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
74. Wang, Y., Guo, L., Li, Z., Huang, J., Wang, P., Wen, B., Wang, J.: Training-free text-guided image editing with visual autoregressive model. arXiv preprint arXiv:2503.23897 (2025)
75. Wang, Y., Lin, Z., Teng, Y., Zhu, Y., Ren, S., Feng, J., Liu, X.: Bridging continuous and discrete tokens for autoregressive visual generation. arXiv preprint arXiv:2503.16430 (2025)
76. Weber, M., Yu, L., Yu, Q., Deng, X., Shen, X., Cremers, D., Chen, L.C.: Maskbit: Embedding-free image generation via bit tokens. TMLR (2025)
77. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025)
78. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: ICLR (2025)
79. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. In: NeurIPS (2023)

80. Yao, Z., Li, J., Zhou, Y., Liu, Y., Jiang, X., Wang, C., Zheng, F., Zou, Y., Li, L.: Car: Controllable autoregressive modeling for visual generation. arXiv preprint arXiv:2410.04671 (2024)
81. Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J.Y., Ermon, S.: Tfg: Unified training-free guidance for diffusion models. In: NeurIPS (2024)
82. You, Z., Ou, J., Zhang, X., Hu, J., Zhou, J., Li, C.: Effective and efficient masked image generation models. arXiv preprint arXiv:2503.07197 (2025)
83. Yu, H., Luo, H., Yuan, H., Rong, Y., Zhao, F.: Frequency autoregressive image generation with continuous tokens. arXiv preprint arXiv:2503.05305 (2025)
84. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. TMLR (2022)
85. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: ICCV (2023)
86. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: CVPR (2023)
87. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. arXiv preprint arXiv:2411.00776 (2024)
88. Zhang, Q., Dai, X., Yang, N., An, X., Feng, Z., Ren, X.: Var-clip: Text-to-image generator with visual auto-regressive modeling. arXiv preprint arXiv:2408.01181 (2024)
89. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
90. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
91. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327 (2025)
92. Zhuang, X., Xie, Y., Deng, Y., Yang, D., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt-v1. 1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning. arXiv preprint arXiv:2504.02949 (2025)