HIDDEN IN PLAIN SIGHT: RECOVERING LATENT GENERATIVE DIVERSITY IN VISUAL AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

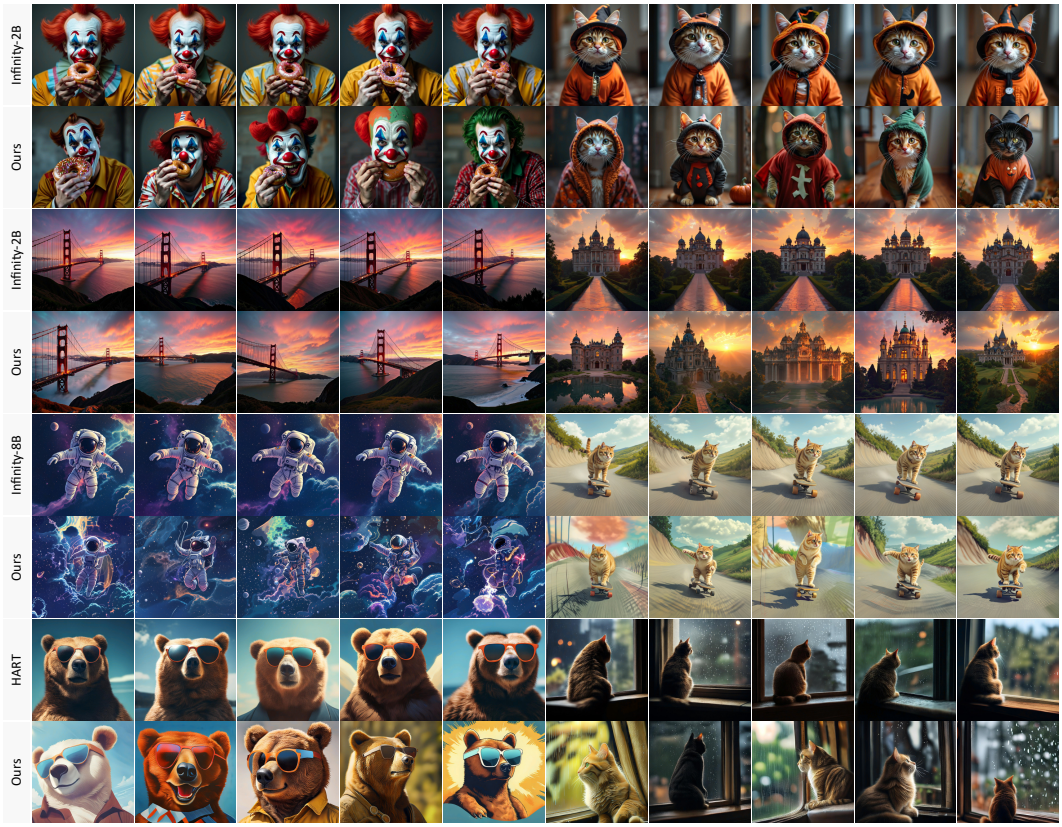

Figure 1: Multiple generation samples from the vanilla VAR models (*1st*, *3rd*, *5th*, and *7th* rows) and our *DiverseVAR* (*2nd*, *4th*, *6th*, and *8th* rows). While vanilla VAR models suffer from diversity collapse, our method produces more diverse outputs while maintaining strong image–text alignment. The text prompts used are as follows: "A man in a clown mask eating a donut", "A cat wearing a Halloween costume", "Golden Gate Bridge at sunset, glowing sky, ...", "A palace under the sunset", "A cool astronaut floating in space", "A cat riding a skateboard down a hill", "A bear wearing sunglasses on a sunny day", and "A cat sitting on a window ledge watching rain".

ABSTRACT

Visual Autoregressive (VAR) models have recently garnered significant attention for their innovative next-scale prediction paradigm, offering notable advantages in both inference efficiency and image quality compared to traditional multi-step autoregressive (AR) and diffusion models. However, despite their efficiency, VAR models often suffer from the diversity collapse i.e., a reduction in output variability, analogous to that observed in few-step distilled diffusion models. In this paper, we introduce **DiverseVAR**, a simple yet effective approach that restores the generative diversity of VAR models without requiring any additional training. Our analysis reveals the pivotal component of the feature map as a key factor governing diversity formation at early scales. By suppressing the pivotal component in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the model input and amplifying it in the model output, **DiverseVAR** effectively unlocks the inherent generative potential of VAR models while preserving high-fidelity synthesis. Empirical results demonstrate that our approach substantially enhances generative diversity with only neglectable performance influences. Our code will be publicly released.

1 INTRODUCTION

Autoregressive (AR) models (Lee et al., 2022; Razavi et al., 2019; Sun et al., 2024; Yu et al., 2022) have emerged as a powerful next-token prediction paradigm (Sun et al., 2024; Wang et al., 2025), achieving competitive results in visual generation but suffering from substantial generation latency caused by their numerous sequential decoding steps. To alleviate the latency, pioneering studies investigated various representation patterns (Huang et al., 2025; Pang et al., 2024; Jang et al., 2024) and parallel decoding (Chang et al., 2022; Teng et al., 2024). Recently, unlike conventional next-token prediction in AR models, visual autoregressive (VAR) models adopt a next-scale prediction paradigm (Tian et al., 2024; Han et al., 2024), achieving efficient and high-quality image generation within approximately ten scale steps. However, although VAR models perform inference with far fewer steps than traditional AR models, they suffer from the *diversity collapse* problem under numerous scenarios (Fig. 1, 1st, 3rd and 5th rows).

In this work, we address the *diversity collapse* problem in VAR models by unlocking their inherent generative diversity. Through a detailed analysis of the coarse-to-fine next-scale prediction process in pretrained VAR models, we arrive at the following key observations. ***(a) Structural Formation in Early Scales:*** As shown in Figs. 2 and 3, structural formation predominantly occurs at early scales, while later scales mainly refine and stabilize existing structures. This finding directs our attention toward the early-stage prediction dynamics. ***(b) Diversity Governed by the Pivotal Token:*** As demonstrated in Figs. 4 and 5, the pivotal token dominates the formation of structure while the auxiliary token carries both semantic information and image fidelity.

Based on the these insights, we propose **DiverseVAR**, a training-free framework that effectively unleashes the inherent generative diversity of VAR models. Without any additional training, DiverseVAR strategically intervenes during the inference process, guided by our analytical findings. Our approach consists of two complementary sampling steps: ***(a) Soft-Suppression Regularization:*** We identify the pivotal component of the model input that governs diversity formation and apply a soft-suppression strategy to attenuate components contributing to structural redundancy, thereby mitigating diversity collapse. ***(b) Soft-Amplification Regularization:*** To further promote controlled diversity, we enhance the pivotal component of the model output through a soft-amplification mechanism, ensuring that the expanded diversity remains consistent text–image alignments. Together, these two regularizations enable *DiverseVAR* to unleash the inherent potential of VAR models while preserving high image fidelity and faithful semantic alignment. Our contributions are as follows:

- We conduct a systematic analysis to identify the scale factors that govern the *diversity collapse* problem in VAR models. Our findings reveal that the VAR model diversity is predominantly influenced by the pivotal component at early scales.
- Based on our observations, we introduce **DiverseVAR**, a simple yet effective training-free method to enable the emergence of the inherent generative diversity of VAR models. It primarily incorporates two regularization terms during VAR inference: soft-suppression regularization and soft-amplification regularization.
- Experiments on the COCO, GenEval, and DPG benchmarks show that our method **DiverseVAR** unleashes the VAR model diversity while maintaining the sampled image fidelity, with slight influences on the generation performance according to numerous evaluation metrics.

2 RELATED WROK

2.1 VISUAL AUTOREGRESSIVE GENERATION

Traditional autoregressive (AR) methods (Lee et al., 2022; Sun et al., 2024; Yu et al., 2022) follow the next-token prediction paradigm, which requires a large number of iterative steps to generate high-quality images. Recently, Visual Autoregressive (VAR) modeling (Tian et al., 2024) has

adopted a next-scale prediction paradigm, enabling progressive generation across different resolutions to produce high-quality images. In contrast to AR methods that require massive steps to produce a high-resolution image (Lee et al., 2022; Sun et al., 2024; Yu et al., 2022), VAR can achieve comparable quality within only around ten steps (Tian et al., 2024; Han et al., 2024; Ma et al., 2024; Tang et al., 2024). Despite the promising performance of VAR models, current VAR models remain exhibit a notable gap: their generated samples for a given text prompt tend to be highly similar. Such indicated limited diversity is termed as *diversity collapse* problem in this paper.

2.2 Alternative Visual Generation Paradigms

Diffusion models (Rombach et al., 2022; Podell et al., 2023) trained on large-scale text-image datasets (Schuhmann et al., 2022) are able to generate high-quality and diverse images, but they face a time-consuming challenge due to multi-step sampling. To address this, distillation techniques (Luo et al., 2023; Sauer et al., 2023b; Lin et al., 2024; Dao et al., 2024) have been applied to diffusion models, enabling few-step sampling in distilled student models. Despite the improved sampling speed, the resulting images often exhibit *diversity collapse* (Kang et al., 2024). The exploration of diverse generation has recently attracted increasing attention in the context of few-step diffusion models. For example, Diffusion2GAN (Kang et al., 2024) aligns its generation trajectory with that of the teacher diffusion model, effectively reducing diversity collapse. Loopfree (Li et al., 2025b) employs a 4-step parallel decoder in the UNet architecture combined with Variational Score Distillation (VSD) (Wang et al., 2023) to generate images that exhibit both high diversity and fidelity. Hybrid (Gandikota & Bau, 2025) uses the base model only for the first step to seed diversity, and then switches to the distilled model for subsequent generation. C3 (Han et al., 2025) enhances generative diversity by amplifying internal feature representations using automatically selected amplification factors. Despite improving diversity, the requirement for training remains a bottleneck (Kang et al., 2024; Li et al., 2025b), the simultaneous use of both teacher and student models further increases memory consumption (Gandikota & Bau, 2025), and depend on creativity-oriented prompts to stimulate diverse outputs (Han et al., 2025). By contrast, **AR models** often demonstrate strong generative diversity (Xiong et al., 2024). Recent researches (Ma et al., 2025; Yu et al., 2024) focus on enhancing the balance between content quality and diversity, rather than specifically addressing diversity in AR models. In contrast to conventional AR models, *VAR* performs next-scale prediction to enable efficient generation, but suffers from diversity collapse. Nevertheless, prior efforts addressing diversity in diffusion and AR models are not directly transferable to VAR, and exploration of diversity in VAR remains scarce.

