Exploring Position Encoding Mechanism in Diffusion U-Net for Training-free High-resolution Image Generation

Feng Zhou[1*], Pu Cao[1,2*], Yiyang Ma[1], Lu Yang[1], Yonghao Dang[1], Jianqin Yin[1†]

[1]Beijing University of Posts and Telecommunications
[2]Beijing Hydrogen Intelligence Technology Co. Ltd.
{zhoufeng, caopu, yym2024, soeaver, dyh2018, jqyin}@bupt.edu.cn

Abstract

Denoising higher-resolution latents using a pre-trained U-Net often results in repetitive and disordered image patterns. In this work, we are motivated to reveal the intrinsic cause of such pattern disruption in high-resolution image generation. Through theoretical analysis and empirical studies, we reveal that the pre-trained U-Net fails to provide sufficient positional information for tokens at high-resolution. Specifically, 1) zero-padding serves as a critical mechanism for position encoding but lacks robustness across varying resolutions; and 2) tokens located farther from the feature map boundaries have increasing difficulty acquiring positional awareness, leading to pattern disruptions. Inspired by these findings, we propose a novel training-free approach for high-resolution generation, introducing a Progressive Boundary Complement (PBC) method. It creates dynamic virtual image boundaries inside the feature map to supplement position information at high resolution, enabling high-quality and rich-content high-resolution image synthesis. Extensive experiments show that our method significantly improves high-resolution image synthesis in terms of visual quality and content richness, achieving state-of-the-art performance.

1 Introduction

Text-to-image generation has gained significant attention due to its wide range of applications in real-world scenarios, such as digital content creation (Poole et al. 2022; Cao et al. 2025; Wang et al. 2025) and personalized media generation (Cao et al. 2024b; Ruiz et al. 2023; Gal et al. 2022). Among various generative paradigms, Latent Diffusion Models (LDMs) (Rombach et al. 2022) have emerged as a popular and powerful approach, showing impressive results. Traditional LDMs, like Stable Diffusion (SD) (Rombach et al. 2022), employ U-Net for latent-space denoising with a relatively fixed training resolution (*e.g.*, 64×64 in *SD-2.1*). Recent research (He et al. 2023; Jin et al. 2023) shows that the pre-trained U-Net is highly sensitive to the latent resolutions; higher-resolution latent inputs often result in repetitive patterns and disordered layouts, as shown in Figure 1. Intuitively and fundamentally, this phenomenon can be attributed to the wrong position information encoded in tokens, breaking the positional relation in images.

Figure 1: Directly generating high-resolution images using a pre-trained Latent Diffusion Model results in disordered (*left*) and repetitive (*right*) patterns.

Hence, in this study, we are driven to explore why token's representations are disrupted by resolution scaling, from the position encoding perspective. We first look into how U-Net constructs the position information across elements or tokens. While convolution operation inherently encodes relative position information, absolute position information is expected to originate solely from the zero-padding mechanism in the convolution layers (Islam, Jia, and Bruce 2020). Note that position embedding is not used in the attention operations in pre-trained U-Net. Through initial empirical observations, we observe that zero-padding is the crucial mechanism governing spatial arrangement, yet it lacks robustness across resolutions. Furthermore, to assess this phenomenon, we design comparative quantitative experiments to verify the significant impact of zero-padding on positional information—and how this information deteriorates at higher resolutions, where tokens far from the boundaries fail to capture sufficient positional information.

Most of the existing methods, based on position information absence perspective, can be regarded as broadly designed to address this challenge in high-resolution generation by aligning position information across different resolutions. For instance, dilated convolution-based method (He et al. 2023; Huang et al. 2025) leverages dilating the convolution kernel to accelerate the propagation of position information in high-resolution generation. Multi-stage lifting methods (Cao et al. 2024a; Qiu et al. 2024; Kim et al. 2024),

*These authors contributed equally.
†Corresponding Author

which progressively increase image resolution by multiple stages, maintain position information by lifting resolution from the original level. However, these approaches have notable limitations: 1) they adopt an empirically designed strategy to maintaining positional consistency at higher resolutions and often introduce complex architectures. 2) this rigid alignment constrains content diversity across different resolutions, making it difficult to generate more enriched and complex elements at high resolutions.

