SynVAR: Synergizing Spatial and Semantic Alignment in Visual Autoregressive Model

Zhennan Chen[1*], Tianxing Shi[1*], Pengcheng Xu[2], Kepan Nan[1]
, Qian Wang[3], Zili Yi[1], Jian Yang[1],, and Ying Tai[1†]

[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] Western University
[3] JIUTIAN Research, CMCC, China
https://github.com/NJU-PCALab/SynVAR

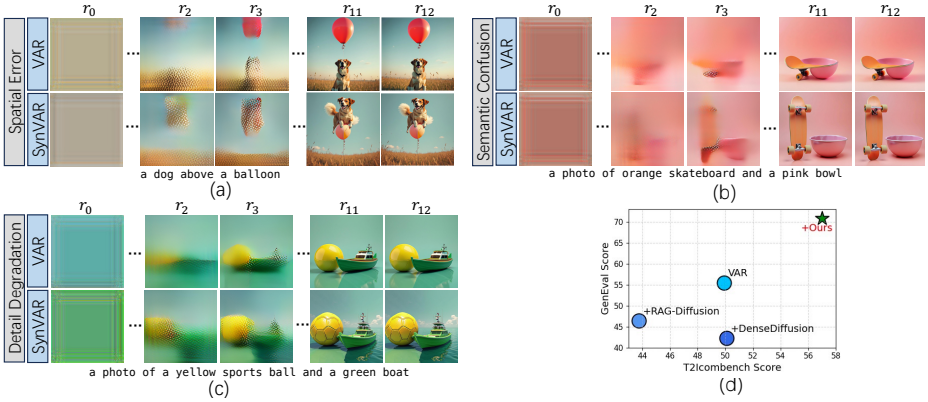

Fig. 1: (a)(b)(c): In complex scenes, existing VAR models are prone to spatial error, semantic confusion, and detail degradation. (d) Due to different formulations and modeling of the image, the solutions in existing diffusion models cannot play a positive role, while our method can significantly enhance the generation performance.

Abstract. VAR has gained widespread popularity due to its next-scale prediction paradigm. However, it faces substantial performance bottlenecks when handling complex scenes with multiple objects and attributes. Existing diffusion-based enhancement methods fail to adequately address the unique challenge of cross-scale error propagation and accumulation in VAR. To this end, we propose SynVAR, the first training-free enhancement framework specifically tailored for the VAR paradigm, which introduces a spatial-semantic collaborative control strategy to effectively suppress propagation error and improve generation quality. SynVAR comprises three key components: (1) Global guidance to ensure reasonable spatial structure in the early stages, (2) Receptive field constraints to

*Equal Contribution.
†Corresponding Author.

mitigate early-stage semantic confusion, (3) High-frequency compensation to recover fine-grained details. Extensive quantitative and qualitative experiments demonstrate the significant improvements in the ability of SynVAR to enhance the VAR's capability for complex scene modeling.

Keywords: Visual Autoregressive Model · Complex Scene Generation · Training-free Method

1 Introduction

The Visual Autoregressive (VAR) model [31] revolutionizes the generation paradigm of "next-token prediction" in the previous AR models with "next-scale prediction". This new formulation significantly advanced generated images' quality and sampling efficiency and induced new AR-based large-scale text-to-image models such as Infinity [18], Switti [35], and VARGPT [49]. Though thriving in quality and efficiency, they show deficiencies in spatial and semantic correctness in complex scenes, especially involving multiple objects, attributes, and spatial relations, which is essential and crucial for high-quality image generation. We demonstrate these issues in Figure 1(a)(b)(c). Specifically, the VAR may not accurately control the spatial relation of objects, and demonstrate visual semantics, and imagery details according to the text guidance in complex scenes. These indicate that the spatial and semantic relations in the VAR are not well aligned. To address these, we investigate the pattern within the VAR transformer during generation and propose a train-free framework for more precise and controllable generation of complex scenes, enhancing the spatial and semantic alignment in the VAR paradigm.

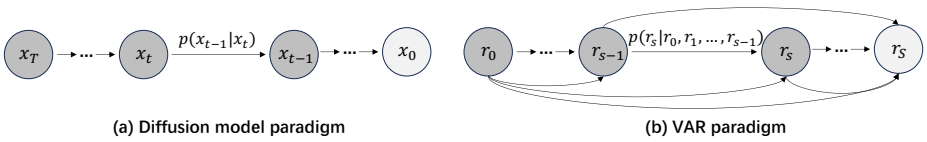

Fig. 2: Graphical models for diffusion and VAR inference models. Compared with the diffusion model, VAR is more prone to error accumulation and propagation.

In our investigation, simply extending diffusion-based enhancement methods such as iterative refinement [2, 9] and local attention [21, 38] to VAR faces a significant degradation, as shown in Figure 1(d). This is because VAR uses the different formulation and modeling of images from diffusion. As shown in Figure 2, in VAR, the prediction of each image scale is conditioned on all previous scales, whereas the image latent in diffusion is only conditioned on that previous one. Such a difference causes the spatial and semantic errors at the low-resolution scale of VAR more prone to propagate and accumulate and difficult

to restore fine texture details. This characteristic of VAR amplifies the degradation of VAR's spatial and semantic misalignment in complex scenes, which invalidates the techniques to improve the image quality. Thus, this motivates us to rethink spatial and semantic relations and enhance the generation rationality of VAR in complex scenes.

