arXiv:2511.21415v1 [cs.CV] 26 Nov 2025

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

Mingue Park* Prin Phunyaphibarn* Phillip Y. Lee Minhyuk Sung
KAIST
{kicikicik, prin10517, phillip0701, mhsung}@kaist.ac.kr

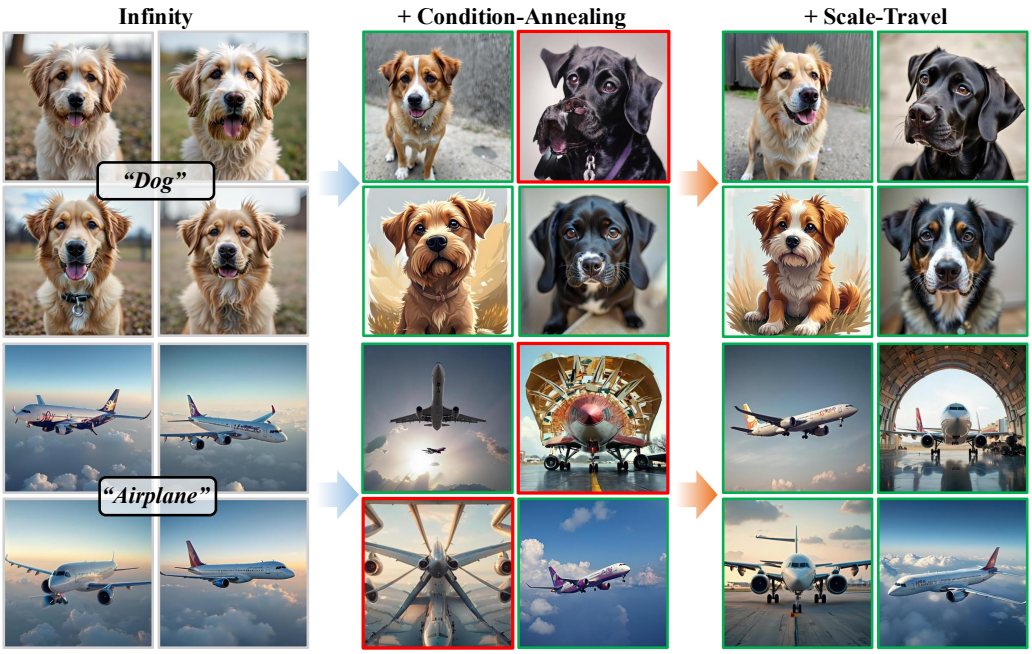

Figure 1. **DIVERSEVAR.** Infinity [21], a text-conditioned VAR model, exhibits a clear lack of per-prompt diversity (left). Adapting condition-annealing techniques from diffusion models [28, 55] enhances diversity yet noticeably degrades quality, as displayed in the red boxes (middle). DIVERSEVAR pairs these annealing strategies with a novel VAR-specific latent refinement technique, finding a sweet spot between diversity and quality of VAR-based image generation (right).

Abstract

We introduce DIVERSEVAR*, a framework that enhances the diversity of text-conditioned visual autoregressive models (VAR) at test time without requiring retraining, fine-tuning, or substantial computational overhead. While VAR models have recently emerged as strong competitors to diffusion and flow models for image generation, they suffer from a critical limitation in diversity, often producing nearly identical images even for simple prompts. This issue has largely gone unnoticed amid the predominant focus on image quality. We address this limitation at test time in two stages. First, inspired by diversity enhancement techniques in diffusion models, we propose injecting noise into the text embedding. This introduces a trade-off between diversity and image quality: as diversity increases, the image quality sharply declines. To preserve quality, we propose* ***scale-travel****: a novel latent refinement technique inspired by "time-travel" strategies in diffusion models. Specifically, we use a multi-scale autoencoder to extract coarse-scale tokens that enable us to resume generation at intermediate stages. Extensive experiments show that combining text-embedding noise injection with our scale-travel refinement significantly enhances diversity while minimizing image-quality degradation, achieving a new Pareto frontier in the diversity–quality trade-off.*

*Equal contribution.

1. Introduction

Visual autoregressive (VAR) models [21, 26, 35, 59, 61, 83, 85] have recently emerged as strong competitors to diffusion and flow models for image generation, achieving comparable image quality. Until recently, the inherent unidirectional bias of autoregressive models [58] was considered inadequate for capturing bidirectional correlations among image patches. However, a recent breakthrough in VAR lies in shifting the prediction paradigm from next-token prediction to *next-scale* prediction. 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, 32] have further improved the speed of VAR, allowing for fast and high-quality image generation.

Incorporating text conditions into VAR models is straightforward, as existing text-conditioning techniques from diffusion models can be directly applied in a similar manner. For example, the cross-attention mechanism used by latent diffusion models [54] for conditional generation can be directly applied to VAR. Classifier-Free Guidance (CFG) [23] can also be directly applied to VARs [21, 59, 61]; specifically, just as diffusion models predict and combine scores from conditional and unconditional cases, VARs can likewise predict and combine conditional and unconditional patch classification logits.

Despite the improved realism achieved by VAR, text-to-image VARs still lack diversity, often generating similar images when given the same textual prompt (Fig. 1). This lack of diversity has been observed in other generative models such as diffusion models [55, 57] and GANs [38] as well. However, while various techniques such as CFG annealing [28] and CADS [55] have been proposed to improve the diversity in diffusion models, the lack of diversity in VAR text-to-image generation remains an open area of research.

In this work, we address the lack-of-diversity issue in text-conditioned VARs through simple inference-time modifications, without requiring model retraining, fine-tuning, or substantial compute at generation. Inspired by diversity-enhancement techniques in diffusion models [28, 55], we first explore the impact of these techniques when applied to VAR. The diversity-enhancement methods from diffusion models mainly fall into two categories. The first is CFG weight scheduling [23, 28]: just as CFG weights vary during the denoising process in diffusion models, similar scheduling can be applied during the autoregressive next-scale generation in VAR. The second technique is noise injection into the text embedding (also known as condition-annealing [55]), which can likewise be applied directly 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 (noise injection) is more effective for VAR models but results in reduced image quality.

To further mitigate the reduced image quality, we propose a novel latent refinement framework called **Scale-Travel**. Motivated by image editing [42], inpainting [39], and reward alignment techniques [41] for diffusion models—which involve a "time-travel" mechanism by briefly running the generative process in reverse (*i.e.*, the forward process in diffusion) before resuming generation—we explore an analogous mechanism for VAR based on the analogy between the denoising process in diffusion models and the scale-wise autoregressive generation in VAR. Although directly reversing the autoregressive process is infeasible, we propose leveraging the latent multi-scale autoencoder for approximate reversion. 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 noise injection.

2. Related Work

2.1. Autoregressive Visual Generation

Building on the success of the transformer-based autoregressive (AR) modeling in LLMs [6, 65], researchers have increasingly explored similar techniques for image 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, 15, 16, 29, 50–52, 56, 58, 62, 63, 70, 79]. Early works like PixelCNN [62] perform autoregressive generation in the pixel space, while more recent works including VQGAN [15] and LlamaGen [58] perform it in the latent space [64]. Follow-up works propose alternative token ordering schemes to address the inflexibility of a fixed raster-scan order [10, 46]. 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, 25, 30, 34, 40, 45, 71, 73, 77, 81]. Represented by MaskGIT [7], this parallel mask prediction introduces significant inference acceleration and also pro-

vides a connection between autoregressive modeling and discrete diffusion [2, 18], as demonstrated by Xie *et al.* [73].

Unlike the AR methods above, which predict individual tokens, visual autoregressive (VAR) model [61] 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 [15, 29], masked AR [7], along with diffusion-based methods [14, 48] on ImageNet [13]. Recent VAR-based models—such as Infinity [21] and Switti [66]—demonstrate that VAR scales effectively to large datasets, and ongoing work continues to introduce VAR variants [11, 20, 24, 26, 53, 78, 82, 83] and their adaptation to downstream applications like editing [67, 69], personalization [12], controllable generation [35, 47, 75], and style transfer [44]. Other works [20, 32, 33] have also improved the efficiency in VAR-based models, significantly decreasing memory cost and runtime.

2.2. Diversity in Image Generation

Enhancing diversity is a crucial task in image 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 [58, 60]. 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 [28, 68] has proven effective for diffusion models. Another diversity enhancement technique for diffusion models is CADS [55], 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, 61] 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—the limited diversity of the generated images.

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 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. [61] 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 [64]). 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:

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

where $\mathtt{up}(\cdot, \cdot)$ 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, 59, 85, 86]. 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 [54]. 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 VARs 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 [27, 72] 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. [55] definition of diversity—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 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 by drawing insights from the diffusion model literature. 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 in text-to-image diffusion models as shown in Fig. 2: (1) custom classifier-free guidance (CFG) schedules [23], and (2) condition-annealing by injecting noise into the condition embedding [55].