In this work, we propose a simple but effective training-free high-resolution image generation method, simplifying the task by directly addressing it from the perspective of position information completion. We introduce a Progressive Boundary Complement (PBC) method, designed to enhance position information from the boundary to the central regions. To achieve this, we construct virtual boundaries within the feature map, a specialized form of unidirectional padding, to facilitate position information while maintaining feature aggregation. By hierarchically placing these virtual boundaries within the feature map, our method not only corrects position encoding inconsistencies in high-resolution features, thereby mitigating spatial distortions, but also effectively extends the perceived image boundaries. This enables the generation of high-resolution images with richer details and greater content diversity. Our extensive experiments demonstrate its effectiveness.

To sum up, our contributions are as follows:

1. We thoroughly investigate the mechanism of positional encoding in diffusion U-Net and arrive at the following conclusion: absolute positional information originates from zero-padding, which governs spatial arrangement. Based on this finding, we identify that at high resolutions, positional encoding becomes inadequate — particularly for tokens in the central region — leading to spatial disorder.

2. To address this issue, we propose Progressive Boundary Complement (PBC), a training-free method that enhances position information by introducing hierarchical virtual boundaries within the feature map. It extends the perceived image boundaries, enabling the synthesis of high-resolution images with richer content.

3. Extensive quantitative and qualitative experiments demonstrate the superiority of our method in generating high-resolution images with enhanced content richness.

2 Related Works

High-Resolution Image Synthesis with Pre-trained Diffusion Models. Diffusion models (Rombach et al. 2022; Peebles and Xie 2023) have significantly propelled the development of text-to-image generation. However, as these models are trained at relatively fixed resolutions, they struggle to effectively generate higher-resolution images. Several methods address this limitation by fine-tuning pre-trained models, converting them into dedicated high-resolution image generators (Guo et al. 2025). In contrast, many approaches tackle this issue without additional training (Hwang, Park, and Jo 2024; Huang et al. 2025). For instance, certain studies have explored adapting model architectures specifically for high-resolution generation. AttnEntro (Jin et al. 2023) adjusts the attention scaling factor based on attention entropy considerations during high-resolution generation. ScaleCrafter (He et al. 2023) increases the receptive field of convolutional layers to facilitate high-resolution image synthesis. HiDiffusion (Zhang et al. 2025) dynamically modifies feature map dimensions to mitigate object duplication. Another category of training-free methods adopts a multi-stage resolution-lifting paradigm, initially generating images at the base resolution and progressively upsampling them to the desired higher resolutions (Cao et al. 2024a; Kim et al. 2024; Wu et al. 2024; Yang et al. 2024a; Zhang, Li, and Zhang 2024; Qiu et al. 2024). Additionally, some methods emphasize refining text prompts at the patch level to enhance high-resolution image generation (Bar-Tal et al. 2023; Lin et al. 2025; Liu et al. 2024).

Mechanisms of Text-to-Image Diffusion Models. Although text-to-image diffusion models have achieved remarkable success across various applications, their internal mechanisms have not yet been fully explored (Si et al. 2024). For instance, FreeU (Si et al. 2024) analyzes the roles played by the U-Net's backbone and skip connections from a frequency-domain perspective. Additionally, several studies have examined properties of text encoders and cross-attention mechanisms, shedding light on the interactions between textual prompts and the diffusion process (Toker et al. 2024; Yang et al. 2024b; Yi et al. 2024). In this study, we aim to further investigate the underlying mechanisms from the perspective of positional information construction of U-Net.

Position Encoding in Visual Neural Networks. In transformer-based visual neural networks (Dosovitskiy 2020; Oquab et al. 2023; Zhou, Yin, and Li 2024), explicit positional embeddings are incorporated into the attention mechanism, providing a clear source of positional information. In contrast, traditional convolution-based networks typically lack explicit positional embeddings. Although Islam, Jia, and Bruce demonstrated that convolutional networks inherently possess positional information closely related to the zero-padding mechanism, their work did not uncover how this positional information propagates or explicitly manifests within generative models, such as diffusion models.

3 Analysis of the Position Encoding Mechanism

In this section, we examine how the U-Net architecture establishes position encoding, specifically identifying which components and how they contribute to position information. Building on these analyses, we aim to deduce the cause of the disorder pattern in high-resolution image generation. All experiments in this section are conducted with *SD-2.1*.

