A.4 Analysis of the Progressive Upsampling Generation Process in AP-LDM

To clearly illustrate the progressive upsampling process of AP-LDM, we set $\eta_2 = [0.2, 0.2, 0.2]$ to generate 4096×4096 images. As shown in Fig. 14, the images generated at different sub-stages of AP-LDM exhibit a high degree of consistency, with only minor differences in details. Since our task focuses on generating HR images rather than traditional image super-resolution, these differences in details are reasonable.

Figure 14: **Illustration of the progressive upsampling generation process.** The inference speed is evaluated on a single NVIDIA 3090 GPU.

Another noteworthy observation is that even though the progressive upsampling generation sub-stages involve only a small number of denoising steps (*e.g.*, 10 steps), the majority of the generation time is still consumed in these sub-stages. This is because the time required for denoising models to perform inference increases dramatically with the image size. For each denoising step, the time required for HR images is several times that for low-resolution images. Consequently, repeating a full denoising process at high resolution is extremely time-consuming (Du et al., 2024; Lin et al., 2024). Considering that HR and low-resolution images should share the same low-frequency structure, and that DMs naturally generate low-frequency structures first during denoising (Yu et al., 2023; Teng et al., 2023), AP-LDM effectively leverages the prior knowledge of low-frequency structures in low-resolution images. This significantly reduces the number of denoising steps needed at high resolution, thereby substantially accelerating the image generation process.

A.5 How Does PFSA Work?

In this section, we further elaborate on the working mechanism of PFSA. Specifically, the functionality of PFSA can be described in two aspects: (**i**) clustering the related tokens in the latent representations; (**ii**) adjusting the amplitude of the high-frequency and low-frequency components in the latent representations.

A.5.1 PFSA Clusters Tokens of Latent Representations

PFSA reorganizes tokens based on their similarities. Intuitively, this enables PFSA to perform token clustering, which enhances the structural consistency of latent representations. To demonstrate the clustering effect of PFSA, we calculated the deviation of the tokens' mean (DTM) of the latent representations $\tilde{\boldsymbol{z}}_t$ and $\boldsymbol{z}_t$. Concretely, assuming $\boldsymbol{z}_t \in \mathbb{R}^{h \times w \times c}$, and $\boldsymbol{Z}_t = \text{Flatten}(\boldsymbol{z}_t) = [\boldsymbol{y}_{t1}, \ldots, \boldsymbol{y}_{tN}] \in \mathbb{R}^{N \times c}$, where $N = h \times w$, we calculate DTM as:

$$\text{DTM} = [\text{mean}(\boldsymbol{y}_{ti}) - \text{mean}(\boldsymbol{Z}_t) \quad \text{for} \quad i = 1, \ldots, N] \tag{5}$$

To provide an intuitive illustration of the clustering effect of PFSA, we visualize the DTM based on token indices (*i.e.*, $i = 1, \ldots, N$) when t is relatively large. As shown in columns (A) and (B) of Fig. 15, compared to the DTM of $\boldsymbol{z}_t$ (blue points), the DTM of $\tilde{\boldsymbol{z}}_t$ (red points) becomes more dispersed and exhibits distinct stripe patterns, indicating that PFSA indeed clusters the tokens of the latent representations. This clustering effect can be more directly demonstrated when t is smaller. As shown in the heatmaps in columns (C) and (D) of Fig. 15, it is evident that PFSA clusters semantically related tokens.