In this work, we conduct a systematic analysis to identify which scales govern diversity in VAR models. Building on these observations, we further characterize the properties of these scales that are closely associated with diversity. Finally, leveraging these properties, we propose a diversity enhancement for VAR inference that preserves both semantic consistency and visual fidelity.

3 Method

VAR revisits the conventional next-token prediction paradigm in AR image generation and introduces a coarse-to-fine next-scale prediction framework (Sec. 3.1), which substantially accelerates sampling while generating high-quality images, yet suffers from *diversity collapse*. In this section, we explore how the mechanisms underlying the coarse-to-fine generation process affect diversity (Sec. 3.2). Building on these insights, we propose *DiverseVAR* that fosters the emergence of diversity, enabling VAR models to achieve higher diversity while maintaining fidelity (Sec. 3.3).

3.1 Preliminaries

Visual Autoregressive (VAR) modeling (Tian et al., 2024) reformulates the conventional autoregressive (AR) framework (Lee et al., 2022; Razavi et al., 2019; Sun et al., 2024; Yu et al., 2022) to the visual domain by transitioning from next-token to next-scale prediction. Under this paradigm, each autoregressive step produces a token map corresponding to a specific resolution scale, instead of predicting tokens one by one. Given an image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, a continuous image feature map $\boldsymbol{F} \in \mathbb{R}^{h\times w\times d}$ is obtained through a visual tokenizer. VAR then quantizes this feature map into K multi-scale token maps $\boldsymbol{R} = (\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_K)$, where each token map corresponds to a predefined spatial resolution (h_k, w_k) for $k = 1, \ldots, K$. Here, (h_K, w_K) denotes the final scale, which

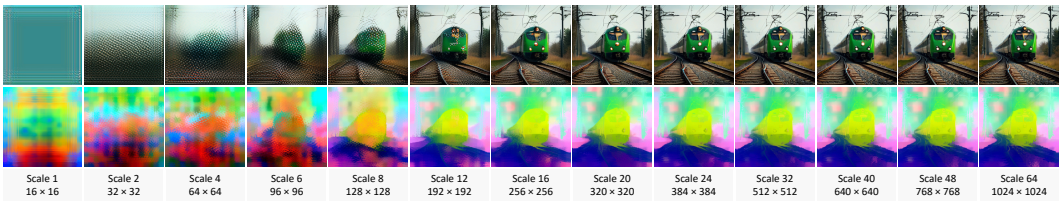

Figure 2: Visualization of samples across all scales (1st row) and their DINO features (2nd row).

is identical to the spatial size (h, w) of the feature map $\boldsymbol{F}$. This sequence of residuals progressively approximates the continuous feature map $\boldsymbol{F}$ as:

$$\boldsymbol{F}_k = \sum_{i=1}^{k} \mathrm{up}(\boldsymbol{R}_i, (h, w)), \tag{1}$$

where $\mathrm{up}(\cdot)$ denotes the upsampling operation. The transformer autoregressively predicts the next-scale $\boldsymbol{R}$ conditioned on the previously generated residuals, with the overall likelihood as:

$$p(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_K) = \prod_{k=1}^{K} p(\boldsymbol{R}_k \mid \boldsymbol{R}_1, \ldots, \boldsymbol{R}_{k-1}). \tag{2}$$

The transformer block in VAR takes the text embeddings at the first scale as input to predict $\boldsymbol{R}_1$. In the subsequent k-th scale, the feature map from the previous $(k$-1$)$-th scale is downsampled to match the spatial resolution of the input at the k-th scale:

$$\widetilde{\boldsymbol{F}}_{k-1} = \mathrm{down}(\boldsymbol{F}_{k-1}, (h_k, w_k)), \tag{3}$$

where $\mathrm{down}(\cdot)$ denotes the downsampling operation, and the transformer block receives the downsampled feature $\widetilde{\boldsymbol{F}}_{k-1}$ as input to predict $\boldsymbol{R}_k$.

3.2 Motivation

In this subsection, we present the motivation for our proposed method, based on two key observations[1] derived from vanilla text-to-image generative models (Han et al., 2024). First, the structural formation in vanilla next-scale models occurs at early scales, motivating us to explore the mechanisms that influence model diversity during these early scales (Figs. 2 and 3). Second, model diversity is primarily influenced by the pivotal token at early scales, presenting an opportunity to further promote the emergence of diversity while preserving generation fidelity (Figs. 4 and 5).

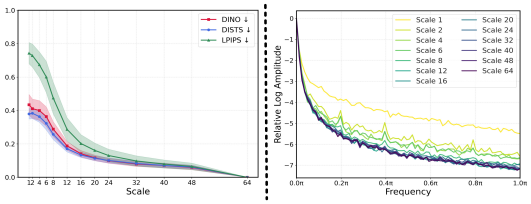

Figure 3: (Left) Statistics of structure evolution on all scale steps. (Right) The relative log amplitude of frequency components across different scales.

Observation 1: Structural Formation in Early Scales. We experimentally observe that the structural formation emerges at early scales, whereas later scales exhibit already stabilized structures (Figs. 2 and 3). More specifically, given a pretrained VAR model, it progressively produces intermediate predictions $\boldsymbol{R}_k$ and constructs the corresponding feature maps $\boldsymbol{F}_k$ (Eq. (1)) as the scale gradually increases (Fig. 2 (1st row)).

Motivated by (Tumanyan et al., 2022; 2023; Park et al., 2024), we adopt DINO features (Oquab et al., 2023) as a representation of the structural characteristics of the feature maps (Fig. 2 (2nd row)). As illustrated in Fig. 2 (2nd row), the structural formation of the image persists up to a specific scale (i.e., 12), after which the structure stabilizes.

Here, we further utilize DINO structure distance (Tumanyan et al., 2022; 2023) to measure whether the feature maps exhibit a well-formed overall structure comparable to that of the final scale (Fig. 3). As illustrated in Fig. 3 (Left), the curve quickly fall below 0.2 at early scales (i.e., 12), indicating

[1]The observations are drawn by statistical experiments on generated images, using 100 random text prompts from the COCO dataset (Lin et al., 2014).

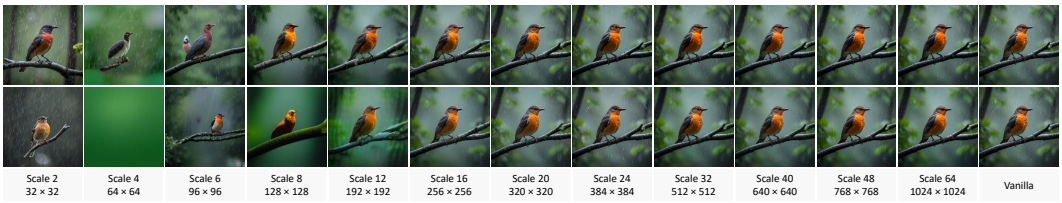

Figure 4: Visualization of samples when zeroing out the pivotal (1st row) or auxiliary (2nd row) tokens across all scales except the 1st scale (1st–12th columns), with the vanilla results (last column).

ongoing structural formation. The curve then converges at the remaining scales, implying that the structure has already stabilized. We also employ LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) to evaluate structural formation. Fig. 3 (Left) shows that the curves of both LPIPS and DISTS exhibit a similar trend to that of the DINO structure distance.

Furthermore, we perform frequency-domain analysis of the DINO features to evaluate the progression of structural formation (Fig. 3 (Right)). The frequency components decrease sharply at the early scales and gradually converge at the later scales, indicating that they become more stable and consistent as the scale increases. In conclusion, the early scales can be exploited to modify the structure, thereby promoting the emergence of diversity.

Observation 2: Diversity Governed by Pivotal Token. The above observation and analysis indicate that structural formation primarily occurs at early scales. This finding motivates us to further examine the components that influence this formation.

Thus, we investigate the internal composition of the feature map $\widetilde{\boldsymbol{F}}_{k-1}$ at scale k (Eq. (3)) by dividing it into *pivotal* and *auxiliary* components. Specifically, we follow (Guo et al., 2025) and define the pivotal score $s_{k,i} = \|\widetilde{\boldsymbol{F}}_{k-1,i} - \bar{\boldsymbol{F}}_{k-1}\|_2$ using the L2 norm, where $\bar{\boldsymbol{F}}_{k-1}$ represents the mean feature map obtained by averaging $\widetilde{\boldsymbol{F}}_{k-1}$ across scale dimensions, and $\widetilde{\boldsymbol{F}}_{k-1,i}$ denotes each token along the scale dimension. Based on their pivotal scores, all tokens in the feature map are ranked and then categorized into *pivotal* and *auxiliary* tokens, with the partition determined by the elbow point computed using the Maximum Distance to Chord (MDC) method (Douglas & Peucker, 1973) (See *Suppl.* for details). We simply regard the *pivotal* and *auxiliary* tokens as the *pivotal* and *auxiliary* components, and refer to this straightforward partition approach as *Naive Component Partition* (NCP).

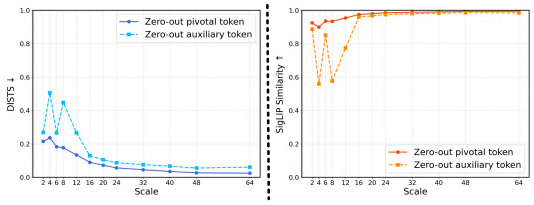

Figure 5: Structural (Left) and semantic (Right) evaluation when pivotal and auxiliary tokens are zeroed out.