Zero-padding is Crucial for Spatial Arrangement

In traditional U-Net architectures, such as the U-Net in *SD-1.x, SD-2.x*, and *SD-XL*, no position-embedding is utilized in both self- and cross-attention layers, suggesting that position information is solely introduced in convolution layers. Convolution operation models relations between local

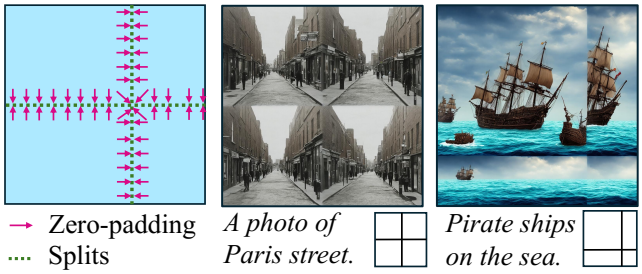

Figure 2: Trench-style Zero-padding Technique. The left-side graph diagram illustrates the process of applying bi-directional zero-padding to the feature map in the convolution operation. The two images on the right show 1024×1024 resolution outputs using this technique, with the corresponding split sketches displayed below.

Figure 3: Padding Type Analysis. We evaluate the effect of different padding types.

tokens, providing relative position information in features. Meanwhile, the absolute position information is introduced to the edge tokens through the padding mechanism and is propagated by other layers that enable interactions between tokens (Islam, Jia, and Bruce 2020). Since the convolution operation behavior is similar across varied resolutions, we hypothesize that padding plays a crucial role in spatial arrangement.

To validate this, we first design a trench-style zero-padding technique, which divides the latent feature into multiple convolution regions, as illustrated on the left side of Figure 2. Specifically, we apply bi-directional zero-padding to the trenches (splits) of the feature map during convolution operations. The added zero-padding effectively partitions the generated images. Regions with downward padding generate elements such as oceans and land, whereas regions with upward padding produce elements such as the sky, while regions with overly low resolution fail to produce image elements. This demonstrates that zero-padding governs spatial arrangement in a pre-trained U-Net.

To further investigate the role of zero-padding, we replaced it with several common padding modes, including reflect padding, replicate padding, and circular padding. The results are shown in Figure 3, where only the zero-padded images exhibit a reasonable layout.

Figure 4: Position Information Correction. We applied unidirectional zero-padding to the central region of the feature map to facilitate faster propagation of position information. The images were generated at a resolution of 1024×1024, with the central region measuring 512×512. Prompt: *A photo of a teddy bear riding a bike in Times Square.*

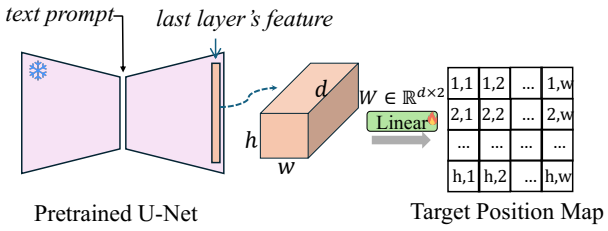

Figure 5: Position Information Quantification. We extract the feature from the last layer of U-Net and map it to a target position map with a trainable linear layer. The loss of the linear year reflects how much position information the feature contains.

Insufficient Position Encoding in High-resolution Generation

Building on the above, a key question arises: why is the element arrangement is corrupted in high-resolution generation? We speculate that it is because the position encoding becomes inadequate when resolution increases in a pre-trained U-Net, especially for the tokens far from zero-paddings at feature boundaries.

To validate this hypothesis, we conduct an experiment aimed at improving positional information in central regions by applying unidirectional zero-padding specifically at the center, as shown in Figure 4. This method successfully resolves the chaotic arrangement in the central region, correcting issues such as disorganized grounds and buildings, as well as the repetitive generation of bears.

To confirm the above conclusions, we design an artful quantitative experiment aimed at exploring how much position information U-Net can encode under different circumstances, as shown in Figure 5. We extract the last layer's feature map from the U-Net at the first denoising time-step and map it to a target position map of the same size, containing only position coordinates. Only a simple trainable linear layer is used for this mapping, aimed at minimizing its fitting

Mode	Resolution	Loss ↓
Random Feature	512^2	0.077
Zero Padding (default)	512^2	**0.006**
Circular Padding	512^2	0.033
Zero Padding (default)	1024^2	0.023
Dilated Convolution	1024^2	0.023
PBC (ours)	1024^2	**0.017**
Zero Padding (default)	Central 512^2 in 1024^2	0.039
Dilated Convolution	Central 512^2 in 1024^2	0.025
PBC (ours)	Central 512^2 in 1024^2	**0.010**

Table 1: Comparison of padding and convolution methods across resolutions.

capability. After training the linear layer to convergence, the final loss reflects the extent of position information encoded in the feature. Greater position information results in a lower loss. The details and thorough ablation studies of the experiment strategies can be found in the Appendix.