To address the issues above, we propose a collaborative correction mechanism for spatial and semantic control termed as SynVAR, which to our knowledge is the first training-free framework tailored for text-to-image models under the VAR paradigm. Specifically, the framework improves the spatial and semantic control from three aspects: 1) The global guidance optimizes the spatial partitioning strategy at the low-resolution scale stage and introduces explicit spatial priors in the cross-attention mechanism, ensuring the structural integrity of the image. 2) The receptive field constraint dynamically adjusts the interaction range of the front self-attention layers, effectively suppressing semantic confusion. 3) The high-frequency compensation enhances the high-frequency components of the features, enhancing the detail generation. Global guidance and receptive field constraint align spatial and semantic elements effectively, preventing error accumulation and ensuring that the final image accurately reflects the desired content. On this basis, the high-frequency compensation mechanism further enhances the detail performance of the image. Experimental results show that SynVAR consistently improves performance on various models such as Infinity and Switti, achieving average improvements of 19.6% and 7.9% on the overall score of the Geneval [15] and T2I-CompBench [20], respectively.

Our contributions are summarized as follows:

- We conducted a detailed investigation into the failure mechanisms of the VAR model in complex scenes. Based on these findings, we introduce SynVAR, the first training-free enhancement framework designed specifically to address these issues in the VAR paradigm.
- We propose an innovative approach for spatial and semantic alignment that combines global spatial guidance, receptive field constraints, and high-frequency compensation. This collaborative mechanism effectively reduces decision errors during the generation process and prevents the propagation of early-stage mistakes.
- As a plug-and-play solution, SynVAR significantly improves the performance of various VAR-based text-to-image models on different benchmarks, proving its effectiveness.

2 Related Work

Visual Autoregressive Modeling. Inspired by the success of language models [1, 3, 22, 30, 32, 33] in the visual generation domain, a discrete representation-based autoregressive paradigm has gradually been established. Early efforts, such as VQ-VAE [34] and VQGAN [14], established a strong foundation for visual autoregressive modeling by introducing the concept of encoding images into discrete visual tokens. These methods enabled the generation of images through a

sequence of discrete symbols, similar to how text is processed in language models. Subsequent efforts [4, 5, 12, 27] introduced this paradigm into text-to-image synthesis. Building on these developments, VAR proposed the "next-scale prediction" strategy, transforming traditional one-dimensional token prediction into multi-scale image block prediction, thereby significantly enhancing both generation quality and semantic consistency. Recent advances such as Infinity [18], Switti [35], and STAR [23] have further optimized image detail representation and text-image alignment from different perspectives. VARGPT [49] also adopts the next-scale prediction paradigm for visual autoregressive generation, as part of broader multimodal visual understanding and generation efforts. Despite these successes in standard scenarios, existing VAR models still struggle with challenges in spatial and semantic alignment when handling complex prompts involving multiple objects and intricate attribute combinations.

Complex Scene Generation. With the development of diffusion model [8, 10, 11, 19, 24–26, 28, 29, 39–45, 47, 48], various solutions have been proposed in the diffusion model domain for complex scene synthesis. To improve the results of combined generation, some methods [2,9,13,21,46] focus on integrating instance-level control through multi-stage sampling, where the generation process is iteratively refined to ensure the correct placement and attributes of objects. Some methods [7, 37] have explored the use of a scoring function to guide the generation process, enabling zero-shot layout control that adapts to complex scene requirements. Another category of solutions [6, 17] optimizes the initial noise to obtain better generation results. However, due to differences in image modeling approaches, these solutions are not suitable for VAR. In this paper, we introduce a training-free framework, SynVAR, tailored for the cross-scale pattern in the VAR paradigm. which establishes a collaborative correction mechanism of spatial and semantics, effectively enhancing the generation performance of VAR in complex scene synthesis.

3 Method

3.1 Preliminaries

VAR redefines autoregressive learning for images via next-scale prediction, inspired by human perception (coarse-to-fine hierarchy). Instead of token-by-token generation, VAR generates multi-scale token maps $(r_0, r_1, ..., r_S)$ autoregressively, starting from the coarsest resolution (1×1) to the target resolution $(H \times W)$. At each step s, the model predicts the entire token map r_s conditioned on all previous scales $r_0, r_1, ..., r_{s-1}$:

$$p\left(r_0, r_1, \ldots, r_S\right) = \prod_{s=1}^{S} p\left(r_s \mid r_0, r_1, \ldots, r_{s-1}\right) \cdot p(s_0) \tag{1}$$

Using this sequence of residuals, the feature f_s at the s-th scale step can be obtained from the following formula:

$$f_s = f_{s-1} + \phi\left(r_{s-1}, (h_s, w_s)\right) \tag{2}$$

where $\phi(\cdot)$ represents upsampling r_{s-1} to resolution $\{h_s, w_s\}$. f_s represents the cumulative sum of the upsampled sequence $(r_0, r_1, ..., r_{s-1})$. In the VAR generation process, once spatial or semantic errors occur in the early stage, this cross-scale dependency modeling can easily lead to the propagation and accumulation of errors, which is particularly common in complex scenarios.

3.2 Overview

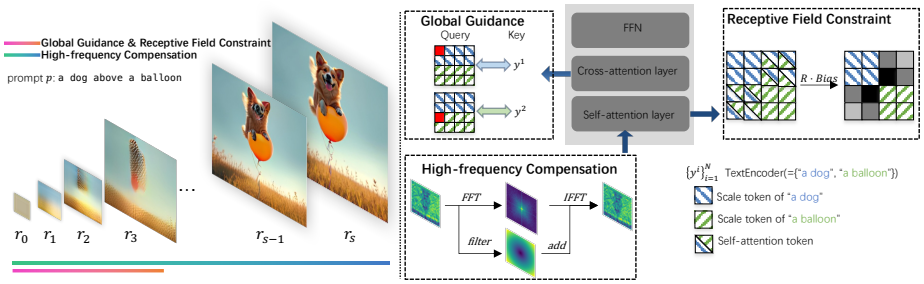

Fig. 3: Overall Framework of SynVAR. Global guidance and receptive field constraint are completed in the early steps of generation, while high-frequency compensation is continuously applied throughout the entire generation process. Specifically, global guidance introduces spatial priors into the cross-attention layers, receptive field constraint adjusts the perceptual scope of different regions in the self-attention layers, and high-frequency compensation enhances the high-frequency components of the features.