Option 1: CFG Scheduling. CFG [23] modulates the influence of the condition in both diffusion and VAR [21, 61]. Formally, CFG adjusts the model's prediction via

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

where $\mathbf{Q}_\mathbf{c}$ 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. [28] 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 have 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 [28]: $\omega(k) = \mathbb{1}[k \in \mathcal{K}]\omega_1 + \mathbb{1}[k \notin \mathcal{K}]\omega_K$ where $\mathcal{K} := \{k_{\text{min}}, \dots, k_{\text{max}}\}$, and ω_1 and ω_K are tunable hyperparameters.
- Interpolation [68]: $\omega(k) = (1-\gamma(k))\omega_1 + \gamma(k)\omega_K$ Here, $\gamma(1) = 0$ and $\gamma(\ell) = 1$ for all $\ell \geq k_{\text{max}}$, and we test several variants of the scheduling function $\gamma(\cdot)$ suggested by Wang et al. [68]. 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. [55] 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. [55], 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 found 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. For this, we first revisit the time-travel technique widely used for diffusion models to enhance image quality, noticing that the same approach is not directly applicable to VAR models. In turn, we propose a novel analogous technique specialized to VAR based on its *multi-scale* nature, thus effectively removing visual artifacts introduced by condition-annealing (Sec. 4.1) with minimal computation cost.

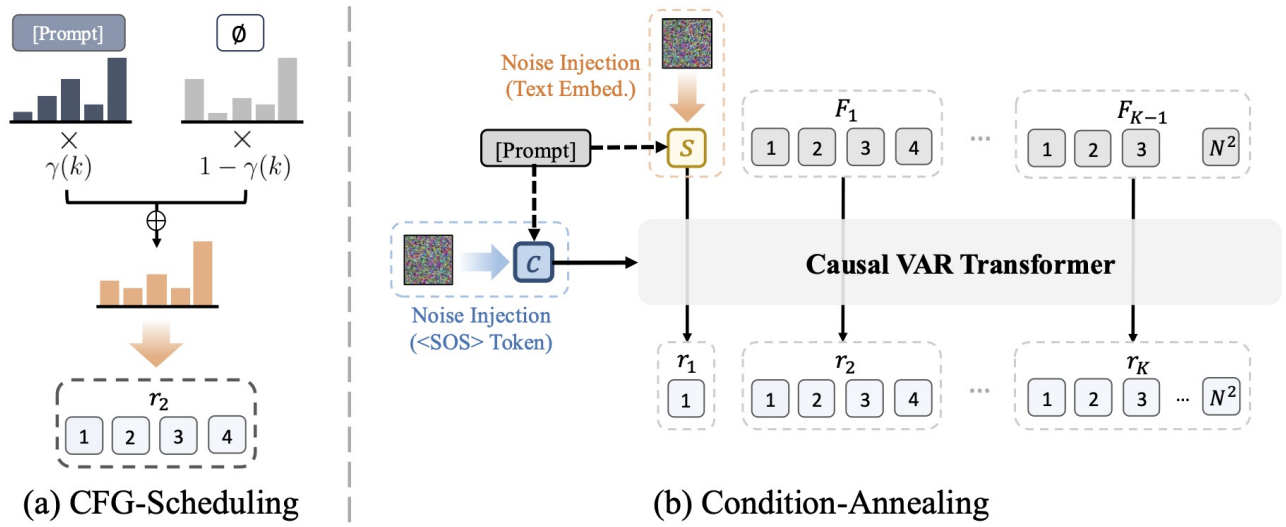

Figure 2. **Diversity-Enhancement Techniques.** We explore two main options for diversity enhancement in VAR: (a) CFG-scheduling and (b) condition-annealing. CFG-scheduling modulates the CFG scale over sampling steps to mitigate mode collapse. Condition-annealing injects noise into the text-embedding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

Time-Travel in Diffusion Models. To gain insights on image (or latent) refinement strategies, we revisit diffusion models, where *time-travel*—a noise-and-denoise step—has become a de-facto technique for refining intermediate or final outputs [4, 39, 42, 76, 80]. Consider a pretrained diffusion model π_θ and an intermediate *corrupted* latent $\mathbf{x}_t$ at timestep $t \in [1, T]$. Time-travel "rewinds time" by adding noise to send the latent to the previous timestep $t+1$, where a denoising step can then be run to refine the latent:

$$\texttt{Time-Travel:} \quad \tilde{\mathbf{x}}_{t+1} \sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_t, (1-\alpha_t)\mathbf{I}) \tag{4}$$

$$\texttt{Refine:} \quad \hat{\mathbf{x}}_t = \pi_\theta(\tilde{\mathbf{x}}_{t+1}, t+1) \tag{5}$$

where α_t follows a predefined variance schedule. This extra denoising pass is the key to time-travel: it projects the corrupted latent $\mathbf{x}_t$ back toward the model's learned image manifold, yielding $\hat{\mathbf{x}}_t$ with higher quality compared to $\mathbf{x}_t$ [42]. Although highly effective for diffusion models, this time-travel mechanism builds on the denoising scheme of diffusion models, making it inapplicable to VARs. Thus we raise the following question: ***Can we define a refinement technique for VARs, similar to time-travel in diffusion models?***

Scale-Travel in VARs. We present ***scale-travel***, a playback strategy for repairing corrupted latents in VAR. 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. Analogous to time-travel in diffusion models, scale-travel *rewinds* the feature map $\mathbf{Z}_l$ to an earlier (*i.e.*, coarser) scale $m < l$ and then resumes generation.

During training of VAR, an image I is first encoded with a VQ-VAE [64] to obtain a dense feature map $\mathbf{Z}$, which is passed through a *multi-scale encoder* to be decomposed into scale-wise token grids

$$\mathbf{r}_{1:K} = \texttt{MultiScaleEnc}(\mathbf{Z}, \mathbf{s}_{1:K}) \tag{6}$$

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 [26, 69], 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

$$\texttt{Scale-Travel:} \quad \tilde{\mathbf{r}}_{1:m} = \texttt{ScaleTravel}(\mathbf{Z}_l, \mathbf{s}_{1:m}, m) \tag{7}$$

$$\texttt{Refine:} \quad \hat{\mathbf{r}}_{m+1:K} \sim \prod_{k=m+1}^{K} p(\mathbf{r}_k | \tilde{\mathbf{r}}_1, \ldots, \tilde{\mathbf{r}}_m, \mathbf{r}_{m+1}, \ldots, \mathbf{r}_{k-1}) \tag{8}$$

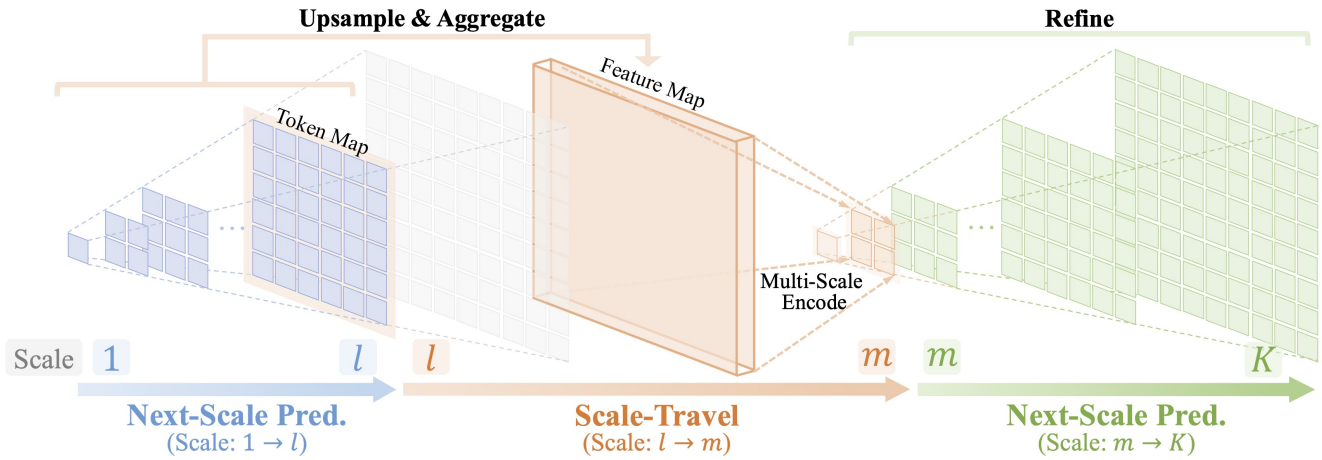

Figure 3. **Latent Refinement via Scale-Travel.** We introduce scale-travel, a novel refinement strategy that leverages the multi-scale structure of VAR models. By reverting intermediate representations to a coarser scale and running generation without noise injection, clean and coherent details can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