We observe that the pivotal token primarily contributes structural formation, whereas the auxiliary token carries both semantic information and image fidelity. For example, when provided with the text prompt "A bird perched on a tree branch in the rain", VAR generates an image consistent with the given description (Fig. 4 (last column)). When zeroing out the pivotal token at one of the early scales (e.g., scale 4 within scales 2–8), the generated image exhibits noticeable structural changes, while its semantic meaning remains consistent (Fig. 4 (1st row, 1st-4th columns). In contrast, zeroing out the auxiliary token at one of these early scales severely degrades both semantic information and image fidelity (Fig. 4 (2nd row, 1st-4th columns).

Furthermore, we employ DISTS (Ding et al., 2020) and SigLIP (Tschannen et al., 2025) to quantitatively evaluate the modified images in comparison with the vanilla VAR outputs. As shown in Fig. 5, DISTS and SigLIP display consistent trends at early scales: zeroing out pivotal token (solid lines) results in slight degradation, while zeroing out auxiliary token (dashed lines) results in a drastic deterioration in both metrics. For example, when zeroing out the pivotal token, the variances of DISTS and SigLIP are below 0.3 and 0.1, respectively, whereas the corresponding values for the auxiliary token exceed 0.5 and 0.4. Finally, when zeroing out either the pivotal or auxiliary token at one of the later scales (e.g., scales 16-64), both the structural and semantic characteristics remain consistent with those of the vanilla generation (Fig. 4 (6th-12th columns) and Fig. 5). These results indicate that the model's diversity is predominantly governed by the pivotal token at early scales.

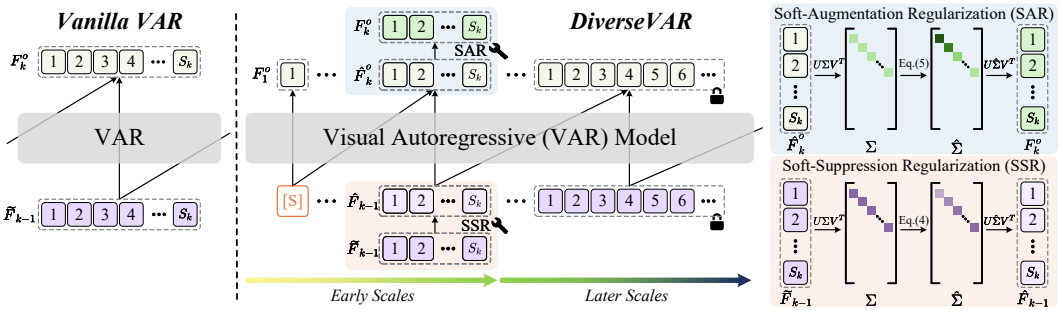

Figure 6: The overall framework of **DiverseVAR**. We explore the diversity emergence of VAR models at the early scales, while retaining the original VAR inference process at the later scales.

However, using the *Naive Component Partition* (NCP) to identify the pivotal component and apply its suppression (i.e., zeroing out) can promote the emergence of generative diversity, yet this naive approach may also induce unexpected variations (Fig. 4, 1st row, 2nd and 3rd columns).

3.3 Diverse Visual Generation for VAR

Based on above observations, we propose **DiverseVAR** for VAR to trigger the emergence of generative diversity while preserving generation fidelity in the VAR inference (Fig. 6 and Algorithm 1 (See *Suppl.*)). As shown in Fig. 6 (Left), the vanilla VAR inference processes the intermediate feature $\widetilde{\boldsymbol{F}}_{k-1}$ through a VAR block to produce the output feature $\boldsymbol{F}_k^o$ as well as its predicted logits, which is subsequently quantized into $\boldsymbol{R}_k$ (The predicted logits, the quantization step, and $\boldsymbol{R}_k$ are omitted in Fig. 6 for brevity). Our method operates at the block level (e.g., 8 blocks in the Infinity backbone).

Diversity emergence via Pivotal Component. For the pivotal component, it is difficult to disentangle it within the pivotal token, as the self-attention mechanism in Transformer-based VAR models inherently entangles information across tokens (Vaswani et al., 2017; Han et al., 2024) in the feature map $\widetilde{\boldsymbol{F}}_{k-1}$. Inspired by (Gu et al., 2014), we assume that the dominant singular values of the feature map $\widetilde{\boldsymbol{F}}_{k-1}$ correspond to the fundamental information, i.e., the pivotal component. To avoid directly using the pivotal token as the pivotal component, we instead regard the dominant singular values as determining the pivotal component.



Figure 7: Dominant singular values correspond to the pivotal component influence diversity.

Specifically, given the feature map $\widetilde{\boldsymbol{F}}_{k-1} \in \mathbb{R}^{S_k \times D}$ at the k-th (early) scale [2], we decompose it using Singular Value Decomposition (SVD): $\widetilde{\boldsymbol{F}}_{k-1} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} = diag(\sigma_1, \cdots, \sigma_n)$, the singular values $\sigma_1 \geq \cdots \geq \sigma_n$, $n = \min(S_k, D)$. Then, our objective is to suppress the pivotal component, thereby triggering the emergence of diversity. To suppress the pivotal component of the feature map $\widetilde{\boldsymbol{F}}_{k-1}$, we introduce the *Soft-Suppression Regularization* (**SSR**) for each singular value:

$$\hat{\sigma} = \alpha e^{-\beta\sigma} \times \sigma. \tag{4}$$

here, e represents the exponential function, while α and β are parameters constrained to be positive. For brevity, the subscript of σ is omitted. The feature map is then reconstructed as $\hat{\boldsymbol{F}}_{k-1} = \mathbf{U}\hat{\boldsymbol{\Sigma}}\mathbf{V}^T$, with $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$ denoting the updated singular values. The feature map $\hat{\boldsymbol{F}}_{k-1}$ is passed through a block to generate the output $\hat{\boldsymbol{F}}_k^o$. In a special case, we reset the top-K singular values to 0 (here, K=2). Interestingly, as shown in Fig. 7 (2nd row), removing the dominant components leads to improved diversity but reduces image fidelity. This supports our assumption that the dominant singular values of the feature map correspond to the pivotal component that influence generative diversity (Fig. 7 (2nd row)), and also shows that our method suppresses the pivotal component more effectively than directly zeroing it out (Fig. 7 (3rd row)).

[2] i.e., $S_k = h_k \times w_k$, $D = 2048$ in the Infinity model (Han et al., 2024)

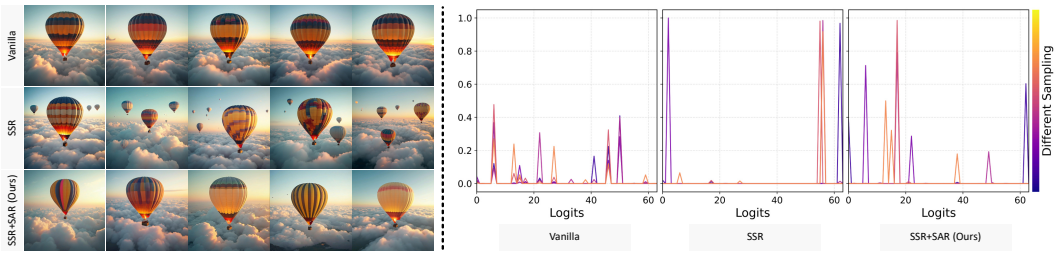

Figure 8: (Left) The generated images often fail to reflect the number described in the text prompt when using only SSR. For example, given the prompt "**A hot air balloon** floating above the clouds", SSR fails to generate the correct quantity of the **hot air balloon** (2nd row). (Right) The logits distribution under different samplings of vanilla model (Left), SSR (Middle), and SSR+SAR (Right).

Guided Diversity Formation. *Soft-Suppression Regularization* facilitates diversity emergence via the pivotal component. However, while enhancing diversity, it weakens the alignment with the text semantics, especially in cases involving numerical descriptions (Fig. 8 (Left (2nd row))). Motivated by (Ma et al., 2025), we analyze the distribution of the predicted logits for the output feature $\boldsymbol{F}_k^o$ to understand the underlying cause of this phenomenon. We observe that the logits distributions under different samplings are similar in the vanilla VAR (Fig. 8 (Right (Left))), while *Soft-Suppression Regularization* leads to more dispersed logits distributions across different samplings (Fig. 8 (Right (Middle))). Specifically, the logits distributions observed in the vanilla VAR reveal that the probability peaks of different samplings nearly coincide (Fig. 8 (Right (Left))), leading to highly similar samples and a lack of diversity in the generated images. In contrast, incorporating *Soft-Suppression Regularization* yields more dispersed logits distributions (Fig. 8 (Right (Middle))), with peaks varying significantly in both position and magnitude, thereby promoting the emergence of diversity in the generated images. However, we observe that independent and very high probability peaks can distort the representation of numerical attributes, leading to suboptimal image-text alignment.

Here, we aim to further improve image-text alignment, especially reducing inaccuracies in numerical attributes, while maintaining diversity. Based on the above analysis, we introduce an augmentation for the output feature $\hat{\boldsymbol{F}}_k^o$ to further guide the formation of diversity. Specifically, we perform singular value decomposition (SVD) on $\hat{\boldsymbol{F}}_k^o$, yielding singular values $\hat{\sigma}_1 \geq \cdots \geq \hat{\sigma}_n$. We then augment each singular value according to the following rule, as *Soft-Amplification Regularization* (**SAR**):

$$\tilde{\sigma} = \hat{\alpha} e^{\hat{\beta}\hat{\sigma}} \times \hat{\sigma}, \tag{5}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are positive parameters controlling the scaling strength. The recovered feature $\boldsymbol{F}_k^o = \mathbf{U}\widetilde{\boldsymbol{\Sigma}}\mathbf{V}^T$, where $\widetilde{\boldsymbol{\Sigma}} = diag(\tilde{\sigma}_1, \cdots, \tilde{\sigma}_n)$. The *Soft-Augmentation Regularization*, by encouraging more dispersed logit distributions and avoiding the formation of isolated peaks across different samplings, can guide the formation of diversity (Fig. 8 (Left (3rd row) and Right (Right))).