As shown in Figure 3, we propose SynVAR, a spatial-semantics collaborative correction mechanism based on the generative characteristics of VAR. We use global guidance to explicitly introduce spatial priors to enhance the accuracy of object localization and recognition. Through receptive field constraints, we reduce interdependencies between different regions, preventing excessive information cross-over. We also adopt a high-frequency enhancement strategy to improve image details further.

3.3 Global Guidance

Some studies [16,31,36] have shown that VAR primarily generates coarse-grained information in the early stages. As shown in Figure 4, our further experiments reveal that there is an early fixation of semantic and spatial information during the VAR generation process. Specifically, when we artificially modify the spatial positions and semantic embeddings at various time steps during the generation, only changes made in the earlier stages have a meaningful impact on the final output. Quantitative analysis in Figure 5 indicates that spatial alignment and semantic consistency stabilize early in the process, with minimal adjustments taking place in the later stages. Therefore, to ensure structural integrity during

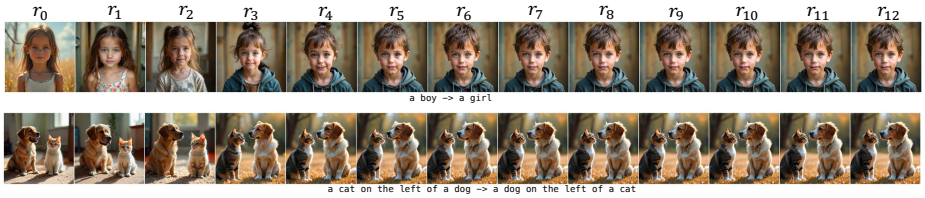

Fig. 4: Early fixation of semantic and spatial information. We force the replacement of semantic or spatial information in the VAR generation process, which can only affect the result in the early stage.

the model's generation process and avoid generation errors, we apply a global guidance in the early stages of generation.

In global guidance, we break the original input prompt p into a set of individual concepts $\{c^i\}_{i=1}^N$, where N denotes the number of concepts in the prompt. We assign a global region $\{h^i, w^i\}_{i=1}^N$ for each concept and each individual concepts $\{c^i\}_{i=1}^N$ is encoded into a text embedding $\{y^i\}_{i=1}^N$ via text encoder, which is then projected into K^i and V^i. Simultaneously, a query vector Q^i is extracted the token map r_{s-1}. The detailed formulation is as follows:

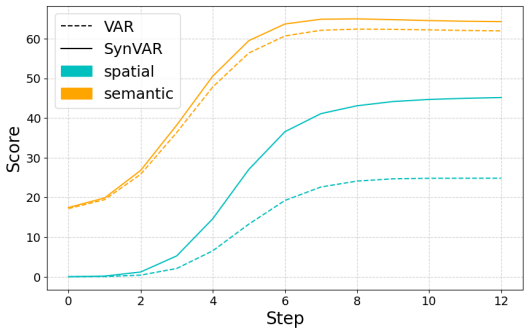

Fig. 5: Quantitative metrics at different steps. On the Geneval benchmark, the semantic and spatial related indicators are close to convergence in the early stage.

$$
\begin{aligned}
Q^i = \ell_Q\left(r_{s-1}\right), K^i = \ell_K(y^i), V^i = &\ell_V(y^i) \\
r_{s-1}^i = \text{Softmax}\left(\frac{Q^i K^{iT}}{\sqrt{d}}\right) V^i,&
\end{aligned}
\tag{3}
$$

where ℓ_Q, ℓ_K, ℓ_V are linear projections. Then, we crop and concatenate the token maps r_{s-1}^i containing conceptual information according to their respective corresponding regions to achieve the injection of spatial information:

$$
r_{s-1}^i(h^i, w^i) = \text{Crop}\left(r_{s-1}^i, (h^i, w^i)\right) \tag{4}
$$

$$
r_{s-1}^{cat} = \text{Concat}(r_{s-1}^i(h^i, w^i)_{i=1}^N) \tag{5}
$$

Finally, the sequence after the global guidance operation is used to obtain the output of the model:

$$
f_s = f_{s-1} + \phi\left(r_{s-1}^{cat}, (h_s, w_s)\right) \tag{6}
$$

This guidance strategy significantly improves early space errors, reduce generation deviations, and enable the model to accurately render and correctly position multiple objects in complex scenes.

3.4 Receptive Field Constraint

The VAR model enhances the receptive field range effectively by using two-dimensional sequences, which allows for more effective global contextual information. However, as illustrated in Figure 6, we observe that when multiple target regions are present, the attention dependencies between different regions remain high, especially when they are close to each other. This leads to a high coupling of information between different object regions, resulting in semantic confusion. To address this issue, we propose a receptive field constraint mechanism to regulate the attention dependencies between tokens, helping VAR obtain the correct semantic information.

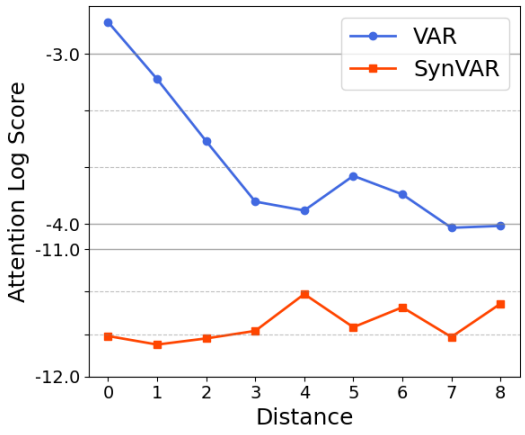

Fig. 6: Attention scores between different targets. the level of dependency among them remains consistently high, particularly in adjacent regions.