Algorithm 1 Multi-Scale Encoding

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}
```

Figure 4. **Scale-Travel Pseudocode.** Side-by-side pseudocode for **(left)** multi-scale encoding process in VAR [61] and **(right)** our proposed scale-travel step in Sec. 4.2. $\mathtt{up}(\cdot,\cdot)$ and $\mathtt{down}(\cdot,\cdot)$ denote bilinear upsampling and downsampling, respectively.

5. Results

In this section, we present the experiment results of our DIVERSEVAR, comparing both diversity and quality with existing diversity enhancement techniques on VARs.

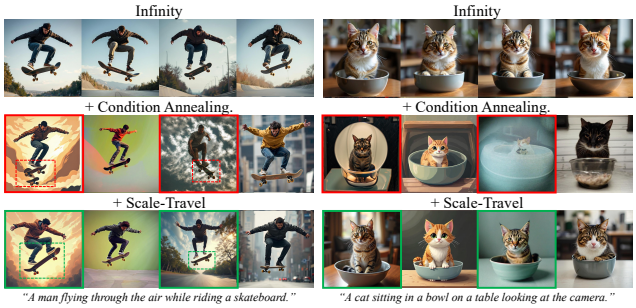

Figure 5. **Qualitative Comparisons.** Infinity produces images with little variation and diversity (row 1). Injecting noise into the text-embedding increases diversity (row 2) but results in visual artifacts (red box). Applying our **scale-travel** refinement technique (row 3) fixes these visual artifacts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

5.1. Evaluation Setup

VAR Models. For text-conditioned VAR models, we use Infinity [21] and Switti [66]. Infinity trains a bitwise transformer for scaling the vocabulary to infinite size, and Switti improves the efficiency and scalability of VAR [61] by solving training instability and using non-causal self-attention for faster sampling.

Benchmarks and Evaluation Metrics. We evaluate on MS-COCO [36] and MJHQ-30K [31] 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 [43]. We use multiple metrics to measure image quality and diversity as follows:

- Diversity: **LPIPS-MPD** [1] computes the mean pairwise LPIPS [84] distance between images to measure per-prompt diversity.
- Diversity: **Vendi-Score** [17] computes the effective num-

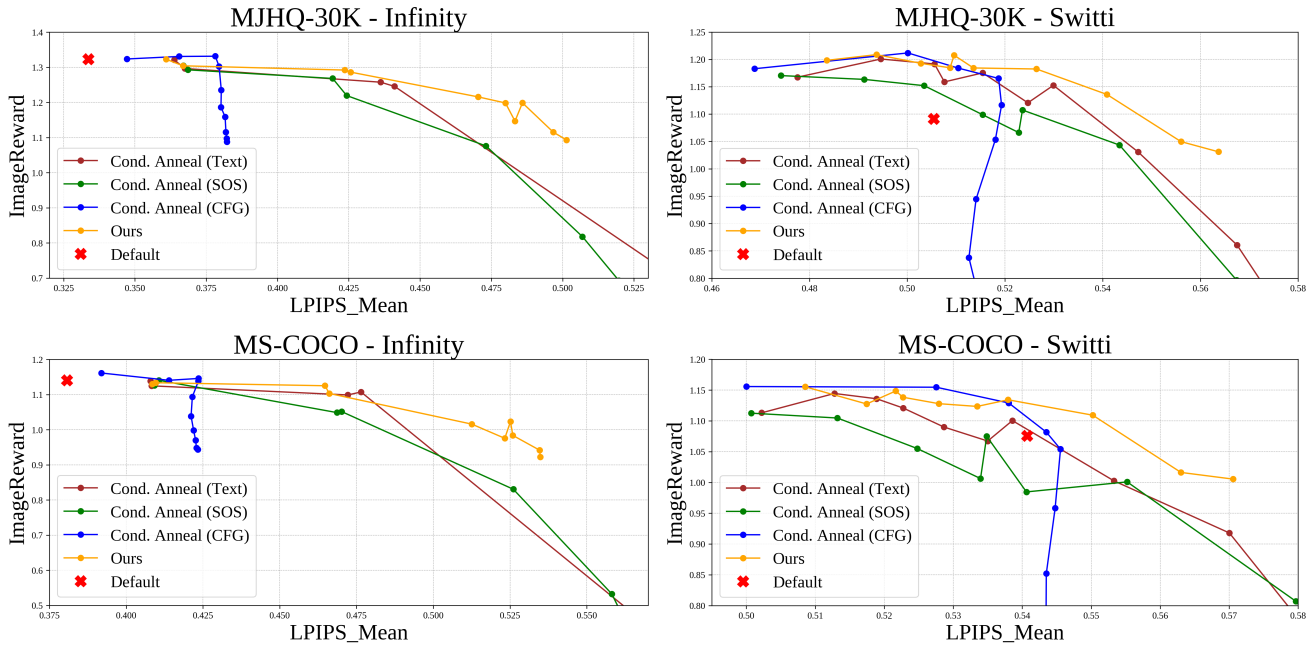

Figure 6. **Pareto Fronts for Diversity-Quality Trade-Off on Infinity [21] and Switti [66]**. Each curve is obtained by varying hyperparameters for each method: noise scales for SOS and CADS, guidance schedules for CFG, and target scales for DIVERSEVAR (Ours). The default setting of each model is included as a reference. On both MJHQ-30K [31] and MS-COCO [36], DIVERSEVAR consistently sets the Pareto frontier, demonstrating the best balance between diversity and quality in VAR.

ber of "dissimilar" elements in a sample of images. Following CADS [55], we use SSCD [49] to measure pairwise dissimilarities.

- Quality: **ImageReward** [74] evaluates image quality and alignment based on human preference.
- Quality & Diversity: **FID** [22] simultaneously captures both diversity and quality with respect to a reference set of images.

5.2. Pareto Fronts on Diversity-Quality Trade-Off

We assess DIVERSEVAR's diversity–quality balance 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 [31, 36], 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 $k_{\text{min}} = 1$ and vary $k_{\text{max}} \in \{2, 3, 4, 5, 6\}$ 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 $k_{\text{max}} \in 1, 2, \ldots, 10$ 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 are provided in the **supplementary**. We also report the specific hyperparameters used for each tested method.

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.

Method	MJHQ-30K [31]				MS-COCO [36]			
	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 [66]	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.** We evaluate four metrics: ImageReward (IR) [74] for quality, FID [22] for both quality and diversity, and LPIPS-MPD (MPD_L) [84], Vendi [17] for diversity. For both Switti [66] 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.

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 $k_{\text{max}} < K$, as done in [55].

As shown in Tab. 1, our method generalizes to different text-conditioned 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 on the LPIPS-MPD [1] for MJHQ-30K [31] 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 noised samples, achieving a $+120.41\%$ improvement on ImageReward [74], 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 the image quality to achieve a substantial improvement in diversity over standard text-conditioned VARs, with only a minor decrease in quality.

Moveover, in Fig. 5, we present qualitative results from each stage of our DIVERSEVAR pipeline. The outputs from the default model (row 1) exhibits a severe lack of diversity across various prompts. 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**.

Supplementary. Further details and extended results are provided in the supplementary material. We include empirical analyses of the multi-scale encoding strategy and the refinement stage. We report the computational cost of DIVERSEVAR, showing that DIVERSEVAR introduces minimal overhead in runtime. We also report additional metrics on text-alignment, comparison with prompt-rewriting, and results of varying the CFG weights, sampling temperature, and top-p sampling parameters. The supplementary also presents additional qualitative results and ablations of the Pareto curves. Moreover, we provide comparisons between different noise injection strategies, along with ablations on various noise injection schedulers.

6. Conclusion