4 Experiment

4.1 Experimental Setup

Evaluation Datasets and Metrics. We evaluate **DiverseVAR** on the text-to-image VAR models Infinity-2B and Infinity-8B (Han et al., 2024), generating images at a resolution of 1024×1024. Following the vanilla Infinity-2B and Infinity-8B models (Han et al., 2024), we evaluate **DiverseVAR** on two widely used benchmarks (Han et al., 2024; Li et al., 2025a; Chen et al., 2025): GenEval (Ghosh et al., 2023) and DPG (Hu et al., 2024). We further evaluate the generative diversity of our method in comparison with the vanilla model on the zero-shot text-to-image benchmarks COCO 2014 and COCO 2017 (Lin et al., 2014). On the COCO 2014 benchmark, we adopt the conventional evaluation protocol (Sauer et al., 2023a; Kang et al., 2023d; Saharia et al., 2022; Rombach et al., 2022), using the 30K text prompts selecte by GigaGAN (Kang et al., 2023c) to generate 30K corresponding images. On the COCO 2017 benchmark, we generate 5K images from the 5K provided text prompts. Following prior works (Kang et al., 2023a; Han et al., 2025), we evaluate the diversity of the synthesized images using Fréchet Inception Distance (FID) (Heusel et al., 2017), Recall (Kynkäänniemi et al., 2019), and Coverage (Cov.) (Naeem et al., 2020). To assess the text-image alignment, we employ CLIPScore (CLIP) (Hessel et al., 2021), where ViT-B/32 is used.

Table 1: Quantitative comparison between the vanilla model and our DiverseVAR on COCO2014 (30K prompts) and COCO2017 (5K prompts), evaluated using FID, Recall, and Coverage (Cov.) for generative diversity, and CLIP-Score (CLIP) for text–image alignment.

Dataset	COCO2014-30K				COCO2017-5K			
Method	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
HART	0.262	0.518	27.46	0.310	0.345	0.728	37.71	0.309
+Ours	**0.293**	**0.557**	**25.67**	0.395	**0.480**	**0.766**	**33.21**	0.309
Infinity-2B	0.316	0.651	28.48	0.313	0.408	0.832	39.01	0.313
+Ours	**0.385**	**0.690**	**22.96**	0.313	**0.480**	**0.860**	**33.39**	0.313
Infinity-8B	0.451	0.740	18.79	0.319	0.563	0.892	29.47	0.319
+Ours	**0.497**	**0.748**	**14.26**	0.315	**0.585**	**0.892**	**25.01**	0.316



Figure 9: The vanilla model tends to produce results with similar styles under the same prompt, leading to limited diversity, while our results are more consistent with those from AFHQ.

Implementation Details. Following our observations that diversity formation primarily occurs at early scales, we apply at scales $\{4, 6\}$ to unleash diversity, while retaining the vanilla inference process in the remaining scales $\{1, 2, 8, 12, 16, 20, 24, 32, 40, 48, 64\}$ to maintain fidelity. Our operation is applied at the block level, spanning all 8 blocks of the Infinity backbone. We set the parameters as $\alpha = 1.0$ and $\beta = 0.01$ in Eq. (4), and $\hat{\alpha} = 1.0$ and $\hat{\beta} = 0.001$ in Eq. (5). All experiments are performed on a single NVIDIA A100 GPU equipped with 40 GB of memory. See *Suppl.* for details.

4.2 Main Results

Diversity Comparison. To demonstrate the diversity of **DiverseVAR**, we report quantitative evaluations using Recall, Coverage, and FID. Generally, Recall and Coverage are commonly used to assess the diversity by comparing the support of real and generated data distributions (Kang et al., 2023a; Naeem et al., 2020).

As reported in Tab. 1, our method achieves higher Recall, Coverage, and FID than the vanilla Infinity on both benchmarks, while maintaining comparable CLIP scores. Fig. 1 shows generations from the vanilla Infinity and our method, qualitatively demonstrating that our approach produces more diverse outputs. Furthermore, to evaluate the diversity under multiple samplings, we use booth AFHQ (Choi et al., 2020) and CelebA-HQ (Karras et al., 2017). We use three prompts corresponding to the three categories of AFHQ and two prompts corresponding to CelebA-HQ, each generating approximately 5,000 images. During evaluation, the training sets of both AFHQ and CelebA-HQ are used as the ground truth. As reported in Tab. 2, compared with the vanilla Infinity-2B and Infinity-8B, our method achieves superior scores in terms of Recall, Coverage, and FID on both datasets, except for Recall on the Infinity-8B model. Fig. 9 illustrates that our generated results (3rd row) more closely resemble the real images from AFHQ (1st row). See *Suppl.* for details.

Overall Comparison. We evaluate the performance of our method against the vanilla VAR model (Han et al., 2024), diffusion-based models (Betker et al., 2023; Podell et al., 2023; Chen et al., 2024), and AR models (Xie et al., 2024; Wang et al., 2024; Tang et al., 2024; Sun et al., 2024). Tab. 3 summarizes the comparison results on the GenEval (Ghosh et al., 2023) and DPG (Hu et al., 2024) benchmarks. As shown in Tab. 3, our method surpasses most competing approaches except, except DALL-E 3 on the DPG benchmark, while maintaining performance on par with the vanilla VAR model. Specifically, both our method and Infinity achieve over 0.7 on GenEval and around 83.0 on DPG, outperforming most competing methods except DALL-E 3, which demonstrates strong overall performance and consistency with the vanilla Infinity. These results indicate that our method generates images with significantly higher diversity while maintaining text-image alignment and visual fidelity compared to the vanilla Infinity model. See *Suppl.* for additional results.

Table 2: Quantitative comparison between the vanilla model and our **DiverseVAR** under multiple sampling runs on AFHQ and CelebA-HQ.

Dataset	AFHQ				CelebA-HQ			
Method	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
Infinity-2B	0.0	0.007	91.85	0.276	0.003	0.007	81.16	0.259
+Ours	0.012	0.020	78.88	0.278	0.017	0.128	62.07	0.257
Infinity-8B	0.0	0.001	126.47	0.271	0.0	0.0	149.95	0.241
+Ours	0.0	0.005	109.73	0.270	0.0	0.013	139.73	0.235

Table 3: Quantitative comparisons on GenEval and DPG benchmarks.

Method \ Metrics	Steps↓	Params↓	GenEval				DPG		
			Two Obj.	Position	Color Attri.	Overall↑	Global	Relation	Overall↑
SDXL (Podell et al., 2023)	40	2.6B	0.74	0.15	0.23	0.55	83.27	86.76	74.65
LlamaGen (Sun et al., 2024)	1024	0.8B	0.34	0.07	0.04	0.32	-	-	65.16
Show-o (Xie et al., 2024)	1024	1.3B	0.80	0.31	0.50	0.68	-	-	67.48
PixArt-Sigma (Chen et al., 2024)	20	0.6B	0.62	0.14	0.27	0.55	86.89	86.59	80.54
SD3-medium (Esser et al., 2024)	28	2.0B	0.74	0.34	0.36	0.62	-	-	-
DALL-E 3 (Betker et al., 2023)	-	-	-	-	-	0.67	90.97	90.58	83.50
Emu3 (Wang et al., 2024)	-	8.5B	0.81	0.49	0.45	0.66	-	-	81.60
HART (Tang et al., 2024)	14	0.7B	0.50	0.16	0.82	0.48	81.76	87.50	74.76
+ Ours	14	0.7B	0.48	0.16	0.82	0.49	83.28	87.74	74.48
Infinity-2B (Han et al., 2024)	13	2.0B	0.84	0.45	0.55	0.73	84.80	93.04	82.97
+ Ours	13	2.0B	0.85	0.41	0.53	0.70	85.11	92.26	83.02
Infinity-8B (Han et al., 2024)	13	8.0B	0.90	0.61	0.68	0.79	85.10	94.50	86.60
+ Ours	13	8.0B	0.89	0.59	0.66	0.76	84.80	94.93	86.78

Table 4: Ablation study of each component in **DiverseVAR**. We evaluate perceptual quality (GenEval, DPG) and generative diversity (COCO2014, COCO2017). † and ‡ denote operations applied to input and output features respectively.

Method \ Metrics	GenEval				DPG			COCO2014-30K				COCO2017-5K			
	Two Obj.	Position	Color Attri.	Overall↑	Global	Relation	Overall↑	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
① SSR†	0.81	0.39	0.53	0.68	83.89	92.45	83.01	0.375	0.689	23.77	0.313	0.440	0.858	34.65	0.313
② SAR†+SSR‡	0.84	0.40	0.53	0.69	84.19	92.57	82.79	0.352	0.686	24.98	0.313	0.456	0.857	35.36	0.313
③ SSR†+SAR†	0.83	0.41	0.51	0.68	82.67	92.80	82.77	0.380	0.694	23.18	0.313	0.461	0.861	33.88	0.313
④ SAR†+SSR†	0.81	0.41	0.51	0.68	82.97	91.37	82.72	0.374	0.688	23.14	0.313	0.471	0.871	33.85	0.313
⑤ SSR†+SAR‡ (**Ours**)	**0.85**	**0.41**	**0.53**	**0.70**	**85.11**	92.26	**83.02**	**0.385**	**0.690**	**22.96**	0.313	**0.480**	**0.860**	**33.39**	0.313

4.3 Additional Analysis

Ablation Study of Components. Based on our analysis and observation, **DiverseVAR** applies *Soft-Suppression Regularization* (i.e., **SSR**) to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$, and *Soft-Augmentation Regularization* (i.e., **SAR**) to the output feature $\boldsymbol{F}_k^o$ (i.e., ⑤ in Tab. 4). We perform an ablation study on these components, exploring alternative designs and comparing performance. The ablated designs include: ① SSR†: The SAR is removed, and only the SSR is applied to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$. ② SAR†+SSR‡: Apply the SAR to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$, and the SSR to the output feature $\boldsymbol{F}_k^o$. ③ SSR†+SAR†: First, apply SSR and then SAR to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$. ④ SAR†+SSR†: First, apply SAR and then SSR to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$.