In receptive field constraint, we introduce a gaussian mask into the self-attention to calculate the relative spatial proximity of tokens in the early stage of generation. Specifically, a bias term is computed based on their Euclidean distance in the feature map, the formula is as follows:

$$Bias[q,k] = \exp\left(-\frac{||p_q - q_k||_2^2}{\sigma^2}\right) \tag{7}$$

where p_q and q_k represent the coordination of the corresponding token of the query and key in the feature map. σ is a predefined decay factor controlling the spatial sensitivity. Then we define a binary mask to represent the region relationship of the query and key:

$$R[q,k] = \mathbb{I}\left(\forall m, q \notin \mathcal{I}_m \vee k \notin \mathcal{I}_m\right) \tag{8}$$

where $\mathcal{I}_m$ represents the set of token indices belonging to the m-th region. Lastly, we get the gaussian mask and use it for the modulation of the attention map:

$$A' = \text{Softmax}(\frac{QK^\top + (R \cdot Bias)}{\sqrt{d}}) \tag{9}$$

The advantage of this method is that it can retain the necessary context information while performing semantic decoupling. The greater the distance, the greater the attenuation, thus avoiding excessive cross-region interactions. We also tried a variety of different methods to implement the receptive field constraint, see Supplementary Material for details.

3.5 High-frequency Compensation

We decompose the feature map into low-frequency and high-frequency components to better observe the evolution of image details during the generation process. In the early stages, the model primarily generates low-frequency components, which capture the broad structure and general shape of the image. As the generation process progresses, the proportion of high-frequency information gradually increases, focusing on finer details and textures. This behavior is illustrated in Figure 7, where the model begins with low-frequency information and progressively shifts towards high-frequency optimization, enhancing the image's texture and details. To this end, we propose controlling the intensity of the high-frequency components in the feature maps during the generation process, thereby obtaining high-fidelity images with more richly detailed content.

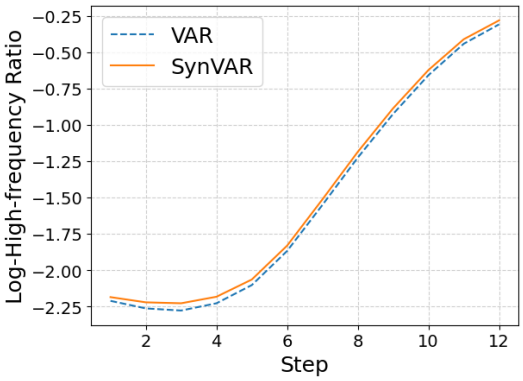

Fig. 7: High-frequency ratio. As generation progresses, high-frequency components increase. A small intensity coefficient is used due to its sensitivity.

We begin by defining a high-pass filter with a linear gain controlled by the enhancing ratio δ, which allows us to adjust the strength of the high-frequency components:

$$D(u, v) = \sqrt{(u - u_0)^2 + (v - v_0)^2} \tag{10}$$

$$HF(u, v) = 1 + \delta \cdot D(u, v) \tag{11}$$

where (u_0, v_0) denotes the center of the frequency spectrum, typically $u_0 = W/2$ and $v_0 = H/2$ for an image of size $H \times W$. We then perform a Fourier transform on the input of each transformer layer to convert it to the frequency domain and enhance its high-frequency components through the above high-pass filter. Finally, we perform an inverse Fourier transform on the filtered signal to restore it back to the spatial domain:

$$\tilde{\mathbf{F}} = \mathcal{F}^{-1}\left[\mathcal{F}[\mathbf{F}] \cdot HF\right] \tag{12}$$

where $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ denotes the input feature map. $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ denote the 2D Fourier and inverse Fourier transform, respectively. $HF \in \mathbb{R}^{H \times W}$ represents a high-pass filter mask in the frequency domain. $\tilde{\mathbf{F}}$ is the filtered feature map after enhancing high-frequency components. In this way, we are able to significantly enhance the details and texture information of the image, ensuring the richness and fineness of the details in the later stages of generation.

The pseudocode for SynVAR is shown in Algorithm 1.

Algorithm 1 SynVAR Algorithm Overview

Input: input prompt p, global regions $\{h^i, w^i\}_{i=1}^N$, individual concepts $\{c^i\}_{i=1}^N$, Transformer block input feature $\mathbf{F}$.

Hyperparameters: selected steps S_{steps}, resolution of scales $\{h_s, w_s\}_{i=1}^S$, decay factor σ, enhancing ratio δ.

Model: linear projections ℓ_Q, ℓ_K, ℓ_V , Text Encoder TE$(\cdot)$. $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ is the 2D Fourier and inverse Fourier transform.