In this work, we first report a significant lack of diversity in text-conditioned VAR models—an observation that has not been addressed in previous works. 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 (please refer to the **supplementary** for more details). We also observe that the

most diverse results still suffer from visual artifacts, occurring in approximately 6% of the generated samples.

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 6, 8

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021. 3

[3] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng YAN. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *ICLR*, 2025. 2

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 5

[5] Victor Besnier, Mickael Chen, David Hurych, Eduardo Valle, and Matthieu Cord. Halton scheduler for masked generative image transformer. In *ICLR*, 2025. 2

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 2, 3

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. 2023. 2

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020. 2

[10] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. 2018. 2

[11] Zigeng Chen, Xinyin Ma, Gongfan Fang, and Xinchao Wang. Collaborative decoding makes visual auto-regressive modeling efficient. *arXiv preprint arXiv:2411.17787*, 2024. 3

[12] Jiwoo Chung, Sangeek Hyun, Hyunjun Kim, Eunseo Koh, MinKyu Lee, and Jae-Pil Heo. Fine-tuning visual autoregressive models for subject-driven generation. *arXiv preprint arXiv:2504.02612*, 2025. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[16] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*, 2024. 2

[17] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 6, 8

[18] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *NeurIPS*, 2024. 3

[19] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 3

[20] Hang Guo, Yawei Li, Taolin Zhang, Jiangshan Wang, Tao Dai, Shu-Tao Xia, and Luca Benini. Fastvar: Linear visual autoregressive modeling via cached token pruning. *arXiv preprint arXiv:2503.23367*, 2025. 2, 3

[21] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025. 1, 2, 3, 4, 6, 7, 8, 10, 11

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 7, 8

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 3, 4

[24] Zhihao Huang, Xi Qiu, Yukuo Ma, Yifu Zhou, Chi Zhang, and Xuelong Li. Nfig: Autoregressive image generation with next-frequency prediction. *arXiv preprint arXiv:2503.07076*, 2025. 3

[25] Jiwan Hur, DongJae Lee, Gyojin Han, Jaehyun Choi, Yunho Jeon, and Junmo Kim. Unlocking the capabilities of masked generative models for image synthesis via self-guidance. In *NeurIPS*, 2024. 2

[26] Siyu Jiao, Gengwei Zhang, Yinlong Qian, Jiancheng Huang, Yao Zhao, Humphrey Shi, Lin Ma, Yunchao Wei, and Zequn Jie. Flexvar: Flexible visual autoregressive modeling without residual prediction. *arXiv preprint arXiv:2502.20313*, 2025. 2, 3, 5

[27] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-time scaling for flow models via stochastic generation and rollover budget forcing. *arXiv preprint arXiv:2503.19385*, 2025. 4

[28] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in

a limited interval improves sample and distribution quality in diffusion models. In *arXiv preprint arXiv:2404.07724*, 2024. 1, 2, 3, 4

[29] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 2, 3

[30] José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *ECCV*, 2022. 2

[31] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 6, 7, 8, 4, 10, 11, 12, 13

[32] Jiajun Li, Yue Ma, Xinyu Zhang, Qingyan Wei, Songhua Liu, and Linfeng Zhang. Skipvar: Accelerating visual autoregressive modeling via adaptive frequency-aware skipping. *arXiv preprint arXiv:2506.08908*, 2025. 2, 3

[33] Kunjun Li, Zigeng Chen, Cheng-Yen Yang, and Jenq-Neng Hwang. Memory-efficient visual autoregressive modeling with scale-aware kv cache compression. *Advances in Neural Information Processing Systems*, 39, 2025. 3

[34] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, 2024. 2

[35] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Zhe Lin, Rita Singh, and Bhiksha Raj. Controlvar: Exploring controllable visual autoregressive modeling. *arXiv preprint arXiv:2406.09750*, 2024. 2, 3

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6, 7, 8, 4, 9

[37] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 3

[38] Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14286–14295, 2020. 2

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2, 5

[40] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024. 2

[41] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 2

[42] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 5

[43] Midjourney Team. Midjourney. https://www.midjourney.com/, 2022. Accessed: 2025-05-16. 6

[44] Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. Csd-var: Content-style decomposition in visual autoregressive models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17013–17023, 2025. 3

[45] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 2

[46] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024. 2

[47] Jihun Park, Jongmin Gim, Kyoungmin Lee, Minseok Oh, Minwoo Choi, Jaeyeul Kim, Woo Chool Park, and Sunghoon Im. A training-free style-aligned image generation with scale-wise autoregressive model. *arXiv preprint arXiv:2504.06144*, 2025. 3