Here, † indicates that the operation is applied to the input feature $\widetilde{\boldsymbol{F}}_{k-1}$, while ‡ indicates that the operation is applied to the output feature $\boldsymbol{F}_k^o$. Performance results and comparisons are presented in Tab. 4. We observe that our method (i.e., ⑤) outperforms other ablated designs (i.e., ①-④) in both perceptual quality and diversity, except for the Relation on DPG and the Coverage metric on COCO2017.



Figure 10: Comparison under a complex prompt.

Additional Results. We further evaluate **DiverseVAR** under long, complex prompts. The result shows that it preserves diversity and semantic alignment (Fig. 10); see *Suppl.* for more results.

5 Conclusion

We explore diversity in text-to-image VAR models and find that the pivotal component critically shapes diversity formation at early scales. Building on this insight, we introduce **DiverseVAR**, a simple yet effective framework that unleashes the intrinsic generative diversity of VAR models while preserving image fidelity, through input-level suppression and output-level augmentation of the pivotal component. We conduct extensive experiments and demonstrate that our approach effectively improves generative diversity while preserving image quality and text–image alignment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing. GPT-based tools were used to improve the language, grammar, and readability of the manuscript without changing the underlying scientific content. The underlying data and numerical results in the figure were not generated or modified by AI.

All AI-assisted outputs were manually reviewed and verified by the authors. The authors take full responsibility for the final manuscript, including all text, figures, experimental results, scientific claims, and conclusions.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Zhuokun Chen, Jugang Fan, Zhuowei Yu, Bohan Zhuang, and Mingkui Tan. Frequency-aware autoregressive modeling for efficient high-resolution image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *European Conference on Computer Vision*, pp. 176–192. Springer, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2):112–122, 1973.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. *arXiv preprint arXiv:2503.10637*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152. Curran Associates, Inc., 2023.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2862–2869, 2014.

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.

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, 2024.

Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, and Jaesik Choi. Enhancing creative generation on stable diffusion-based models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28609–28618, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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. *Advances in neural information processing systems*, 30, 2017.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

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

Doohyuk Jang, Sihwan Park, June Yong Yang, Yeonsung Jung, Jihun Yun, Souvik Kundu, Sung-Yub Kim, and Eunho Yang. Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. *arXiv preprint arXiv:2410.03355*, 2024.

Mingu Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 2020.

Minguk Kang, Woohyeon Shim, Minsu Cho, and Jaesik Park. Rebooting acgan: Auxiliary classifier gans with stable training. In *Neural Information Processing Systems*, 2021.

MinGuk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023a.

Minguk Kang, Joonghyuk Shin, and Jaesik Park. Studiogan: a taxonomy and benchmark of gans for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023c.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for Text-to-Image Synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023d.

Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *European Conference on Computer Vision*, pp. 428–447. Springer, 2024.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *CoRR*, abs/1904.06991, 2019.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11523–11532, 2022.

Kunjun Li, Zigeng Chen, Cheng-Yen Yang, and Jenq-Neng Hwang. Memory-efficient visual autoregressive modeling with scale-aware kv cache compression. *arXiv preprint arXiv:2505.19602*, 2025a.

Senmao Li, Lei Wang, Kai Wang, Tao Liu, Jiehang Xie, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, and Jian Yang. One-way ticket: Time-independent unified encoder for distilling text-to-image diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024.

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*, pp. 740–755. Springer, 2014.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Huaian Chen, and Yi Jin. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 2024.

Xiaoxiao Ma, Feng Zhao, Pengyang Ling, Haibo Qiu, Zhixiang Wei, Hu Yu, Jie Huang, Zhixiong Zeng, and Lin Ma. Towards better & faster autoregressive image generation: From the perspective of entropy, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Yatian Pang, Peng Jin, Shuo Yang, Bin Lin, Bin Zhu, Zhenyu Tang, Liuhan Chen, Francis E. H. Tay, Ser-Nam Lim, Harry Yang, and Li Yuan. Next patch prediction for autoregressive visual generation, 2024.

Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *International Conference on Machine Learning*, 2023a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023b.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

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.

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. *arXiv preprint arXiv:2410.10812*, 2024.

Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10748–10757, 2022.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, June 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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. *arXiv preprint arXiv:2408.12528*, 2024.

Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024.

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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

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

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*, pp. 586–595, 2018.

APPENDIX

A OVERVIEW

This supplementary material provides additional implementation details, formal algorithmic descriptions, comprehensive parameter sensitivity analyses, and extended qualitative results to support the findings presented in the main paper. The contents are organized as follows:

- **Statements (Sec. B)**: Detailed discussion on the limitations, broader impacts, ethical considerations, and reproducibility of our study.
- **Implementation Details (Sec. C)**: Configuration of datasets, evaluation metrics, baseline implementations, and the technical derivation of the elbow point for token categorization.
- **Algorithm Detail of DiverseVAR (Sec. D)**: A step-by-step procedural pseudocode of our proposed training-free diversity enhancement method.
- **Additional Analysis (Sec. E)**: Extensive ablation studies covering the selection of scales, blocks, Top-K singular values, and sensitivity analyses of hyperparameters ($\alpha, \beta, \hat{\alpha}, \hat{\beta}$).
- **Diversity under Varying Aspect Ratios (Sec. F)**: Qualitative comparisons showing that **DiverseVAR** improves diversity across different image aspect ratios.
- **Additional Results (Sec. G)**: Extended qualitative comparisons between the vanilla VAR and **DiverseVAR** across diverse text prompts to further demonstrate the improvement in generative variability.

B STATEMENTS

Limitations. This study focuses on unlocking the diversity inherent in Visual Autoregressive (VAR) models in a training-free manner. While **DiverseVAR** effectively enhances generative diversity and mitigates the diversity collapse without compromising image quality, training-free approaches inherently have an upper performance bound constrained by the pre-trained model's latent space. Additionally, the selection of optimal scales for regularization still requires empirical heuristics. In future work, we will explore training-based strategies to more fundamentally expand the model's creative capacity.

Broader Impacts. **DiverseVAR** improves the output variability of VAR models, which is crucial for creative industries and personalized content generation. However, enhanced diversity also carries potential ethical risks. The ability to generate a wider range of realistic visual content could be misused to create more convincing deepfakes or misleading imagery, potentially contributing to misinformation. Its application to human subjects may also raise privacy concerns. We urge the community to apply these techniques responsibly within established safety and copyright frameworks.

Ethical Statement. We acknowledge the ethical implications of deploying high-fidelity generative models, including risks related to privacy, data misuse, and the potential amplification of societal biases present in training data. All models used in this research (e.g., VAR, Infinity) are publicly available for research purposes. We are committed to fostering transparency and will release our modified code to ensure the results are verifiable and to support the development of responsible AI.

Reproducibility Statement. To ensure the reproducibility of our findings, we provide detailed descriptions of the Soft-Suppression Regularization (SSR) and Soft-Augmentation Regularization (SAR) in Sec. 3. The full source code, pre-trained model configurations, and evaluation scripts used to generate all results in this paper will be made publicly available on GitHub upon acceptance. All experiments were conducted using standard publicly accessible datasets (e.g., ImageNet, MS-COCO).

C Implementation Details

C.1 Configure

Datasets. We utilize the AFHQ (Choi et al., 2020) and CelebA-HQ (Karras et al., 2017) datasets to evaluate diversity under multiple sampling scenarios, as they contain diverse animal and human face images, respectively. We employ text prompts in the following formats: "A face of a <cat/dog/wild animal>" and "A face of a <man/woman>".

Metrics. We leverage the code from the popular GitHub repository "StudioGAN" (Kang et al., 2023b; 2021; Kang & Park, 2020)[3] to compute two metrics: Recall (Kynkäänniemi et al., 2019) and Coverage (Naeem et al., 2020). For FID (Heusel et al., 2017) and CLIP Score (Hessel et al., 2021), we employ the official evaluation code from GigaGAN (Kang et al., 2023d)[4] We use the official evaluation code for GenEval (Ghosh et al., 2023)[5] (also provided in the official Infinity code) and DPG (Hu et al., 2024)[6]

Baseline Implementations. We use the official implementation of Infinity-2B and Infinity-8B (Han et al., 2024)[7]. For Infinity-2B, we use all the hyperparameters at their default settings. For Infinity-8B, due to computational resource limitations, using A100 40GB GPUs leads to out-of-memory (OOM) errors. Therefore, we conduct both qualitative and quantitative evaluations on the results at scale 48.

C.2 The Pivotal and Auxiliary Token

In our ***Observation 2***, we employ the defined pivotal score and the Maximum Distance to Chord (MDC) to identify the *pivotal* and *auxiliary* token. Specifically, we follow (Guo et al., 2025) and define the pivotal score $s_{k,i} = \|\widetilde{\boldsymbol{F}}_{k-1,i} - \bar{\boldsymbol{F}}_{k-1}\|_2$ using the L2 norm, where $\bar{\boldsymbol{F}}_{k-1}$ represents the mean feature map obtained by averaging $\widetilde{\boldsymbol{F}}_{k-1}$ across scale dimensions, and $\widetilde{\boldsymbol{F}}_{k-1,i}$ denotes each token along the scale dimension. Then, to determine the boundary between *pivotal* and *auxiliary* tokens, we follow the MDC method (Douglas & Peucker, 1973). Given the sorted sequence $\{s_{k,i_n}\}_{n=1}^{L}$ in descending order, we construct a chord connecting the endpoints $(1, s_{k,i_1})$ and (L, s_{k,i_L}), where $L = S_k$ is the scale dimension. The perpendicular distance of each intermediate point (i_n, s_{k,i_n}) to this chord is computed as

$$d_{i_n} = \frac{\left|(s_{k,i_L} - s_{k,i_1}) \cdot i_n - (L-1) \cdot s_{k,i_n} + L \cdot s_{k,i_1} - s_{k,i_L}\right|}{\sqrt{(s_{k,i_L} - s_{k,i_1})^2 + (L-1)^2}}. \tag{6}$$

The index $i_n^* = \arg\max_{i_n} d_{i_n}$ corresponding to the maximum distance is defined as the elbow point, which separates the tokens into

$$\mathcal{T}_{\text{pivotal}} = \{1, \ldots, i_n^*\}, \quad \mathcal{T}_{\text{auxiliary}} = \{i_n^* + 1, \ldots, L\}. \tag{7}$$

Tokens before the elbow point exhibit higher deviation from the mean feature and are thus considered *pivotal*, while those after are treated as *auxiliary*.