The loss values under different circumstances are written in Table 1. The top three rows compare random features, zero-padding features (default setting), and circular-padding features, confirming that U-Net encodes position information and zero-padding greatly contributes to it. The middle three rows present the loss values for high-resolution generation (1024^2). In the bottom three rows, we separate the central 512^2 region in the 1024^2 resolution to conduct the experiment for a fair comparison with the original resolution generation (top three rows). This confirms that the encoded position information becomes less in the central region in high-resolution generation, validating our hypothesis. Dilated convolutions (He et al. 2023) alleviate this issue to some extent since they expand the receptive field, making the tokens in the central regions "closer" to the boundaries from the receptive field perspective. Our method (described in the next section) achieves the best performance.

4 Methods

Preliminaries

Latent Diffusion Model (LDM). A pre-trained LDM enables the transfer of standard Gaussian noise z_T into an image latent z_0 aligned with a pre-given text prompt through reverse denoising processes. Here, the latent space $\mathbf{Z} \subseteq \mathbb{R}^{h\times w\times c}$ is characterized by a pre-trained Vector Quantized variational autoencoder (VQ-VAE). One can decode the image latent z_0 into a real image x by the VAE decoder $x = \mathcal{D}(z_0)$.

The reverse denoising process gradually obtains less noisy latent z_{t-1} from the noisy input z_t at each timestep $t \in (1, T)$:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \quad (1)$$

where μ_θ and Σ_θ are determined through a noise prediction network $\epsilon_\theta(z_t, t, C)$ with learnable parameter θ. C is the text prompt embedding. In traditional LDM architecture, the denoising network is realized using a time-conditional U-Net.

Progressive Boundary Complement (PBC)

One straightforward approach to enhance the position information is to insert additional real boundaries inside the feature map, as we discussed in Section 3. In that experiment, we added a unidirectional zero-padding square at the central region of the feature map. However, though it solves the disorder in the central region, this approach will cause the visible split shown at the square edges, we define it as the split effect, as shown in the "After Correction" part of Figure 4.

Virtual Boundary. To address the issue of discontinuity, we propose the Virtual Boundary technique, which aims to construct internal fake boundaries within the feature maps. This approach effectively mitigates the discontinuities introduced by zero-padding while preserving the propagation of positional information.

The cause of the split effect can be easily inferred. Tokens located just inside the zero-padding square cannot gather information from the tokens right outside it during convolution operations, leading to incorrect feature aggregation at the edges of the square. Thus, we proposed a novel padding trick, named valued-padding, designed to replace zero-padding with a value that contains features from the outside token to enhance the feature aggregation. On the other hand, we expect it to have the zero-padding's property to provide the position information. Therefore, we set the padding value as a proportional of the outside token's value, formed as:

$$p = F * \lambda, \quad \lambda \in [0, 1] \quad (2)$$

where p is the padding value, F is the feature of the outside token, and λ is the ratio. Ideally, a lower λ results in the virtual boundary providing stronger position information, as it more closely resembles zero-padding.

Random Boundary Perturbation. While valued-padding helps alleviate the split effect, it does not completely eliminate discontinuities at boundary edges. To further smooth transitions, we introduce random perturbations to the positions of the virtual boundaries within a small range.

$$\tilde{l} = l + \delta, \quad \delta \sim \mathrm{U}(-r, r) \quad (3)$$

where l is the distance between boundary center to the real image boundary. $\tilde{l}$ is the perturbed distance, r denotes the perturbation range.

Hierarchical Virtual Boundary Placement. As discussed in Section 3, U-Net encodes stronger position information near the edges while it gradually diminishes towards the center. Therefore, to simulate this trend, we discretely place multiple virtual boundaries within the feature map, ensuring that those closer to the feature boundaries have a stronger ability to provide position information.

Specifically, we first define the padding ratio as $\lambda = \frac{2l}{s}$, where l represents the distance from the virtual boundary to the real boundary of the feature map, and s denotes the width or height of the feature map, depending on the orientation of the virtual boundary. This formulation ensures that virtual

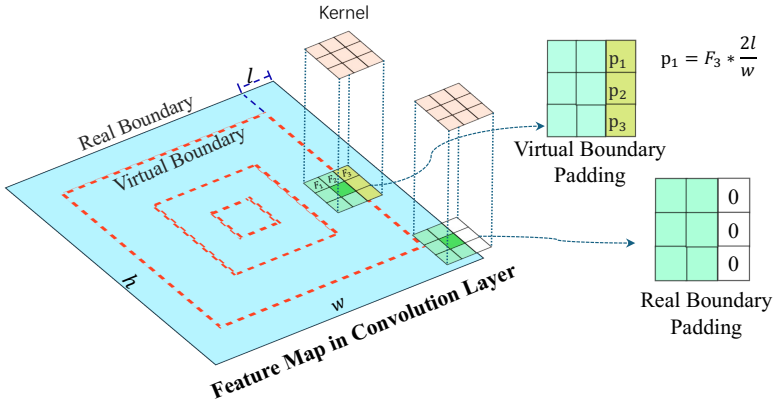

Figure 6: Progressive Boundary Complement (PBC) for training-free high-resolution image generation. PBC enhances the propagation of position information by inserting a series of designed virtual boundaries into the feature map.