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3

[49] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 7

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. 2021. 2

[51] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019.

[52] Scott Reed, Aäron Oord, Nal Kalchbrenner, Sergio Gómez Colmenarejo, Ziyu Wang, Yutian Chen, Dan Belov, and Nando Freitas. Parallel multiscale autoregressive density estimation. 2017. 2

[53] Sucheng Ren, Yaodong Yu, Nataniel Ruiz, Feng Wang, Alan Yuille, and Cihang Xie. M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. *arXiv preprint arXiv:2411.10433*, 2024. 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[55] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *ICLR*, 2024. 1, 2, 3, 4, 7, 8

[56] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *ICLR*, 2017. 2

[57] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[58] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2, 3

[59] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. HART: Efficient visual generation with hybrid autoregressive transformer. In *ICLR*, 2025. 2, 3

[60] Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Accelerating autoregressive text-to-image generation with training-free speculative jacobi decoding. In *ICLR*, 2024. 3

[61] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*, 2024. 2, 3, 4, 6

[62] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 2

[63] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. 2016. 2

[64] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 3, 5

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[66] Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024. 3, 6, 7, 8, 1, 9, 12, 13

[67] Jiahuan Wang, Yuxin Chen, Jun Yu, Guangming Lu, and Wenjie Pei. Editinfinity: Image editing with binary-quantized generative models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[68] Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule CANI, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. 2024. 3, 4, 1, 2

[69] Yufei Wang, Lanqing Guo, Zhihao Li, Jiaxing Huang, Pichao Wang, Bihan Wen, and Jian Wang. Training-free text-guided image editing with visual autoregressive model. *arXiv preprint arXiv:2503.23897*, 2025. 3, 5

[70] Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. *arXiv preprint arXiv:2503.16430*, 2025. 2

[71] Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. 2025. 2

[72] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. 4

[73] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025. 2, 3

[74] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 7, 8

[75] Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024. 3

[76] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024. 5

[77] Zebin You, Jingyang Ou, Xiaolu Zhang, Jun Hu, Jun Zhou, and Chongxuan Li. Effective and efficient masked image generation models. *arXiv preprint arXiv:2503.07197*, 2025. 2

[78] Hu Yu, Hao Luo, Hangjie Yuan, Yu Rong, and Feng Zhao. Frequency autoregressive image generation with continuous tokens. *arXiv preprint arXiv:2503.05305*, 2025. 3

[79] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. 2022. 2

[80] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023. 5

[81] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 2

[82] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024. 3

[83] Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024. 2, 3

[84] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver 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. 6, 8

[85] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025. 2, 3

[86] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Dongchao Yang, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou.

Vargpt-v1. 1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning. *arXiv preprint arXiv:2504.02949*, 2025. 3

Supplementary

In this supplementary document, we first present visualizations and explanations of the VAR sampling trajectory and the multi-scale encoding employed for scale-travel (Sec. A). Next, we compare the various noise-scale schedulers for each diversity-enhancement strategy as introduced in Sec. 4.1 of the main paper (Sec. B), followed by results of applying scale-travel refinement for the different enhancement techniques (Sec. C). We then detail the hyperparameters used for the quantitative results in Tab. 1 of the main paper (Sec. D), report the inference time of DIVERSEVAR (Sec. E), and conclude with additional quantitative (Sec. F) and qualitative results of DIVERSEVAR (Sec. G).

A. Details on Multi-Scale Encoding

As mentioned, in Sec. 4.2 of the main paper, we provide additional details on the multi-scale encoding and the sampling trajectory. Fig. A.1 displays the decoded feature map $\mathbf{Z}_k$ at each scale k during VAR generation. Early scales form the coarse, high-level structure of the final image, while later scales form the fine, low-level details. Notably, by the eighth scale (green box), the global layout and semantic content have already emerged.