C.3 Text Prompts

We list the text prompts used for image generation in this paper below.

Fig. 1: "A man in a clown mask eating a donut", "A cat wearing a Halloween costume", "Golden Gate Bridge at sunset, glowing sky, dramatic perspective, vivid colors", "A palace under the sunset", "A cool astronaut floating in space", "A cat riding a skateboard down a hill", "A bear wearing sunglasses on a sunny day", and "A cat sitting on a window ledge watching rain".

Fig. 2: "a green train is coming down the tracks".

[3] https://github.com/POSTECH-CVLab/PyTorch-StudioGAN
[4] https://github.com/lucidrains/gigagan-pytorch
[5] https://github.com/djghosh13/geneval
[6] https://github.com/TencentQQGYLab/ELLA
[7] https://github.com/FoundationVision/Infinity

Fig. 4: "A bird perched on a tree branch in the rain".

Fig. 7: "A bear fishing with a stick by a calm river".

Fig. 8: "A hot air balloon floating above the clouds".

Fig. 9: "The face of a cat".

Fig. 10: "A bullet time photography of a beautiful 90 year old woman in a colorful futuristic felt dress covered with buttons, wearing a futuristic hat, Gucci style earrings, goggles, bright red lipstick, and surrounded by colored rain".

Fig. S1: "A dog running through water surrounded by flowers" and "A cozy house surrounded by autumn trees".

Fig. S2: "A very cute cat near a bunch of birds", "A cat standing on a hill", "A photo of a cute rabbit holding a cup of coffee in a café", "A cinematic shot of a little pig priest wearing sunglasses", "A dog covered in vines", and "Cute grey cat, digital oil painting by Monet".

Fig. S3: "Editorial photoshoot of an old woman, high fashion 2000s fashion", "An astronaut riding a horse on the moon, oil painting by Van Gogh", "Full body shot, a French woman, photography, French streets", "A boy and a girl fall in love", "An abstract portrait of a pensive face, rendered in cool shades of blues, purples, and grays", and "Cute boy, hair looking up to the stars, snow, beautiful lighting, painting style by Abe Toshiyuki".

Fig. S4: "A table with a light on over it", "A library filled with warm yellow light", "A villa standing on a hill", "A train crossing a bridge over a canyon", "A bridge stretching over a calm river", and "A temple surrounded by flowers".

D ALGORITHM DETAIL OF **DIVERSEVAR**

Algorithm 1: DiverseVAR

Input : Scales $\{S_1, S_2, \cdots, S_K\}$. Scales for **DiverseVAR** consists of m scales $\{l_1, \cdots, l_m\}$. Scales for the vanilla process $\{S_1, S_2, \cdots, S_K\} \setminus \{l_1, \cdots, l_m\}$. The VAR model ϕ, the image decoder $\mathcal{D}$, and the quantizer $\mathcal{Q}$.

Output: The final diverse output $\mathbf{I}$.

$\boldsymbol{F}_0 = 0$;

$\widetilde{\boldsymbol{F}}_0 = \langle \text{SOS} \rangle \in \mathbb{R}^{1\times 1\times d}$; // The start token (Han et al., 2024)

for $k = 1, \cdots, K$ **do**

  if $S_k \in \{S_1, S_2, \cdots, S_K\} \setminus \{l_1, \cdots, l_m\}$ **then**

    $\boldsymbol{F}_k^o = \phi(\widetilde{\boldsymbol{F}}_{k-1})$; // Fig. 6 (Left)

  end

  else

    $\hat{\boldsymbol{F}}_{k-1} \leftarrow \widetilde{\boldsymbol{F}}_{k-1}$; // SSR: Eq. (4)

    $\hat{\boldsymbol{F}}_k^o = \phi(\hat{\boldsymbol{F}}_{k-1})$; // Fig. 6 (Right)

    $\boldsymbol{F}_k^o \leftarrow \hat{\boldsymbol{F}}_k^o$; // SAR: Eq. (5)

  end

  $\boldsymbol{R}_k = \mathcal{Q}(\boldsymbol{F}_k^o)$;

  $\boldsymbol{F}_k = \boldsymbol{F}_{k-1} + \text{up}(\boldsymbol{R}_k, (h_K, w_K))$; // Eq. (1)

  $\widetilde{\boldsymbol{F}}_k = \text{down}(\boldsymbol{F}_k, (h_k, w_k))$; // Eq. (3)

end

$\mathbf{I} = \mathcal{D}(\boldsymbol{F}_K)$

Return The final generated image $\mathbf{I}$

Table S1: Ablation study of different scales l_i on image generation quality (GenEval and DPG) and diversity (COCO 2014 and COCO 2017).

Dataset	GenEval				DPG			COCO2014-30K				COCO2017-5K			
Scales \ Metrics	Two Obj.	Position	Color Attri.	Overall↑	Global	Relation	Overall↑	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
$l_i = \varnothing$ (Vanilla)	0.84	0.45	0.55	0.73	84.80	93.04	82.97	0.316	0.651	28.48	0.313	0.408	0.832	39.01	0.313
$l_i \in \{2\}$	0.85	0.48	0.52	0.733	83.28	92.06	83.018	0.321	0.674	27.26	0.313	0.417	0.843	37.80	0.313
$l_i \in \{4\}$	0.83	0.44	0.58	0.716	83.28	92.49	83.256	0.333	0.660	27.11	0.314	0.430	0.833	37.64	0.313
$l_i \in \{6\}$	0.85	0.42	0.52	0.711	85.41	92.72	83.254	0.337	0.667	26.64	0.314	0.415	0.844	36.87	0.313
$l_i \in \{8\}$	0.84	0.45	0.52	0.725	83.89	92.49	82.891	0.329	0.662	27.35	0.314	0.422	0.838	37.92	0.313
$l_i \in \{2, 4\}$	0.83	0.44	0.52	0.712	82.06	92.22	82.859	0.340	0.680	25.43	0.313	0.435	0.851	35.92	0.312
$l_i \in \{4, 6\}$ **(Ours)**	0.85	0.41	0.53	0.70	85.11	92.26	83.02	0.385	0.690	22.96	0.313	0.480	0.860	33.39	0.313
$l_i \in \{6, 8\}$	0.80	0.40	0.47	0.677	83.89	92.45	82.996	0.366	0.693	23.24	0.313	0.452	0.855	33.84	0.313
$l_i \in \{2, 4, 6\}$	0.77	0.40	0.48	0.669	83.58	92.26	82.388	0.432	0.710	20.05	0.312	0.513	0.876	30.79	0.311
$l_i \in \{4, 6, 8\}$	0.75	0.34	0.42	0.622	83.58	92.37	81.630	0.499	0.700	15.85	0.310	0.567	0.878	26.35	0.310
$l_i \in \{2, 4, 6, 8\}$	0.70	0.35	0.35	0.593	82.37	92.45	80.749	0.544	0.687	14.70	0.306	0.611	0.857	25.67	0.304

E Additional Analysis

E.1 Ablation Study of Scales.

Based on our observation and analysis, early scales have a significant impact on the structural formation. Specifically, we apply our method to the VAR at scales 2 and 4 (i.e., $l_i \in \{4, 6\}$). To quantitatively investigate the impact of scales on image generation quality and diversity, we conducted experiments on GenEval and DPG to compare generation quality, and on COCO 2014 and COCO 2017 to evaluate generation diversity. As shown in Tab. S1 (1st–4th rows), applying **DiverseVAR** with a single early scale leads to consistent improvements in Recall, Coverage, and FID over the vanilla on COCO 2014 and COCO 2017, demonstrating enhanced diversity. Meanwhile, the performance on GenEval and DPG shows only marginal degradation, indicating that image generation quality is well preserved. To further stimulate model diversity, we apply **DiverseVAR** to multiple early scales and their various combinations. As illustrated in Tab. S1 (5th–10th rows), increasing the number of applied scales leads to higher diversity but causes a rapid degradation in generation quality. Notably, applying **DiverseVAR** to four early scales (i.e., $l_i \in \{2, 4, 6, 8\}$) yields the highest diversity metrics but the lowest generation quality, with GenEval decreasing by more than 0.1. Therefore, to achieve a good trade-off between generation quality and diversity, we apply **DiverseVAR** at scales 4 and 6 (i.e., $l_i \in \{4, 6\}$), which are used as the default settings.

E.2 Ablation Study of Blocks.

Our method operates at the block level, i.e., on 8 blocks b_i within the Infinity backbone, where $i \in \{0, 1, 2, 3, 4, 5, 6, 7\}$. Specifically, **DiverseVAR** is applied across all 8 blocks. To further assess the contribution of different blocks, we perform comparative experiments by selectively applying **DiverseVAR** to specific subsets of blocks. As illustrated in Tab. S2 (3rd–7th rows), applying **DiverseVAR** to only partial subsets of blocks results in a significant degradation in either image generation quality or diversity compared to applying it to all blocks. For example, applying **DiverseVAR** to blocks b_i with $i \in \{0, 1, 2, 3, 4, 5, 6\}$ yields the best diversity performance, achieving superior Recall, Coverage, and FID on COCO 2014 and COCO 2017, but results in a severe degradation of image quality (Tab. S2 (3rd row)). While applying **DiverseVAR** to blocks b_i with $i \in \{4, 5, 6, 7\}$ maintains the generation quality, the resulting diversity metrics remain suboptimal (Tab. S2 (6th row)). Therefore, to achieve a good trade-off between generation quality and diversity, we apply **DiverseVAR** at all 8 blocks, which are used as the default settings. For a more comprehensive comparison, we further apply the method at the model level. As illustrated in Tab. S2 (last row), this configuration maintains the generation quality (evaluated on GenEval and DPG), with GenEval remaining at 0.73, but yields almost no improvement in diversity (evaluated on COCO 2014 and 2017).