boundaries closer to the feature edge have larger ratios, effectively capturing stronger positional information. For simplicity, we refer to the virtual boundary as V_λ, indexed by its ratio λ.

Secondly, we define a N discrete hierarchical virtual boundaries extending from the real boundary toward the image center, denoted as:

$$\mathcal{V}_N = \{V_{\lambda_n} \mid \lambda_n = \frac{n}{N+1}, n = 1, 2, \ldots, N\} \quad (4)$$

where λ_n represents the ratio for each virtual boundary.

The overall process of our Progressive Boundary Complement (PBC) is shown in Figure 6.

Image Generation with Enriched Content

Most high-resolution generation methods primarily aim to produce images that closely resemble those at the training resolution but on a larger scale, aligning to some extent with the objective of super-resolution. In other words, these methods essentially generate an upscaled version of the training-resolution image without introducing new semantic content. In contrast, our proposed Progressive Boundary Complement (PBC) method leverages virtual boundaries to expand the image canvas, facilitating the synthesis of high-resolution images with enriched and additional content.

Content Richness Metric. Since existing image generation evaluation metrics do not accurately quantify the richness of content in generated images, we introduce a novel metric, Content Richness (CR), to assess the diversity and richness of details in generated images. Specifically, given a generated image $x \in \mathbb{R}^{H\times W\times 3}$, we first partition it into k^2 ($k = 3$ in our experiment) equally sized image patches, denoted as $\{x_1, x_2, \ldots, x_{k^2}\}$, where each patch satisfies $x_i \in \mathbb{R}^{(H/k)\times(W/k)\times 3}$. Next, we compute the pairwise CLIP (Radford et al. 2021) similarity between all image patches. Let f_i denote the CLIP feature of patch x_i. The overall similarity score is obtained by computing the sum of pairwise cosine similarity across all patch pairs

$$S = \sum_{i=1}^{k^2} \sum_{j=1\ j\neq i}^{k^2} \frac{f_i \cdot f_j}{|f_i||f_j|}. \quad (5)$$

A lower similarity score S indicates higher content diversity within the image. We prove that this metric is human-aligned in Appendix.

5 Experiments

Experimental Setup

Implementation Details. We conduct experiments based on the open-source latent diffusion model *SD-1.5*, *SD-2.1*, and *SD-XL*. The hyper-parameter N, representing the number of virtual boundaries, is set to 31. We adhere to the default settings and utilize a 50-step DDIM (Song, Meng, and Ermon 2020) sampling process. Following (He et al. 2023), our PBC strategy is applied only during the first 25 denoising steps, with FreeU (Si et al. 2024) integrated. We implement our method by unfold-fold technique of convolution, see details in Appendix. All experiments are conducted using a single RTX 3090 GPU.

Dataset and Evaluation Metrics. For test, we randomly select 500 high-quality prompts from Laion-5B (Schuhmann et al. 2022), each prompt is evaluated with 10 seeds. To evaluate the quality of the generated images, we report the following metrics: Content Richness (CR); Kernel Image Distance (KID) (Bińkowski et al. 2018), which measures the semantic similarity between the generated high-resolution images and the original resolution images; Inception Score (IS) (Salimans et al. 2016), which assesses the diversity of the generated images; Human Preference Score (HPS) (Wu et al. 2023) and ImageReward (IR) (Xu et al. 2024), both of which gauge human preferences for the generated images. Besides, we conduct User Study for further evaluation. Fve participants independently rated images on a scale from 1 to 5 for their image visual quality based on 40 randomly