Moreover, Fig. A.2 demonstrates the multi-scale encoding outputs of different intermediate latents $\mathbf{Z}_l$ with varying $l \in [1, K]$. In **scale-travel**, we take the multi-scale encoding of an intermediate latent (Fig. A.2 cyan box), and run sampling initialized from the first few latents. The full scale-travel refinement strategy is visualized in Fig. A.3.

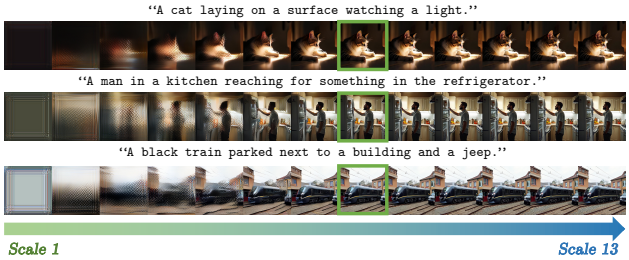

Figure A.1. **Sampling Trajectory of VAR.** Each row displays the decoded feature map at each scale k during VAR generation. The coarse high-level structure of the image is formed at earlier scales, while the fine low-level details are added at later scales.

B. Exploration for CFG Scheduling and Condition-Annealing Noise Injection Scheduling

In this section, we explore different schedulers for both CFG scheduling and noise injection techniques for diversity enhancement introduced in Sec. 4.1 of the main paper. Following previous works on CFG scheduling [28, 68], we provide details on popular schedulers used for each variant in the below Tab. B.1, and visualize each scheduler in Fig. B.3.

Fig. B.1 shows the results for *CFG-scheduling* and *condition-annealing* by varying the choice of scheduling functions $\alpha(k)$, $\beta(k)$, and $\gamma(k)$. We limit the maximum value of $\alpha(k), \beta(k)$ with σ. Cosine and linear schedulers exhibit similar trends and performance on the Pareto front. The constant scheduler yields higher diversity but significantly lower image quality. Based on these results, we adopt the constant scheduler for our main experiments.

C. Comparison on Scale-Travel with various Condition-Annealing Noise Injections

In Fig. C.1, we show Pareto fronts for the results of applying scale-travel refinement to the <SOS> condition-annealing technique. Applying scale-travel refinement to <SOS> noise injection also achieves better Pareto fronts compared to the original <SOS> noise injection respectively. This figure shows that using our scale-travel refinement can improve the Pareto front of any condition-annealing methods. The refinement hyperparameters used here are identical to those in the main Pareto plot in Fig. 6 in the main paper.

D. Details on Experiment Setup (Tab. 1 of the main paper)

In this section, we report the hyperparameter settings used for the main quantitative results in Tab. 1 in Sec. 5.3 of the main paper. For Infinity [21], we adopt the default setting with a CFG weight of 2.0 and $\tau = 1.0$ for pure generation. Additionally, we inject noise into the text embeddings using a cosine scheduler with $k_{\text{max}} = 4$ and $\sigma = 1.0$. For refinement, we apply scale-travel from scale 8 down to scale 3. For Switti [66], we set the CFG weight to 6.0 for pure generation. Noise is injected into the text embeddings using a cosine scheduler with $k_{\text{max}} = 6$ and $\sigma = 1.0$. For refinement, scale-travel is applied from scale 8 to scale 4.

E. Inference Time

Tab. E.1 we provide a comparison of the inference time for refinement hyperparameters used in Tab. 1 in Sec. 5.3 of the main paper. The measurements were conducted on an RTX 3090 GPU with 24GB of VRAM. For all settings in the table, we used a batch size of 1. The reported time costs are averaged over 100 samples. As shown, Infinity [21] takes 2.115 seconds per inference while scale-travel takes slightly longer at 2.638 seconds per inference. Similarly, Switti [66] achieves an inference time of 1.642 seconds, which increases modestly to 1.892 seconds when combined with scale-travel. Notably, this corresponds to only a 15% overhead, while yielding substantial improvements in the diversity–quality trade-off.

Figure A.2. **Multi-scale Encoding.** Multi-scale encoding applied to feature map $\mathbf{Z}_l$ for $l \in [1, K]$ (right-most column). The multi-scale encodings obtained from $\mathbf{Z}_9$ (cyan box) are semantically simliar to the multi-scale encodings obtained from $\mathbf{Z}_K$ (red box), allowing scale-travel to be performed.

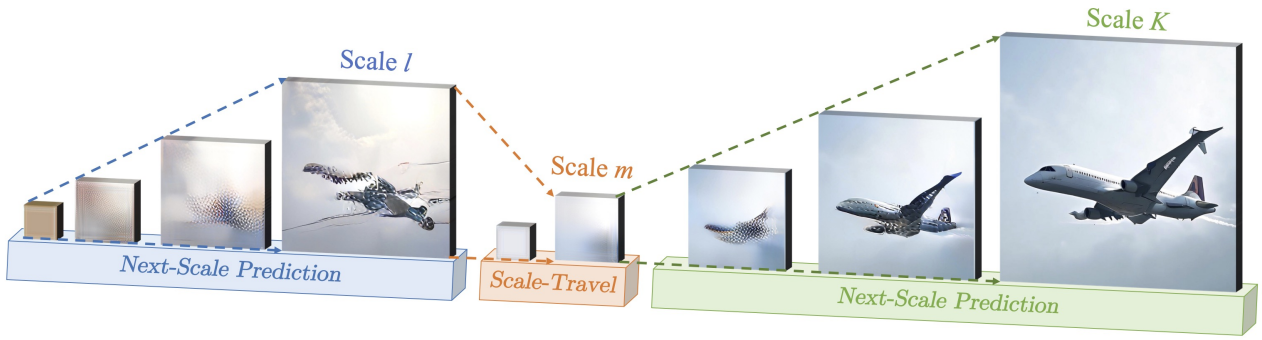

Figure A.3. **Scale-Travel Visualization.** Scale-travel is performed by applying multi-scale encoding to the feature map at stage l, and resuming generation at a coarser scale $m < l$.

Method	Type	Variant	Formula / Parameters
CFG Scheduling	Piecewise Constant [28]	Constant	$\omega_1 = 0, \omega_K = C$
		Constant (Inverse)	$\omega_1 = C, \omega_K = 0$
	Interpolation [68]	Cosine	$\gamma(k) = 1 - \frac{1}{2}\left(\cos(\frac{\pi(k-1)}{k_{\max}-1}) + 1\right)$
		Cosine (Inverse)	$\gamma(k) = \frac{1}{2}\left(\cos(\frac{\pi(k-1)}{k_{\max}-1}) + 1\right)$
		Linear	$\gamma(k) = \frac{k-1}{k_{\max}-1}$
		Linear (Inverse)	$\gamma(k) = 1 - \frac{k-1}{k_{\max}-1}$
		Constant	$\gamma(k) = 0.5$
Condition-Annealing	Text Embedding [55]	Linear	$\alpha(k) = \sigma(1 - \frac{k-1}{k_{\max}-1})$
		Cosine	$\alpha(k) = \frac{\sigma}{2}\left(\cos(\frac{\pi(k-1)}{k_{\max}-1}) + 1\right)$
		Constant	$\alpha(k) = \sigma$
	<SOS> Token	Linear	$\beta(k) = \sigma(1 - \frac{k-1}{k_{\max}-1})$
		Cosine	$\beta(k) = \frac{\sigma}{2}\left(\cos(\frac{\pi(k-1)}{k_{\max}-1}) + 1\right)$
		Constant	$\beta(k) = \sigma$

Table B.1. **Details on Diversity Enhancement Techniques.** Variants of CFG scheduling and condition-annealing strategies [28, 55, 68].