$\{y^i\}_{i=1}^N \leftarrow TE(\{c^i\}_{i=1}^N)$
for $s = 1, \cdots, S$ **do**
  $HF \leftarrow 1 + \delta \cdot \sqrt{(u-u_0)^2 + (v-v_0)^2}$ ▷ **High-frequency Compensation**
  $\tilde{\mathbf{F}} = \mathcal{F}^{-1}[\mathcal{F}[\mathbf{F}] \cdot HF]$
  $r_{s-1} \leftarrow \tilde{\mathbf{F}}$
  if $s \in S_{steps}$ **then**
    $Bias \leftarrow \exp\left(-\frac{||p_q - q_k||_2^2}{\sigma^2}\right)$ ▷ **Receptive Field Constraint** in self-attention
    $R \leftarrow \mathbb{I}\left(\forall m, q \notin \mathcal{I}_m \vee k \notin \mathcal{I}_m\right)$
    $r_{s-1} \leftarrow \text{Softmax}(\frac{QK^\top + (R \cdot Bias)}{\sqrt{d}})V$
    for $i = 1, \cdots, N$ **do** ▷ **Global Guidance** in cross-attentions
      $Q^i \leftarrow \ell_Q(r_{s-1})$; $K^i \leftarrow \ell_K(y^i)$; $V^i \leftarrow \ell_V(y^i)$
      $r_{s-1}^i \leftarrow \text{Softmax}\left(\frac{Q^i K^{iT}}{\sqrt{d}}\right) V^i$
    $r_{s-1}^{cat} \leftarrow \text{Crop}(r_{s-1}^i(h^i, w^i))$
    $f_s \leftarrow f_{s-1} + \phi\left(r_{s-1}^{cat}, (h_s, w_s)\right)$
  else
    $f_s \leftarrow f_{s-1} + \phi\left(r_{s-1}, (h_s, w_s)\right)$
return Decoder(f_s)

4 Experiments

4.1 Experiment Setting

Implementation Details. For the influence step of global guidance and receptive field constraints, we choose to operate at step=2. The decay coefficient σ of the receptive field constraint is set to 0.5. The intensity control coefficient δ for high-frequency compensation is set to 0.01. Other parameters are consistent with the compatible base model. For large-scale quantitative evaluations, we leverage MLLM for automated global region and concept division, see Appendix

Method	Geneval					T2I-CompBench					
	Two_object ↑	*Counting* ↑	*Position* ↑	*Attribute_Binding* ↑	*Overall* ↑	*Color* ↑	*Shape* ↑	*Texture* ↑	*Spatial* ↑	*Complex* ↑	*Overall* ↑
Infinity [CVPR2025]	79.80	58.13	26.00	58.00	55.48	74.99	48.59	62.90	24.13	38.89	49.90
+Densediffusion	75.00	19.69	21.75	52.50	42.23	72.42	48.39	62.64	24.78	38.55	49.36
+RAG-Diffusion	70.71	40.31	30.00	44.75	46.44	59.86	34.85	56.11	27.05	33.06	42.19
+Ours	91.92	58.25	63.00	70.00	70.79	75.21	51.85	66.23	46.00	39.02	55.66
Ours vs Infinity	+12.12	+0.12	+37.00	+12.00	+15.31	+0.22	+3.26	+3.33	+21.87	+0.13	+5.76
Switti [CVPR2025]	73.99	48.12	14.25	28.00	41.09	74.83	52.00	64.42	19.09	37.72	49.61
+Densediffusion	65.91	18.44	19.00	27.75	32.77	71.25	49.60	61.15	17.49	36.68	47.23
+RAG-Diffusoin	36.62	3.44	21.00	18.00	19.76	55.92	37.81	53.65	12.42	31.33	38.22
+Ours	75.25	49.06	16.25	40.50	45.27	78.54	54.67	68.10	22.14	38.77	52.44
Ours vs Switti	+1.26	+0.94	+2.00	+12.50	+4.18	+3.71	+2.67	+3.68	+3.05	+1.05	+2.83

Table 1: Quantitative results on Geneval and T2I-CompBench benchmark.

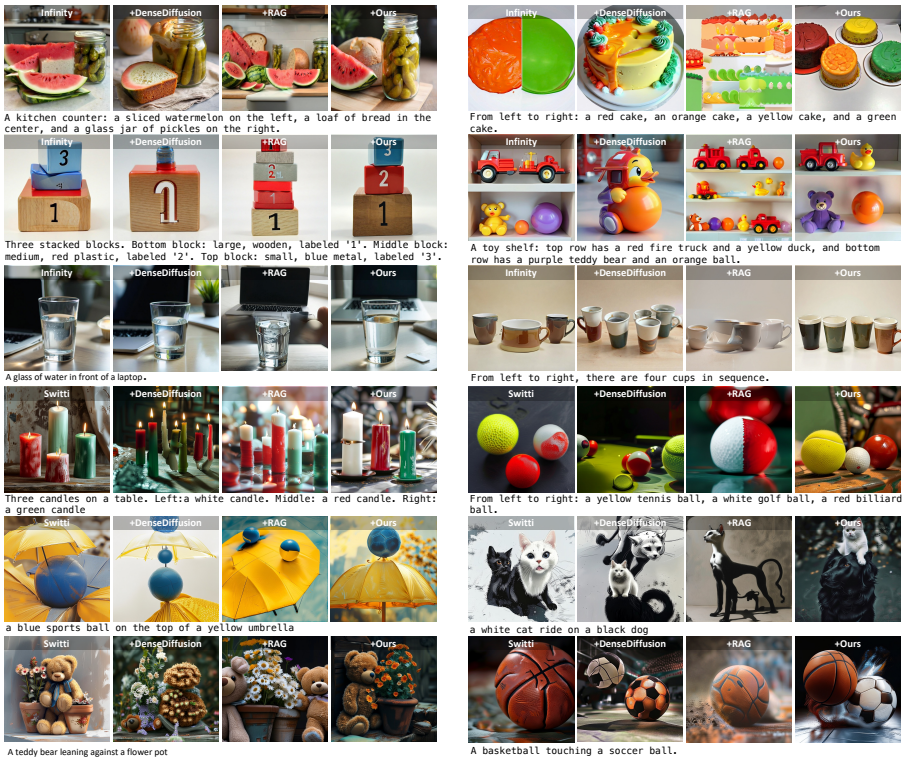

Fig. 8: Qualitative comparison of different methods for adapting VAR-based text-to-image models.

for usage (Notably, our framework is MLLM-free. MLLMs are only employed for large-scale evaluation and user convenience, with additional experiments detailed in the Appendix). All experiments are implemented on a single A6000. We also provide runnable code in the Appendix for understanding and reproduction.