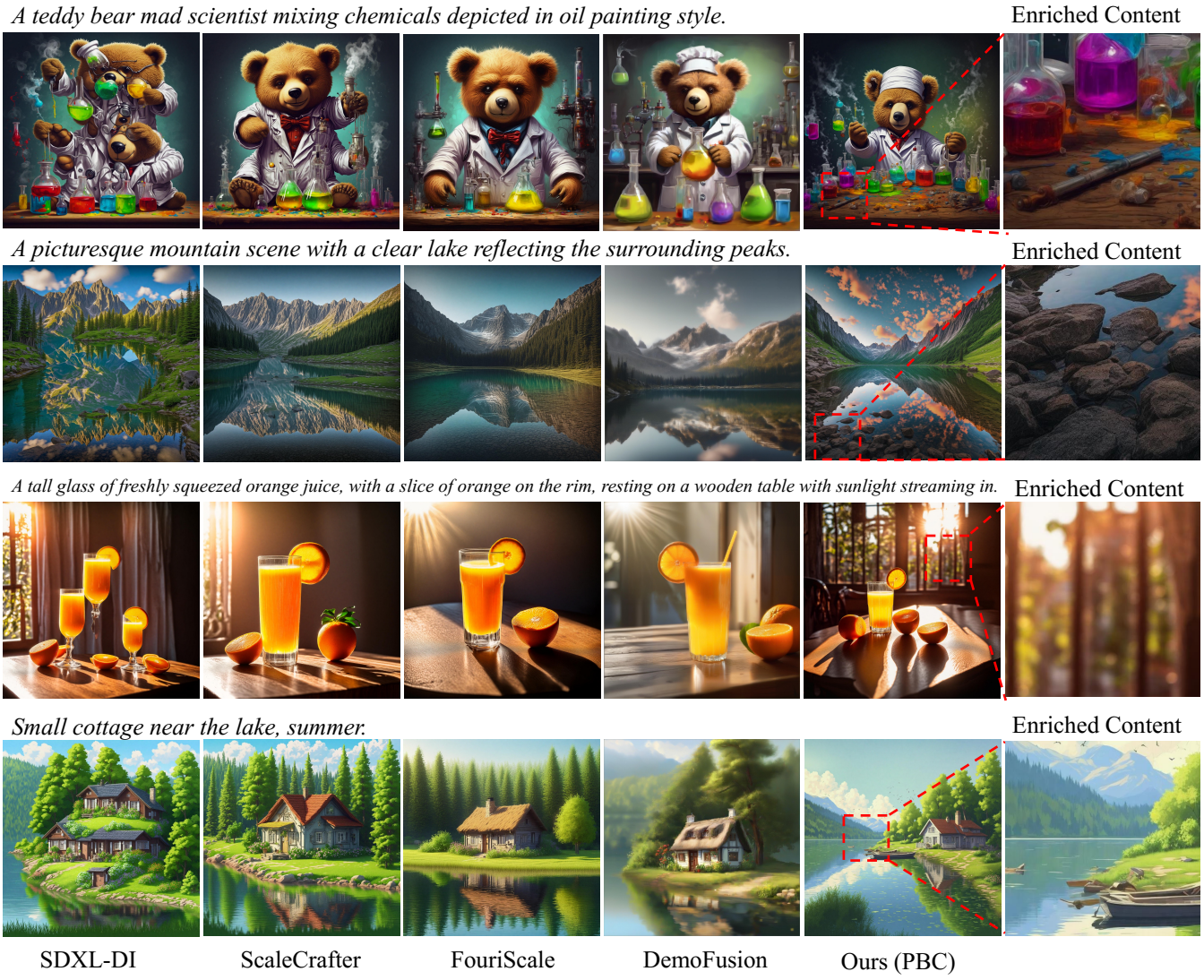

Figure 7: PBC is compared to baseline methods at a resolution of 2048×2048 using *SD-XL*. Our approach demonstrates the ability to generate images with enriched content and high-quality results. Zoom-in for best view.

selected prompts per method . The average user scores are reported.

Baselines. We compare our methods with other training-free high-resolution image generation methods. Single-stage method: (i) *SD-XL* (Rombach et al. 2022) direct inference (DI) (ii) *ScaleCrafter* (He et al. 2023) (iii) *FouriScale* (Huang et al. 2025). Two-stage method: (iv) *DemoFusion* (Du et al. 2024).

Evaluation

Quantitative Results. Quantitative results confirm the superiority of our method. The comparison on *SD-XL* is presented in Table 2, while the results on *SD-1.5* and *SD-2.1* are provided in the Appendix. Our method achieves the best or second-best scores for all quality-related metrics among all single-stage baselines with lower additional time costs. Our method achieves best Content Richness (CR) results among all the baselines, which demonstrates our method can produce content-enriched images.