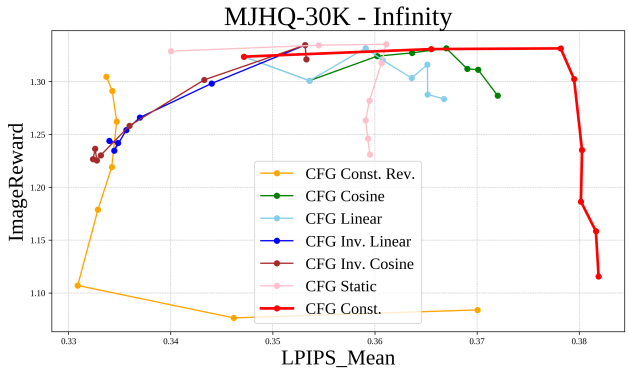

Figure B.1. **Pareto fronts with different CFG schedulers [28, 68].** The plot shows the result of different schedulers applied to CFG scheduling. We find that constant CFG scheduling yields the best results.

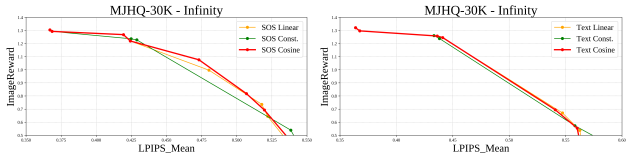

Figure B.2. **<SOS>and Text Embedding Annealing schedulers [55].** The plot shows the result of different scheduler for each of the condition-annealing methods (<SOS> noise injection and Text-embedding noise injection [55]). We find that cosine scheduling yields the best results.

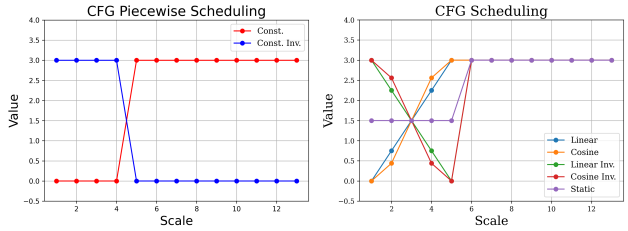

Figure B.3. **Visualization for various CFG schedulers with $k_{\max} = 5$ [28, 68].**

F. Additional Quantitative Results

In this section, we present additional ablation studies and experiments to further validate the effectiveness of our method. Tables F.1–F.4 report results on the MJHQ-30K prompt set used in our main Pareto analysis (see Sec. 5.2). Tables F.6 and F.7 use the same evaluation prompts as in Tab. 1.

F.1. DIVERSEVAR with Varying CFG weights

We provide additional results for different CFG guidance weights ω. As shown in Tab. F.1, DIVERSEVAR improves the diversity-quality tradeoff across all CFG weights.

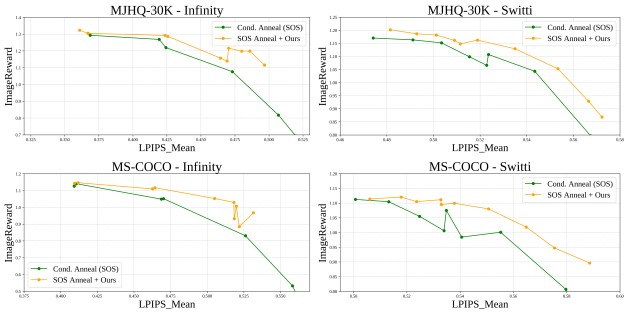

Figure C.1. **Scale-Travel Applied to <SOS> Condition Annealing.** DIVERSEVAR (SOS Anneal + Ours) consistently yields better Pareto frontiers than <SOS> Condition Annealing alone.

Method	Inference time	× Slower (vs. Default)
Infinity	2.115s	–
+ Cond. Anneal.	2.266s	×1.07
+ Scale-Travel	2.638s	×1.24
Switti	1.642s	–
+ Cond. Anneal.	1.692s	×1.03
+ Scale-Travel	1.892s	×1.15

Table E.1. **Inference Time Comparison.** Noise injection increases inference time because KV-caching cannot be applied during the injection process. In Scale-wise VAR, most of the computational cost is concentrated in the final stage, whereas DIVERSEVAR avoids repeated inferences at this stage. Consequently, the additional overhead introduced by scale-travel remains minimal, while still providing substantial improvements in image quality.

F.2. Scale-Travel Parameters for DIVERSEVAR

Scale-travel rewinds the autoregressive generation from stage l to stage m. In Tab. F.2, we vary $l \in \{6, 8, 10\}$ and $m \in \{2, 3, 4\}$. We find that $m = 3$ and $m = 4$ yields the best diversity-quality tradeoff for Infinity and Switti, respectively.

F.3. Temperature and Top-p sampling

Autoregressive models can control the diversity of generation by varying the temperature τ and through nucleus (top-p) sampling. As shown in Tab. F.3, while increasing the temperature τ also increases the diversity, DIVERSEVAR achieves the highest diversity. Importantly, increasing τ comes at the expense of significantly degraded image quality (τ=10 in Tab. F.3) whereas DIVERSEVAR increases diversity with only a minimal drop in image quality. Top-p sampling, on the other hand, does not significantly increase the diversity as shown in Tab. F.4.

F.4. Text-Alignment

We evaluate the text-alignment of DIVERSEVAR using GenEval [19] in Tab. F.5 and using VQAScore [37] in

ω	Method	MPD ↑	Vendi ↑	IR ↑	ω	Method	MPD ↑	Vendi ↑	IR ↑
0.0	Infinity	0.383	4.989	1.075	0.0	Switti	0.528	6.924	0.413
	DIVERSEVAR	0.479	6.262	0.865		DIVERSEVAR	0.591	8.056	-0.291
2.0	Infinity (Default)	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 (default)	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 F.1. **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.

Tab. F.6. While DIVERSEVAR results in slightly worse text-alignment, the decrease is not significant and can be attributed to the increased diversity: increasing the diversity leads to sampling further away from the modes of the distribution, which slightly lowers text-alignment. However, we emphasize that DIVERSEVAR still does not significantly degrade the text-alignment according to eiter metrics.

F.5. Prompt-Rewriting

Prompt-rewriting is a technique to generate higher-quality, more diverse images by using an LLM to rewrite the prompt. In Tab. F.7, we apply prompt-rewriting to VAR models and compare the diversity and quality with DIVERSEVAR. DIVERSEVAR achieves significantly higher diversity than prompt-rewriting while also retaining high image quality. Furthermore, prompt-rewriting can be additionally applied to DIVERSEVAR, resulting in improved image quality.

	MJHQ	MS-COCO
Infinity	0.67	0.869
DiverseVAR	0.663	0.855
Switti	0.682	0.875
DIVERSEVAR	0.671	0.866

Table F.6. **VQAScore of DIVERSEVAR.** DIVERSEVAR does not significantly degrade the VQAScore compared to the baseline VAR models.

G. Additional Qualitative Results

We provide additional qualitative results for MS-COCO [36] in Fig. G.3 and Fig. G.4 and for MJHQ-30k [31] in Fig. G.7 and Fig. G.8. Condition-annealing increases the diversity at the expense of image quality. Adding our Scale-Travel **(Ours)**. refinement technique strikes the balance between diversity and image quality.

Infinity	m = 2 MPD ↑	Vendi ↑	IR ↑	m = 3 MPD ↑	Vendi ↑	IR ↑	m = 4 MPD ↑	Vendi ↑	IR ↑
ℓ = 6	0.385	4.712	1.312	0.485	6.108	1.199	0.556	7.453	0.724
ℓ = 8	0.386	4.671	1.309	0.486	6.087	1.199	0.556	7.441	0.729