Baselines and Benchmark. Since SynVAR is designed as a plug-and-play solution, we integrate it with existing VAR-base methods infinity and Switti for experiments. We apply consistent parameter settings across all models to ensure

fair comparison. For the generation quality, we conduct experimental evaluation on two widely used benchmark Geneval [15] and T2I-CompBench [20].

4.2 Main Results

Quantitative Comparison. As shown in Table 1, existing enhancement methods generally have a negative impact on the performance of the original model, suggesting their limited effectiveness in addressing the inherent weaknesses of VAR-based methods under complex compositional prompts. In contrast, integrating SynVAR leads to consistent and significant improvements. On the Geneval dataset, SynVAR boosts the overall performance of the Infinity model by 27.6% and that of Switti by 10.2%. Notably, SynVAR achieves clear gains on both spatial metrics (e.g., Position in Geneval and Spatial in T2I-CompBench) and semantic metrics (e.g., Attribute_Binding in Geneval and Shape/Texture in T2I-CompBench), demonstrating its capability to simultaneously reduce spatial errors and semantic mismatches.

Qualitative Comparison. As illustrated in Figure 8, existing approaches fail to resolve spatial dislocation or semantic confusion when generating images from complex prompts. For instance, objects often appear in incorrect positions or exhibit inconsistent appearances. By contrast, SynVAR achieves more coherent spatial arrangements and preserves semantic fidelity more effectively. It not only reduces spatial errors and semantic confusion but also produces visually pleasing results with sharper structures and finer textures. Moreover, the generated global scene appears more harmonious and contextually aligned with the input prompt, indicating that SynVAR successfully captures both local detail and holistic scene layout. In addition, we have demonstrated more practical and interesting usages in the Appendix, such as overlapping, multi-objective fine-grained interaction, and the utilization of irregular masks.

User Study. To further validate the effectiveness of SynVAR from a human perspective, we conducted a user study. We randomly selected 20 prompts from the Geneval and T2Icombench to assess the aesthetic quality and text-image alignment (including both semantic and spatial position) of the generated images. During the evaluation, we randomly displayed an image pair and its corresponding text description, and users selected the image that best matched the prompt based on their judgment. The results in Figure 9 show that

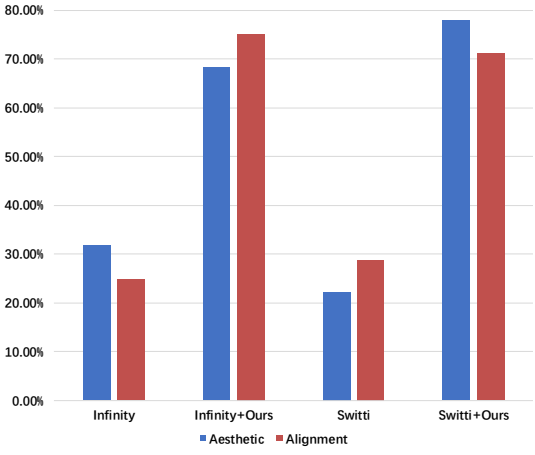

Fig. 9: User study on aesthetics and text-image alignment. Our approach shows a significant improvement.

more than 68% of users chose the image enhanced by SynVAR in terms of both aesthetics and alignment, indicating that SynVAR not only improves objective performance metrics but also aligns more closely with human visual perception and interpretive judgment.

4.3 Ablation Study

Methods	Two_object	Counting	Position	Attribute_Binding	Overall
w/o Global Guidance	74.49	66.25	23.75	49.00	53.37
w/o Receptive Field Constraint	80.30	21.25	36.25	58.25	49.01
w/o High-Frequency Compensation	86.38	55.81	59.50	62.75	66.11
SynVAR	**91.92**	**58.25**	**63.00**	**70.00**	**70.79**

Table 2: Ablation performance of SynVAR on the Geneval.

Effectiveness of Global Guidance. As shown in Table 2,removing the global guidance mechanism reduces the spatial position accuracy by 62.3%, and also causes a decrease in other indicators. This means that global guidance effectively improves the correctness of the spatial structure by explicitly injecting spatial priors, laying a solid foundation for the subsequent generation process, thereby improving the image indicators in all aspects. Figure 10(a) shows the visual effect of global guidance in ensuring the correct spatial structure.

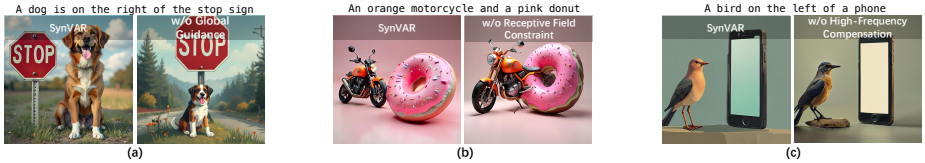

Fig. 10: Qualitative analysis of global guidance, receptive field constraint and high-frequency compensation.

Effectiveness of Receptive Field Constraint. The receptive field constraint helps reduce excessive interactions between regions, preserving object generation independence. Table 2 shows that it significantly lowers interference, allowing each region to generate its object independently, improving the counting metric. Figure 10(b) visually demonstrates this effect, where regions decouple their target objects while preserving semantic integrity.