Qualitative Results. Qualitative results on *SD-XL* comparing our method with the baselines at a resolution of 2048^2 are shown in Figure 7. Additional results on *SD-1.5*, *SD-2.1*, as well as higher resolutions, are provided in the Appendix. The direct inference of *SD-XL* often cause repetitive local image elements and disordered patterns. ScaleCrafter also tends to produce localized repetitions. FouriScale delivers high-quality visual results, but the image content is relatively limited. The images generated by DemoFusion exhibit

Figure 8: PBC is able to generate non-square images. The images are generated by *SD-XL*. Zoom-in for best view.

Method	KID ↓	IS ↑	HPS ↑	IR ↑	CR ↓	Inference Time ↓	User Study ↑
DI	0.0092	9.25	20.76	0.645	30.30	68s	1.1
ScaleCrafter (He et al. 2023)	0.0067	9.53	21.15	**1.078**	29.50	71s (+4%)	3.4
FouriScale (Huang et al. 2025)	0.0103	9.38	21.13	1.023	29.61	130s (+91%)	4.0
PBC (ours)	**0.0067**	**9.95**	**21.20**	1.036	**29.42**	**71s** (+4%)	**4.0**
DemoFusion∗ (Du et al. 2024)	0.0088	11.07	21.33	1.090	29.47	189s (+178%)	2.9

Table 2: Quantitative comparisons with baselines. We evaluate 2048^2 resolution images generated by *SD-XL*. Our method achieves the best or second-best scores for all quality-related metrics among all training-free baselines with lower additional time costs. The best results are marked in **bold**, and the second-best results are marked by underline. $*$ indicates the method is a two-stage method.

Metric	N=1	N=3	N=15	N=31
KID ↓	0.0080	**0.0066**	0.0074	0.0072
IS ↑	7.69	**8.13**	7.95	7.62
CR ↓	30.1769	29.9485	29.6655	**29.5153**

Table 3: Effectiveness of N. The best result is marked in **bold**, and the second best is marked with underline.

noticeable blurriness. Our method not only generates high-quality images but also produces content-rich scenes.

Additional Qualitative Results. Our method can generate non-square high-resolution images by adding virtual boundaries in only one direction, the generated results are shown in the Figure 8. Besides, our method can seamlessly integrate downstream applications, such as spatial-controlled generation, as shown in the Appendix.

Ablation Study

Effect of Virtual Boundary and the Random Perturbation. We evaluate the effectiveness of the random perturbation technique, as shown in Figure 9 right. For better visualization, we set the number of virtual boundaries to $N = 3$ on *SD-2.1*. Without random perturbation, a clear split effect is observed.

Ablation Study on Virtual Boundary Numbers. To assess the impact of the number of virtual boundaries N , we visualize the generated images at a 2048^2 resolution using SD-XL under different values of N, as shown in Figure 9 left. As N increases, the images exhibit richer content and larger scene compositions, indicating that additional virtual boundaries effectively expand the perceived image boundaries. Conversely, an insufficient number of virtual boundaries leads to inconsistencies in image details, resulting in fragmented or incoherent visual elements.

We also quantitatively analyze the effect of N setups in Table 3. It demonstrates that increasing the number of virtual boundary layers results in richer image content. Yet, this

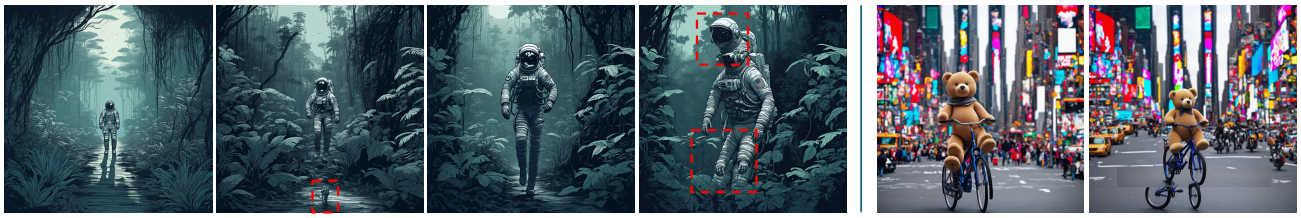

Figure 9: *left*: Effect of virtual boundaries numbers on image quality and content richness. A greater number of virtual boundaries leads to higher image quality and richer content diversity. The images are generated by *SD-XL* with resolution 2048^2. *right*: Effect of Random Perturbation with $N = 3$.

also results in a deviation from the image distribution of the training dataset, leading to a deterioration in the KID metric.