ℓ = 10	0.386	4.668	1.310	0.486	6.091	1.184	0.555	7.430	0.726
Switti	MPD ↑	Vendi ↑	IR ↑	MPD ↑	Vendi ↑	IR ↑	MPD ↑	Vendi ↑	IR ↑
ℓ = 6	0.503	5.737	1.172	0.542	6.307	1.115	0.564	6.695	0.992
ℓ = 8	0.502	5.694	1.161	0.542	6.283	1.121	0.564	6.609	1.031
ℓ = 10	0.502	5.697	1.179	0.538	6.226	1.135	0.563	6.602	1.004

Table F.2. **DIVERSEVAR with different scale-travel parameters.** Scale-travel rewinds the autoregressive generation process from stage l to stage m. We vary $l \in \{6, 8, 10\}$ and $m \in \{2, 3, 4\}$. We find that $\ell = 8$ and $m = 3$ yields the best diversity-quality tradeoff for Infinity and $\ell = 8$ and $m = 4$ yields the best diversity-quality tradeoff for Switti.

Infinity	MPD ↑	Vendi ↑	IR ↑	Switti	MPD ↑	Vendi ↑	IR ↑
$\tau = 0.5$	0.284	3.500	1.305	$\tau = 0.5$	0.437	4.483	1.204
$\tau = 1$ (default)	0.334	4.031	1.323	$\tau = 1$ (default)	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 F.3. **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	MPD ↑	Vendi ↑	IR ↑	Switti	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	$p = 0.95$ (default)	0.444	4.744	1.210
$p = 0.97$ (Default)	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 F.4. **VAR models with varying top-p sampling parameter.** Varying p does not significantly impact the diversity. DIVERSEVAR achieves significantly higher diversity without significantly degrading the image quality.

	Overall	Single Object	Two Object	Counting	Colors	Position	Color Attribution
Infinity	0.686	0.997	0.785	0.622	0.854	0.258	0.603
DIVERSEVAR	0.662	0.981	0.742	0.622	0.835	0.263	0.528
Switti	0.615	0.997	0.773	0.544	0.894	0.118	0.368
DIVERSEVAR	0.586	0.978	0.722	0.494	0.859	0.128	0.333

Table F.5. **GenEval score of DIVERSEVAR.** DIVERSEVAR does not significantly degrade text-alignment.

	MJHQ IR ↑	FID ↓	MPD ↑	Vendi ↑	MS-COCO IR ↑	FID ↓	MPD ↑	Vendi ↑
Infinity +Prompt-Rewriting	1.251	17.162	0.387	4.741	1.124	31.575	0.457	5.541
DiverseVAR	1.085	15.278	0.485	6.071	1.04	28.974	0.516	6.682
Switti +Prompt-Rewriting	1.086	14.639	0.484	5.552	0.947	27.830	0.524	6.457
DiverseVAR	0.93	17.30	0.56	6.68	1.05	23.47	0.57	7.22

Table F.7. **Prompt-rewriting in VAR models.** We apply prompt-rewriting to VAR models and compare the diversity-quality tradeoff against DIVERSEVAR. While prompt-rewriting improves the diversity, DIVERSEVAR achieves higher diversity. Moreover, prompt-rewriting can be applied to DIVERSEVAR to further improve the image quality while still generating diverse outputs.

Figure G.1. **Additional Qualitative Results on MS-COCO [36] using Infinity [21].** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

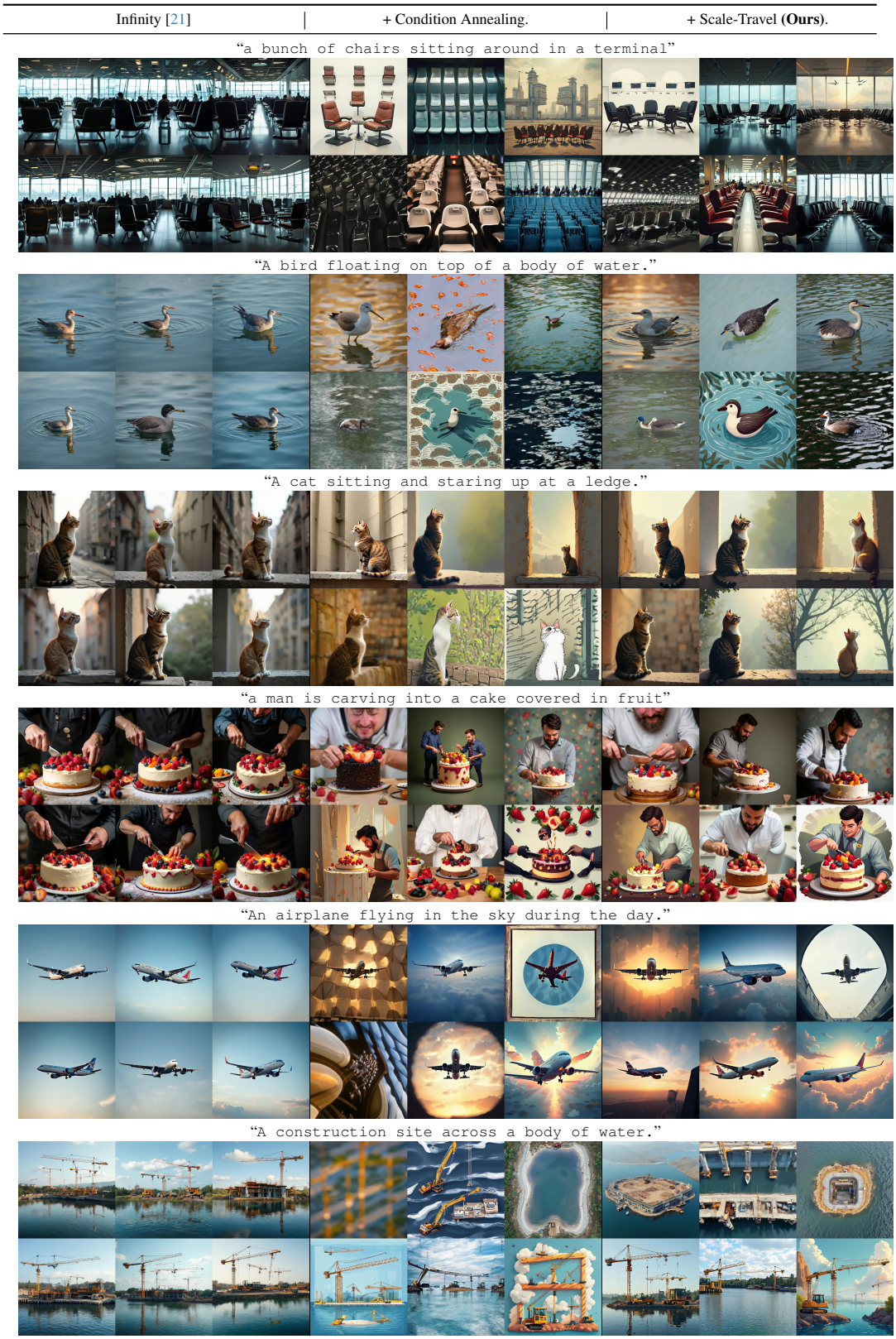

Figure G.2. **Additional Qualitative Results on MS-COCO [36] using Infinity [21] (Cont'd).** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

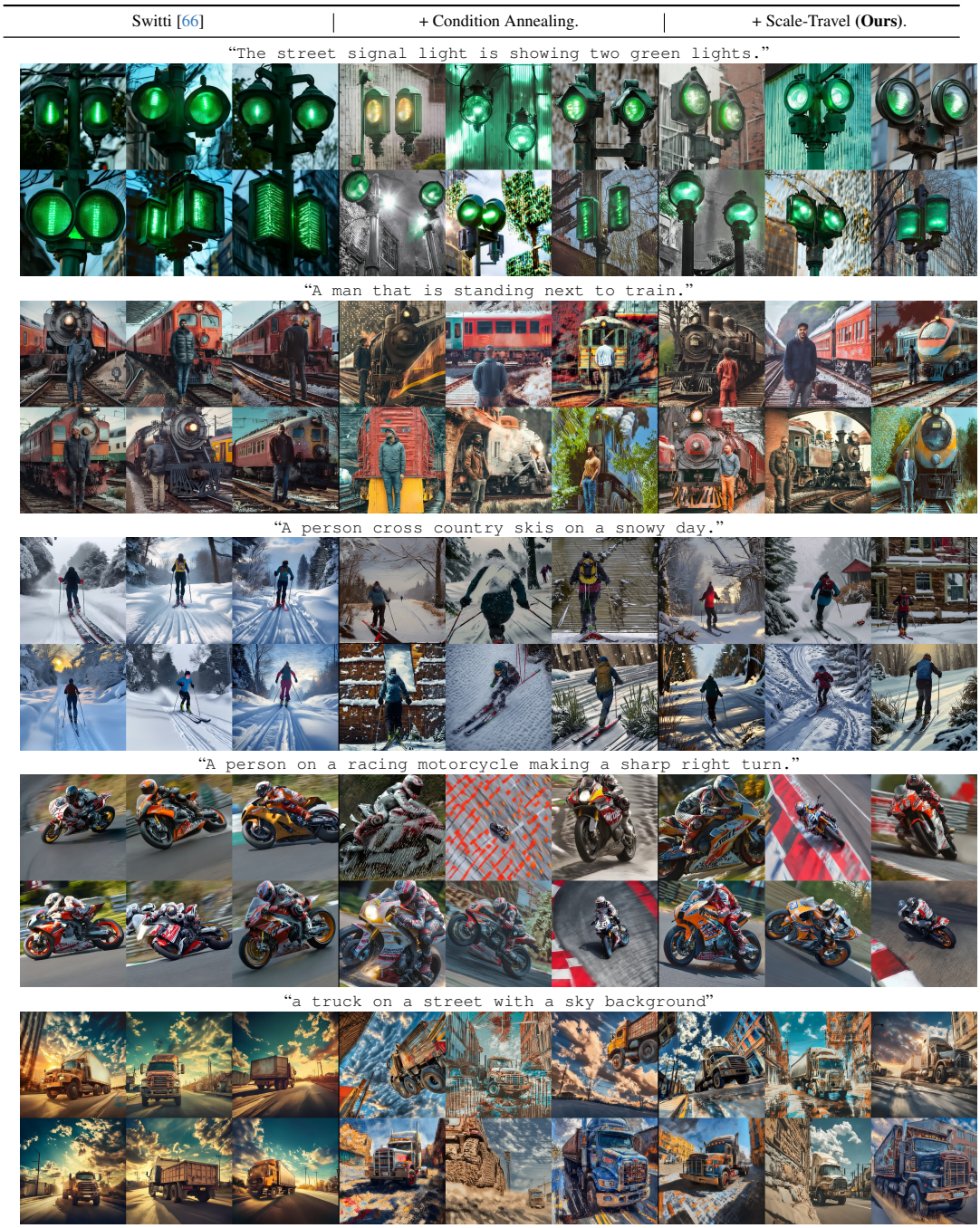

Figure G.3. **Additional Qualitative Results on MS-COCO [36] using Switti [66].** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

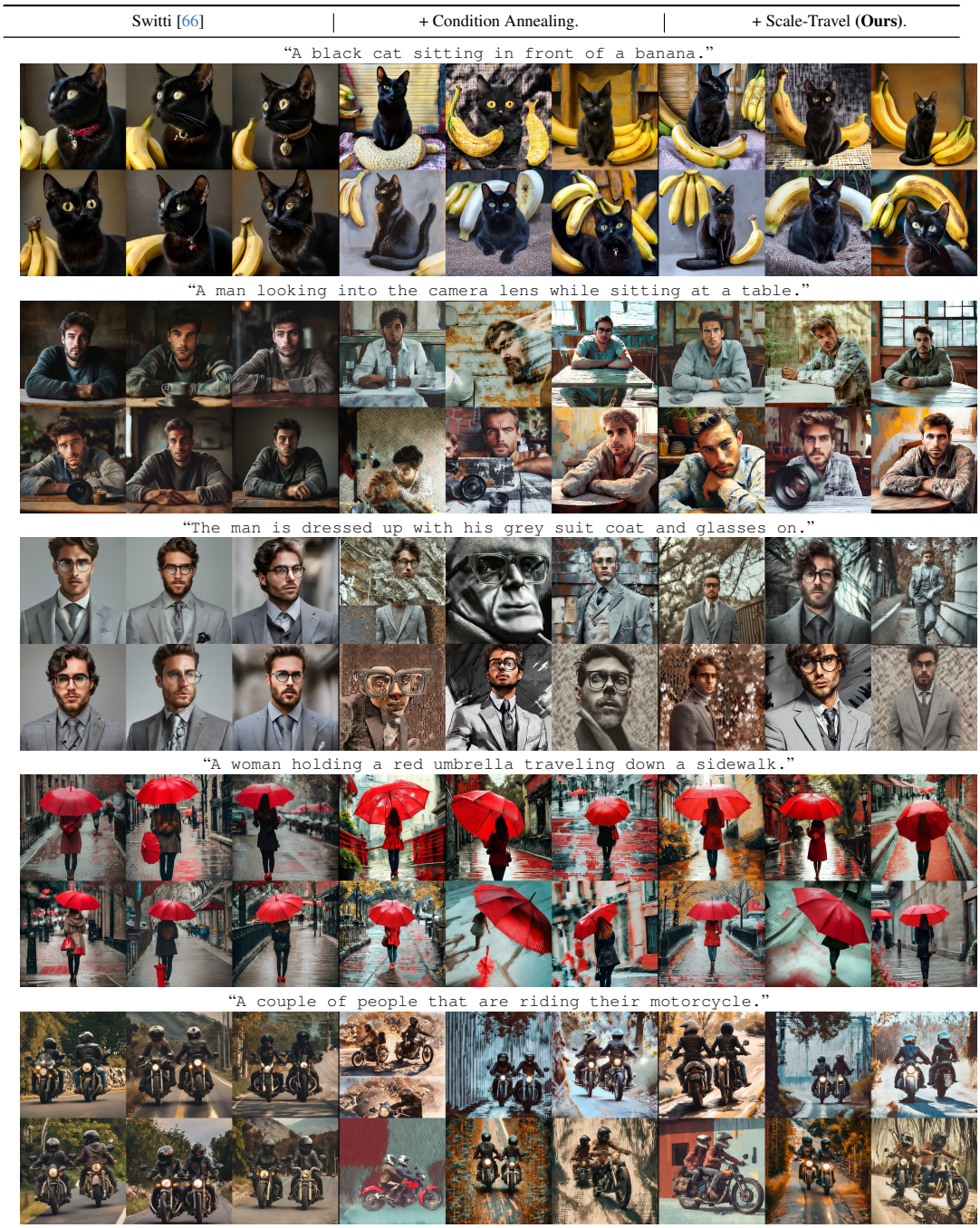

Figure G.4. **Additional Qualitative Results on MS-COCO [36] using Switti [66] (Cont'd).** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

Figure G.5. **Additional Qualitative Results. MJHQ-30K [31] using Infinity [21].** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

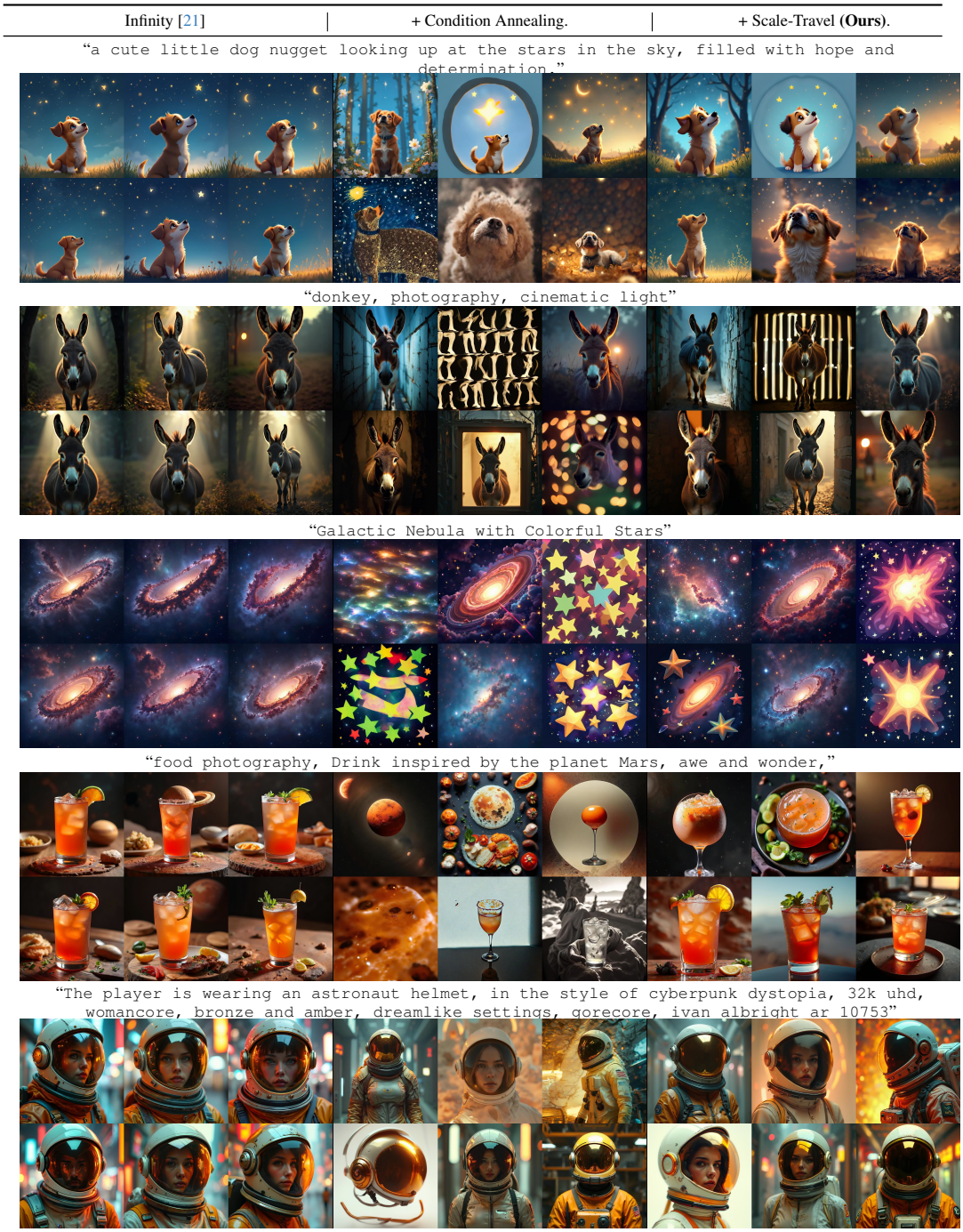

Figure G.6. **Additional Qualitative Results. MJHQ-30K [31] using Infinity [21] (Cont'd).** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

Figure G.7. **Additional Qualitative Results. MJHQ-30K [31] using Switti [66].** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.

Switti [66]	+ Condition Annealing.	+ Scale-Travel (**Ours**).

"A futuristic bottle of champagne , exensive luxury itm, green bottle with label, scifi gadet"

"IMAGE_TYPE Landscape photography GENRE Nature EMOTION Dreamy SCENE An isolated lighthouse on a rocky coastline, with the Southern Lights and a starfilled sky as the backdrop ACTORS None LOCATION TYPE Coastline CAMERA MODEL Fujifilm GFX 100 CAMERA LENSE 23mm f4 SPECIAL EFFECTS Long exposure, dramatic lighting TAGS Auroras, starry skies, nighttime landscapes, Southern Lights, Milky Way, dreamy, lighthouse, rocky coastline"

"road leading into a silhouette city, background sunset warm twlight gradient sky"

"This is an image that glows. , A lotusflower, Flying, Dreamy colors, Colorful light spots, Many beautiful spectra, Magic particle light, Bright, Super clear, Perfect detail, Transparent, Focused light, Optics, Surreal, Magical fantasy, Epic, Perfect composition, 8k,"

"Hyper Realistic Picture of Linda Carter as Wonder Womanr taken with a Canon EOS R6, Fullframe camera with a 50mm f1.8 lens, ISO 100, Aperture Priority Av mode, Autofocus, single point AF, RAW, Shooting in a welllit environment with soft, diffused lighting, subject positioned offcenter and slightly angled toward the camera"

Figure G.8. **Additional Qualitative Results. MJHQ-30K [31] using Switti [66] (Cont'd).** Condition-annealing improves the diversity at the cost of visual artifacts and lower image quality. Applying Scale-Travel **(Ours)** maintains a similar degree of diversity while correcting the visual artifacts.