An ablation study on the decay coefficient σ in the receptive field constraint reveals a non-linear relationship with performance. From the data in Table 3, as σ increases, performance improves initially but declines beyond a point. This

suggests that excessive constraint hinders performance, while a moderate σ balances region independence and necessary information transfer, enhancing output quality. This is shown in Figure 11(a), where a balanced decay coefficient produces better results.

Methods	Two_object	Counting	Position	Attribute_Binding	Overall
$\sigma = 0.1$	80.56	46.25	34.50	57.75	54.76
$\sigma = 0.3$	88.38	54.69	53.25	65.50	65.46
$\sigma = 0.5$	**91.92**	**58.25**	**63.00**	**70.00**	**70.79**
$\sigma = 0.7$	90.66	57.50	60.25	67.50	68.98
$\sigma = 1.0$	90.66	55.00	57.25	67.00	67.48

Table 3: Ablation experiment of decay coefficient σ in receptive field constraint.

Methods	Two_object	Counting	Position	Attribute_Binding	Overall
$\delta = 0.002$	90.40	56.56	57.00	65.00	67.24
$\delta = 0.005$	89.14	55.00	60.50	69.25	68.47
$\delta = 0.01$	91.92	**58.25**	63.00	**70.00**	**70.79**
$\delta = 0.02$	**92.17**	54.37	**65.50**	67.00	69.76
$\delta = 0.03$	90.66	51.25	63.75	66.50	68.04

Table 4: Ablation experiment of intensity control coefficient δ in High-frequency Compensation.

Fig. 11: Qualitative analysis of decay coefficient σ and intensity control coefficient δ.

Effectiveness of High-frequency Compensation. High-frequency compensation enhances fine-grained details like textures and intricate structures. While its impact on overall metrics is modest, it noticeably improves image quality. Figure 10(c) highlights its effect, refining details, especially in complex areas. We also conducted a user study to evaluate our advantages in terms of image quality (e.g., Aesthetics), see the supp for details.

An analysis of the intensity control coefficient δ reveals that while performance remains stable with minor fluctuations, excessive enhancement can introduce artifacts, as shown in Table 4 and Figure 11(b). We chose δ=0.01 as an optimal value, balancing detail enhancement with image integrity.

Selection of Early Steps for Global Guidance and Receptive Field Constraints. We conduct ablation studies to verify the optimal stage for applying global guidance and receptive field constraints during the generation sequence. As shown in Table 5, extending these operations to multiple iterations leads to performance degradation, but the overall trend remained stable.

We also visualize the attention regions during the VAR generation process, as illustrate in Figure 12. It can be observed that the attention focus stabilizes at very low resolutions, with minimal changes in subsequent steps. This supports the findings in Table 5, indicating that injecting structured priors at the critical decision window (step = 2) enables accurate correction of error sources while maintaining the coherence of the VAR generation.

Step	Two_object	Counting	Position	Attribute_Binding	Overall
1	83.33	57.50	45.00	66.00	63.43
2	91.92	**58.25**	63.00	**70.00**	**70.79**
3	81.06	52.19	36.25	61.25	57.69
1, 2	89.14	54.06	68.25	68.00	69.86
1, 3	86.11	50.94	55.00	67.50	64.89
1, 2, 3	92.68	47.81	74.50	63.25	69.56
1, 2, 3, 4	**93.69**	39.38	75.00	60.00	67.02
1, 2, 3, 4, 5	91.41	32.81	**75.25**	58.50	64.49

Table 5: Quantitative comparison of global guidance and receptive field constraints at different steps.

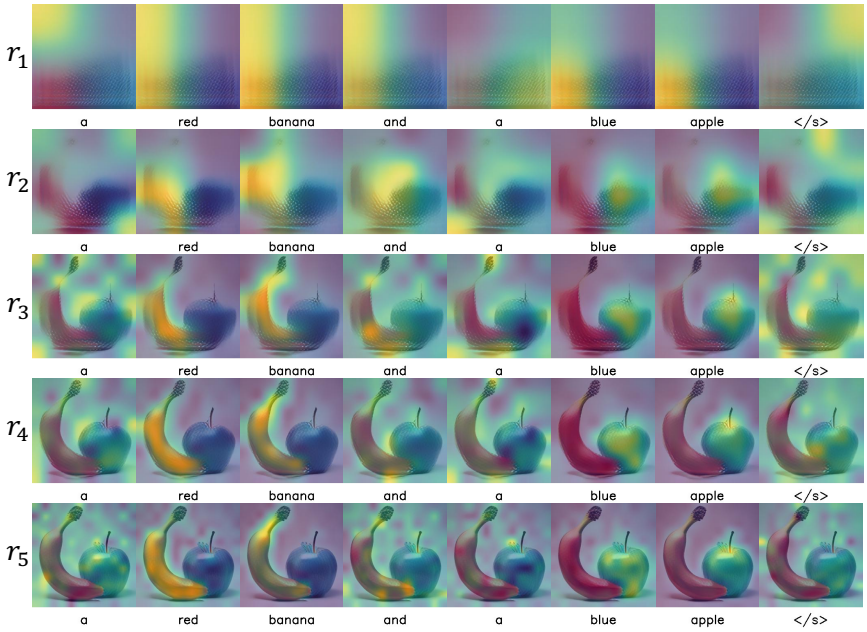

Fig. 12: Visual analysis of the attention area during the generation process.

Inference Time Cost. As shown in Table 6, we tested the single-image inference time consumption of different components. It can be seen that the inference overhead of Global and Receptive Field Constraints is almost negligible compared to Vanilla VAR. High-Frequency Compensation, due to the frequency domain conversion involved, incurs additional time overhead, but this is traded off for improved aesthetic quality. Since our method does not involve any structural changes, SynVAR can seamlessly integrate with existing

VAR acceleration schemes (such as Fastvar [16]), thereby ensuring image quality improvements while avoiding increasing inference overhead. We have also provided performance metrics compatible with Fastvar in the Appendix.