6 Conclusion

In this paper, we explore the position encoding mechanism in diffusion U-Net and identify inconsistent positional encoding as a significant cause of repetitive and disordered patterns in high-resolution image generation. We propose Progressive Boundary Complement (PBC), a training-free method that enhances positional information via hierarchical virtual boundaries. PBC effectively mitigates spatial inconsistencies, expands image boundaries, and enriches content diversity. Extensive experiments demonstrate its superiority.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62173045), the Beijing Natural Science Foundation under Grant F2024203115, the Young Scientists Fund of NSFC (Grant No. 62406035), the China Postdoctoral Science Foundation under Grant Number 2024M750255, and BUPT Excellent Ph.D. Students Foundation (CX20241088).

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Cao, B.; Ye, J.; Wei, Y.; and Shan, H. 2024a. Ap-ldm: Attentive and progressive latent diffusion model for training-free high-resolution image generation. *arXiv preprint arXiv:2410.06055*.

Cao, P.; Zhou, F.; Song, Q.; and Yang, L. 2024b. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*.

Cao, P.; Zhou, F.; Yang, L.; Huang, T.; and Song, Q. 2025. Image is All You Need to Empower Large-scale Diffusion Models for In-Domain Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18358–18368.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6159–6168.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Guo, L.; He, Y.; Chen, H.; Xia, M.; Cun, X.; Wang, Y.; Huang, S.; Zhang, Y.; Wang, X.; Chen, Q.; et al. 2025. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European Conference on Computer Vision*, 39–55. Springer.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Huang, L.; Fang, R.; Zhang, A.; Song, G.; Liu, S.; Liu, Y.; and Li, H. 2025. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European Conference on Computer Vision*, 196–212. Springer.

Hwang, J.; Park, Y.-H.; and Jo, J. 2024. Upsample guidance: Scale up diffusion models without training. *arXiv preprint arXiv:2404.01709*.

Islam, M. A.; Jia, S.; and Bruce, N. D. 2020. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36: 70847–70860.

Kim, Y.; Hwang, G.; Zhang, J.; and Park, E. 2024. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. *arXiv preprint arXiv:2406.18459*.

Lin, Z.; Lin, M.; Zhao, M.; and Ji, R. 2025. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*, 38–53. Springer.

Liu, X.; He, Y.; Guo, L.; Li, X.; Jin, B.; Li, P.; Li, Y.; Chan, C.-M.; Chen, Q.; Xue, W.; et al. 2024. Hiprompt: Tuning-free higher-resolution generation with hierarchical mllm prompts. *arXiv preprint arXiv:2409.02919*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qiu, H.; Zhang, S.; Wei, Y.; Chu, R.; Yuan, H.; Wang, X.; Zhang, Y.; and Liu, Z. 2024. FreeScale: Unleashing the Resolution of Diffusion Models via Tuning-Free Scale Fusion. *arXiv preprint arXiv:2412.09626*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Si, C.; Huang, Z.; Jiang, Y.; and Liu, Z. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Toker, M.; Orgad, H.; Ventura, M.; Arad, D.; and Belinkov, Y. 2024. Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines. *arXiv preprint arXiv:2403.05846*.

Wang, L.; Zhou, F.; Yu, B.; Cao, P.; and Yin, J. 2025. OMEGAS: Object Mesh Extraction from Large Scenes Guided by Gaussian Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wu, H.; Shen, S.; Hu, Q.; Zhang, X.; Zhang, Y.; and Wang, Y. 2024. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3).

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Yang, H.; Bulat, A.; Hadji, I.; Pham, H. X.; Zhu, X.; Tzimiropoulos, G.; and Martinez, B. 2024a. FAM Diffusion: Frequency and Attention Modulation for High-Resolution Image Generation with Stable Diffusion. *arXiv preprint arXiv:2411.18552*.

Yang, T.; Lan, C.; Lu, Y.; et al. 2024b. Diffusion Model with Cross Attention as an Inductive Bias for Disentanglement. *arXiv preprint arXiv:2402.09712*.

Yi, M.; Li, A.; Xin, Y.; and Li, Z. 2024. Towards Understanding the Working Mechanism of Text-to-Image Diffusion Model. *arXiv preprint arXiv:2405.15330*.

Zhang, S.; Chen, Z.; Zhao, Z.; Chen, Y.; Tang, Y.; and Liang, J. 2025. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, 145–161. Springer.

Zhang, Z.; Li, R.; and Zhang, L. 2024. FreCaS: Efficient Higher-Resolution Image Generation via Frequency-aware Cascaded Sampling. *arXiv preprint arXiv:2410.18410*.

Zhou, F.; Yin, J.; and Li, P. 2024. Lifting by image–leveraging image cues for accurate 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7632–7640.