E.3 Ablation Study on *logits*

We observe that the *logits* distributions under different samplings are similar in the vanilla VAR, causing diversity collapse. Our proposed method mitigates this issue by introducing variation into

Table S2: Ablation study of different blocks b_i on image generation quality (GenEval and DPG) and diversity (COCO 2014 and COCO 2017).

Dataset	GenEval							DPG				COCO2014-30K				COCO2017-5K			
Blocks \ Metrics	Two Obj.	Position	Color Attri.		Overall↑	Global	Relation	Overall↑		Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
$i = \varnothing$	0.84	0.45	0.55	0.73		84.80	93.04	82.97		0.316	0.651	28.48	0.313	0.408	0.832	39.01	0.313
$i \in \{0,1,2,3,4,5,6,7\}$ **(Ours)**	0.85	0.41	0.53	0.70		85.11	92.26	83.02		0.385	0.690	22.96	0.313	0.480	0.860	33.39	0.313
$i \in \{1,2,3,4,5,6,7\}$	0.80	0.39	0.50	0.678		85.41	92.57	82.93		0.370	0.691	23.77	0.314	0.456	0.867	34.38	0.313
$i \in \{0,1,2,3,4,5,6\}$	0.80	0.38	0.45	0.657		83.89	92.72	82.60		0.411	0.699	21.06	0.314	0.485	0.868	31.92	0.313
$i \in \{2,3,4,5,6,7\}$	0.85	0.42	0.56	0.715		83.58	92.49	82.86		0.350	0.676	25.78	0.314	0.432	0.851	36.27	0.314
$i \in \{4,5,6,7\}$	0.82	0.42	0.56	0.722		84.19	92.26	83.03		0.338	0.659	27.22	0.314	0.427	0.846	37.15	0.314
$i \in \{1,2,3\}$	0.82	0.39	0.47	0.682		84.49	92.45	83.29		0.352	0.677	24.08	0.313	0.439	0.855	34.17	0.312
Model-level	0.85	0.43	0.56	0.73		84.19	92.80	83.03		0.317	0.651	28.29	0.314	0.402	0.828	38.98	0.313

Table S3: Ablation study on *logits* for image generation quality (GenEval and DPG) and diversity (COCO 2014 and COCO 2017).

Dataset	GenEval							DPG				COCO2014-30K				COCO2017-5K			
Method \ Metrics	Two Obj.	Position	Color Attri.		Overall↑	Global	Relation	Overall↑		Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
Vanilla	0.84	0.45	0.55	0.73		84.80	93.04	82.97		0.316	0.651	28.48	0.313	0.408	0.832	39.01	0.313
SSR in *logits*	0.86	0.45	0.82	0.72		83.28	91.99	83.41		0.320	0.604	28.95	0.313	0.419	0.838	38.91	0.313
SSR+SAR in *logits*	0.86	0.45	0.54	0.73		83.89	92.99	83.12		0.331	0.580	29.06	0.313	0.402	0.828	39.22	0.313
SSR+SAR in blocks **(Ours)**	0.85	0.41	0.53	0.70		85.11	92.26	83.02		0.385	0.690	22.96	0.313	0.480	0.860	33.39	0.313

the distributions, thereby encouraging diverse generations while maintaining faithful text–image alignment and high visual quality. To further examine whether **DiverseVAR** is more effective at the block level than at the logits level, we performed comparative experiments by applying it to the logits (Tab. S3 (2nd-3rd rows)). As shown in Tab. S3, when **DiverseVAR** is applied to the logits, the generation quality is well preserved — as indicated by a GenEval score of 0.72 and 0.73 on GenEval — while the diversity metrics on COCO 2014 and COCO 2017 show only a slight improvement.

E.4 Ablation Study on Parameter Sensitivity

The performance of **DiverseVAR** is governed by the scaling strength parameters α, β and $\hat{\alpha}, \hat{\beta}$. We conduct a comprehensive sensitivity analysis to evaluate their impact across semantic alignment and generative diversity benchmarks, as presented in Tab. S4.

Sensitivity to α and β. The parameters α and β control the strength of the SSR regularization, which balances semantic structure preservation and diversity enhancement. As shown in Tab. S4, when α is set to a low value (e.g., $\alpha = 0.7$), there is a significant drop in perceptual quality, with GenEval Overall decreasing to 0.43 and DPG to 69.73. This indicates that insufficient regularization fails to effectively mitigate the diversity collapse inherent in vanilla VAR. Increasing α to 1.5 or properly adjusting β (e.g., $\beta = 0.015$) substantially restores alignment quality (GenEval Overall ≈ 0.66) and improves diversity, achieving an FID of 19.26 on COCO2014.

Sensitivity to $\hat{\alpha}$ and $\hat{\beta}$. The parameters $\hat{\alpha}$ and $\hat{\beta}$ further regulate the refinement process. We observe that the model remains stable under a wide range of values. For example, varying $\hat{\alpha}$ from 0.7 to 1.5 maintains a consistent GenEval score (0.64 to 0.68) and a stable CLIP score (≈ 0.31). Our finalized configuration (**Ours**) achieves the most balanced performance, reaching the highest GenEval score of 0.70 and the best FID of 22.96 on COCO2014, demonstrating the effectiveness and robustness of our regularization design.

E.5 Ablation Study on Top-K Singular Values

In **DiverseVAR**, we apply regularization to the top-K singular values to modify the latent distribution while preserving essential semantic information. We evaluate the impact of different K values ($K \in \{1, 2, 4, 8\}$) in Tab. S5. The results indicate that **DiverseVAR** is robust to the choice of K.

Table S4: Parameter sensitivity analysis of **DiverseVAR**. We evaluate the impact of hyper-parameters α, β (Stage 1) and $\hat{\alpha}, \hat{\beta}$ (Stage 2) across perceptual quality (GenEval, DPG) and generative diversity (COCO 2014 and COCO 2017).

Dataset	GenEval				DPG			COCO2014-30K				COCO2017-5K			
Settings \ Metrics	Two Obj.	Position	Color Attri.	Overall↑	Global	Relation	Overall↑	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
α=**0.7**	0.50	0.18	0.53	0.43	77.20	89.86	69.73	0.604	0.402	41.60	0.274	0.664	0.590	54.32	0.274
α=**1.5**	0.80	0.38	0.80	0.67	83.28	91.41	82.15	0.342	0.655	26.44	0.312	0.420	0.837	36.90	0.312
β=**0.007**	0.84	0.44	0.84	0.72	82.97	92.10	82.57	0.358	0.667	25.79	0.313	0.449	0.846	36.09	0.314
β=**0.015**	0.79	0.37	0.81	0.66	84.19	93.34	82.72	0.446	0.708	19.26	0.313	0.505	0.876	30.34	0.313
$\hat{\alpha}$=**0.7**	0.79	0.34	0.78	0.64	83.28	92.68	81.54	0.495	0.695	15.87	0.310	0.571	0.871	26.64	0.310
$\hat{\alpha}$=**1.5**	0.83	0.40	0.80	0.69	82.97	92.92	82.94	0.356	0.677	24.61	0.313	0.461	0.854	34.65	0.313
$\hat{\beta}$=**0.0007**	0.82	0.42	0.85	0.69	82.06	92.45	82.07	0.392	0.689	23.10	0.313	0.491	0.872	33.14	0.313
$\hat{\beta}$=**0.0015**	0.81	0.42	0.82	0.69	83.28	92.80	82.18	0.387	0.692	22.97	0.314	0.477	0.862	33.26	0.314
Ours	0.85	0.41	0.53	0.70	85.11	92.26	83.02	0.385	0.690	22.96	0.313	0.480	0.860	33.39	0.313

Table S5: Ablation study on different TopK singular values in **DiverseVAR**. We evaluate the performance across perceptual quality (GenEval, DPG) and generative diversity (COCO 2014 and COCO 2017).

Dataset	GenEval				DPG			COCO2014-30K				COCO2017-5K			
Settings \ Metrics	Two Obj.	Position	Color Attri.	Overall↑	Global	Relation	Overall↑	Recall↑	Cov.↑	FID↓	CLIP↑	Recall↑	Cov.↑	FID↓	CLIP↑
$K = 1$	0.83	0.43	0.84	0.69	84.80	91.99	82.29	0.387	0.692	22.99	0.313	0.454	0.869	33.18	0.313
$K = 2$	0.83	0.38	0.82	0.69	83.28	92.80	82.44	0.385	0.690	23.25	0.314	0.478	0.866	33.60	0.313
$K = 4$	0.85	0.42	0.83	0.69	83.59	92.46	82.55	0.386	0.691	23.33	0.314	0.457	0.865	33.88	0.314
$K = 8$	0.82	0.42	0.84	0.69	83.28	91.95	82.27	0.378	0.696	23.24	0.314	0.477	0.864	33.24	0.314
Ours	0.85	0.41	0.53	0.70	85.11	92.26	83.02	0.385	0.690	22.96	0.313	0.480	0.860	33.39	0.313

Even with $K = 1$, where only the most dominant singular component is regularized, the model already achieves a high DPG score of 82.29 and an FID of 22.99 on COCO2014, effectively mitigating diversity collapse. As K increases to 8, the diversity metrics on COCO2017 remain stable (Recall $\approx$ 0.47, FID $\approx$ 33.24). Our default setting (**Ours**) strikes the optimal balance, yielding the highest perceptual alignment (0.70 on GenEval and 83.02 on DPG) and superior diversity (FID 22.96 on COCO2014). This robustness demonstrates that regularizing a small number of principal spectral components is sufficient to guide the model toward diverse and high-quality generations without requiring exhaustive per-task tuning of K.