Methods	Inference Time Cost (seconds/image)
Vanilla VAR	2.60
w/ Global Guidance	2.61
w/ Receptive Field Constraint	2.61
w/ High-Frequency Compensation	3.08
SynVAR	3.10
SynVAR + Fastvar	2.27

Table 6: Single-image inference time consumption of different SynVAR components.

5 Conclusion

In this paper, we conduct an in-depth analysis of the failure mechanisms of VAR in complex scene generation and propose the first training-free enhancement framework for VAR, dubbed SynVAR. SynVAR implements a collaborative correction mechanism for spatial and semantic control through three key components: global guidance, receptive field constraints, and high-frequency compensation. Global guidance and receptive field constraints work during the early stages of generation, utilizing spatial priors and gaussian masks to capture precise spatial and semantic information, effectively preventing the propagation and accumulation of errors. High-frequency compensation enhances high-frequency features, further improving the detail and generating high-fidelity images. Extensive experimental results demonstrate that SynVAR seamlessly integrates with various VAR-based text-to-image models and significantly improves generation performance. In the future, we will try to extend SynVAR to more types of generative models, hoping to provide more valuable insights to the community.

6 Limitation and Future Work

Our framework mainly has two limitations. First, the capability of our framework to better align the semantic and spatial relations is limited by the prior of the VAR foundation model. If the semantic concepts are missed in the prior or the spatial relation deviation is unreasonable, our correction mechanism is deficient at such imprinted limitations of the foundation model. Second, for the best performance, the optimal early step of global guidance and receptive field constraints through experiments and empirical settings. In the future, we expect to develop a lightweight fine-tuning and auto-selection strategies to edit the model prior and automatically select the optimal step for each single image.

Acknowledgment

This work was supported by Natural Science Foundation of Jiangsu Province: BK20241198, the Gusu Innovation and Entrepreneur Leading Talents: No. ZXL20 24362 and Natural Science Foundation of China: No. 62406135, Nanjing University-China Mobile Communications Group Co. Ltd. Joint Institute.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
5. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
6. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
7. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)
8. Chen, Z., Gao, R., Xiang, T.Z., Lin, F.: Diffusion model for camouflaged object detection. In: ECAI 2023, pp. 445–452. IOS Press (2023)
9. Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., Tai, Y.: Region-aware text-to-image generation via hard binding and soft refinement. arXiv preprint arXiv:2411.06558 (2024)
10. Chen, Z., Zhu, J., Chen, X., Zhang, J., Chen, J., Zeng, Z., Zhang, W., Wang, C., Yang, J., Tai, Y.: L2p: Unlocking latent potential for pixel generation. arXiv preprint arXiv:2605.12013 (2026)
11. Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space. arXiv preprint arXiv:2511.18822 (2025)
12. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024)
13. Du, N., Chen, Z., Chen, Z., Gao, S., Chen, X., Jiang, Z., Yang, J., Tai, Y.: Textcrafter: Accurately rendering multiple texts in complex visual scenes. arXiv preprint arXiv:2503.23461 (2025)
14. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
15. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
16. Guo, H., Li, Y., Zhang, T., Wang, J., Dai, T., Xia, S.T., Benini, L.: Fastvar: Linear visual autoregressive modeling via cached token pruning. arXiv preprint arXiv:2503.23367 (2025)

17. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
18. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. arXiv preprint arXiv:2412.04431 (2024)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
21. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)
22. Liu, D., Zhao, S., Zhuo, L., Lin, W., Qiao, Y., Li, H., Gao, P.: Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. arXiv preprint arXiv:2408.02657 (2024)
23. Ma, X., Zhou, M., Liang, T., Bai, Y., Zhao, T., Chen, H., Jin, Y.: Star: Scale-wise text-to-image generation via auto-regressive representations. arXiv preprint arXiv:2406.10797 (2024)
24. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
26. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
27. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
28. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
31. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
32. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
33. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)

34. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
35. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. arXiv preprint arXiv:2412.01819 (2024)
36. Wang, Y., Guo, L., Li, Z., Huang, J., Wang, P., Wen, B., Wang, J.: Training-free text-guided image editing with visual autoregressive model. arXiv preprint arXiv:2503.23897 (2025)
37. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)
38. Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
39. Yang, S., Chen, Z., Lin, Y., Chen, X., Cai, G., Yu, H., Wu, P., Yang, Q.: A survey on vision-language models for multimodal federated learning tasks. Authorea Preprints (2025)
40. Yang, Z., Jiang, Z., Li, X., Zhou, H., Dong, J., Zhang, H., Du, Y.: D 4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In: European Conference on Computer Vision. pp. 36–52. Springer (2024)
41. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16702–16711 (2025)
42. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
43. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: LUVE : Latent-cascaded ultra-high-resolution video generation with dual frequency experts. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=FjZF2l8ZZF
44. Zhao, C., Dong, C., Cai, W., Wang, Y.: Learning a physical-aware diffusion model based on transformer for underwater image enhancement. IEEE Transactions on Geoscience and Remote Sensing (2026)
45. Zhao, C., Xu, Y., Chen, Z., Gu, E., Zhang, K., Liu, X., Yang, J., Tai, Y.: From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–18 (2026)
46. Zhou, D., Li, M., Yang, Z., Yang, Y.: Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. arXiv preprint arXiv:2503.12885 (2025)
47. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6818–6828 (2024)
48. Zhou, D., Xie, J., Yang, Z., Yang, Y.: 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. arXiv preprint arXiv:2410.12669 (2024)
49. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327 (2025)