F Diversity under Varying Aspect Ratios

The Infinity (Han et al., 2024) model originally supports image generation with varying aspect ratios, and our method, **DiverseVAR**, preserves this property. As shown in Fig. S1, when combined with **DiverseVAR**, it continues to facilitate efficient image generation while producing more diverse samples. These results demonstrate that our method can be readily extended to improve diversity across different image aspect ratios.

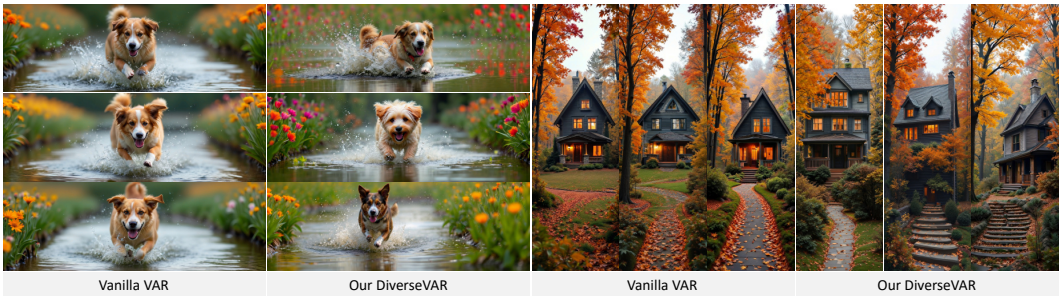

Figure S1: Qualitative comparison between the vanilla model and our method under varying aspect ratios. Our **DiverseVAR** produces more diverse samples while preserving the aspect-ratio generation capability of the vanilla model.



Figure S2: Additional diversity comparison. Our results demonstrate superior diversity.

G Additional Results

Diversity Comparison. As shown in Figs. S2 to S4, we present additional qualitative results for diversity comparison. We observe that the vanilla model tends to produce samples with limited diversity across different text prompts, although it generates high-quality results. In contrast, our method can generate diverse images from multiple samples under different text prompts, while maintaining good text–image alignment, indicating its advantage over the baselines.

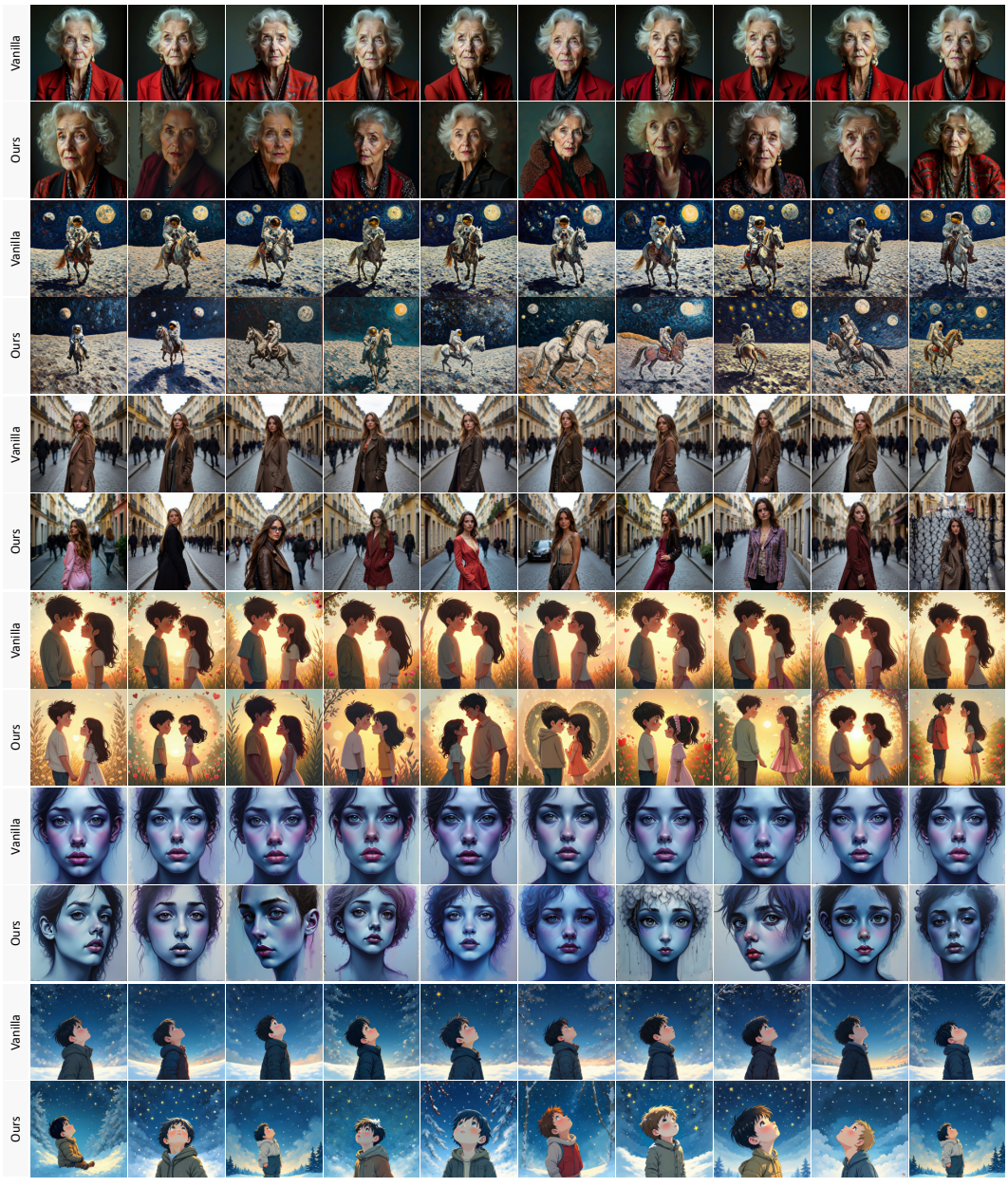

Figure S3: Additional diversity comparison. Our results demonstrate superior diversity.

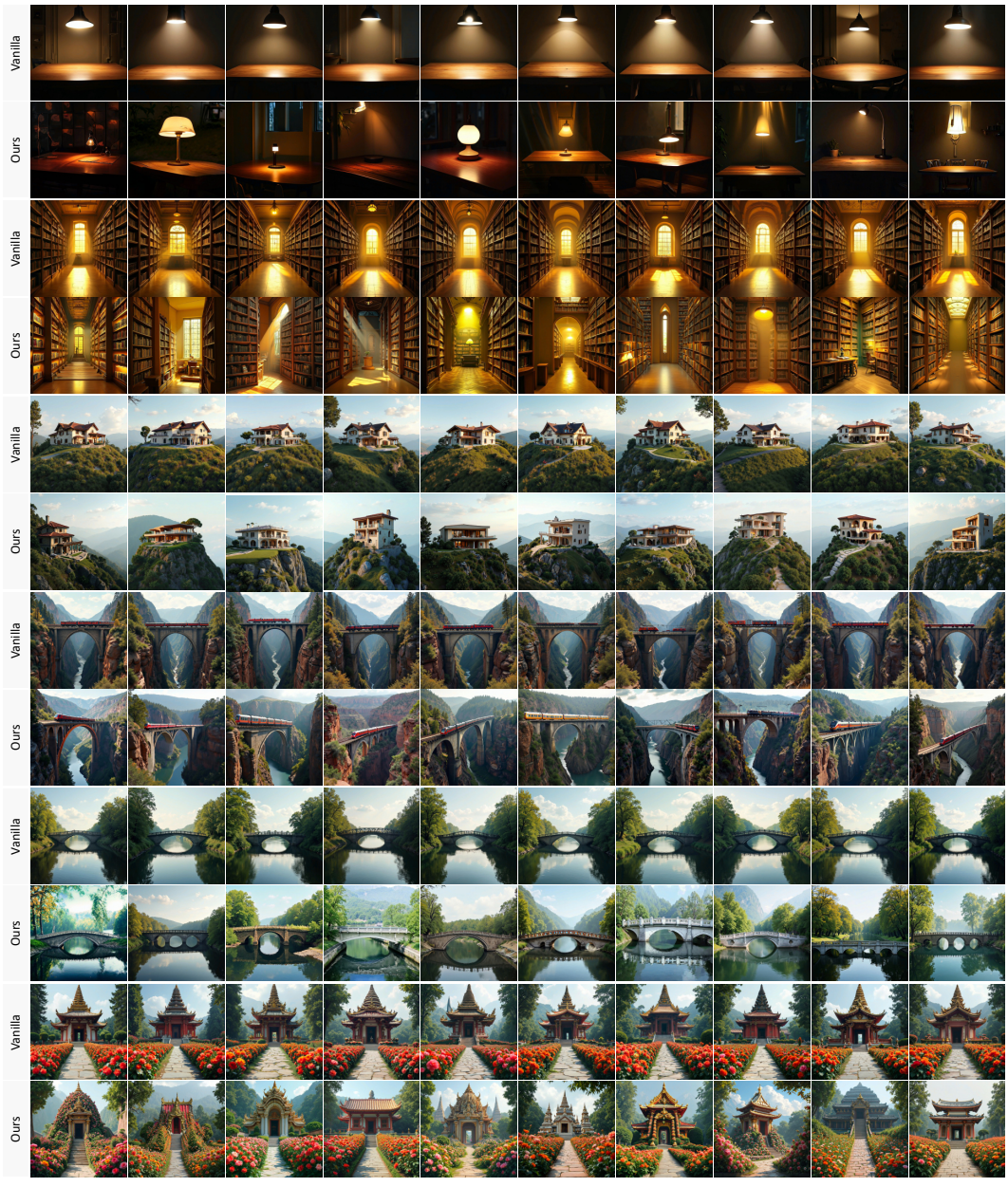

Figure S4: Additional diversity comparison. Our results demonstrate superior diversity.