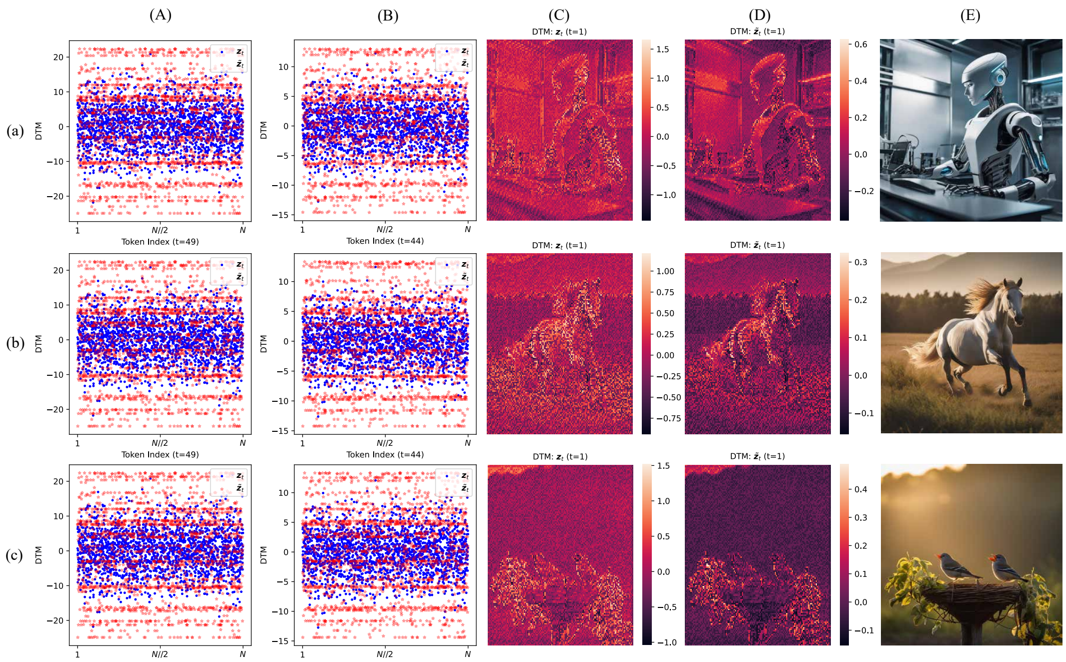

Figure 15: **The clustering effect of PFSA.** Columns (A), (B), (C), and (D) show the DTM of latent representations, while column (E) presents the corresponding generated RGB images.

A.5.2 PFSA Adjusts the Amplitude of High- and Low-Frequency Components in Latent Representations

The aim of this section is to explain: (**i**) why appropriately delaying attentive guidance can resolve structural deformation issues (as shown in Fig. 8), (**ii**) why attentive guidance enhances the details and colors of the image (as shown in Fig. 6, 7, and 12), and (**iii**) why applying attentive guidance in the later stages of denoising does not enhance the image details and colors (as shown in Fig. 9).

To explain the aforementioned three points, as shown in Fig. 16, we calculate the Fourier transforms of $\boldsymbol{z}_t$ (blue solid line) and $\tilde{\boldsymbol{z}}_t$ (red solid line), along with the mean of the standard deviations for all their channels (dashed line). It can be observed that PFSA significantly alters the relative amplitudes of the high- and low-frequency components in the latent representations during the initial denoising steps (from $t = 49$ to $t = 47$), particularly affecting the low-frequency components, which results in structural deformation. During the early and middle stages of denoising (from $t = 44$ to $t = 29$), PFSA increases the amplitudes of high-frequency components in the latent representations, which explains why attentive guidance leads to richer details and colors. In the later stages of denoising (from $t = 28$ to $t = 0$), PFSA slightly suppresses the high-frequency components of the latent representations while almost leaving the low-frequency components unchanged. This explains why applying attentive guidance in the later stages of denoising cannot enrich details and colors of the generated images.

Additionally, Fig. 16 shows that PFSA increases the standard deviation of $\tilde{\boldsymbol{z}}_t$ during the early and middle stages of denoising, while decreasing it in the later stages. The trend of the standard deviation

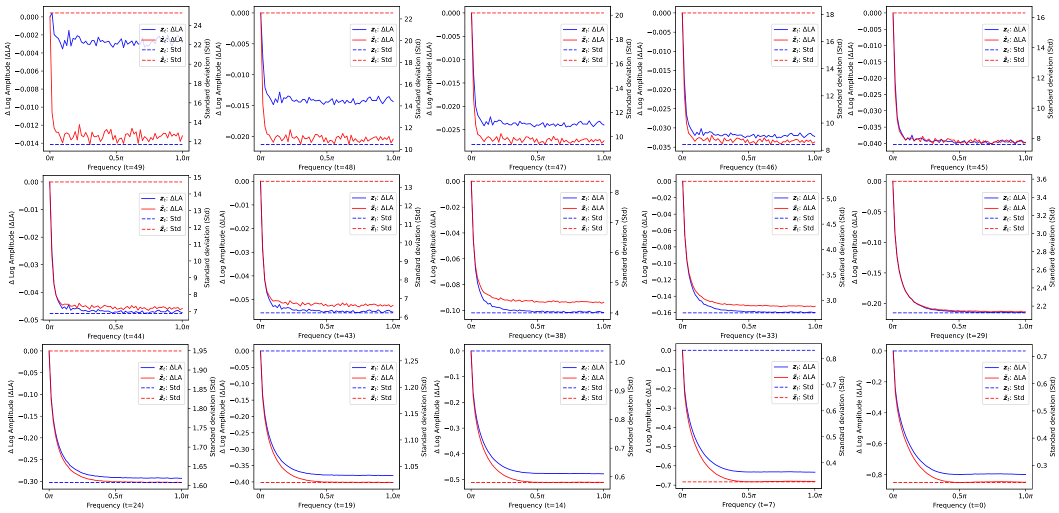

Figure 16: **The Fourier transform of the latent representation and the mean of the standard deviations across all channels.** z_t is represented in blue, while $\tilde{z}_t$ is represented in red; the Fourier transforms are shown as solid lines, and the standard deviations are shown as dashed lines. The results are based on the generation process of 5k images.

changes is closely consistent with the variation in the amplitude of the high-frequency components. We conjecture that this is because the amount of information in the latent representations is positively correlated with the standard deviation, where a larger standard deviation corresponds to more image details and larger high-frequency components.

A.6 Comparison with Additional Baseline Models

In this section, we compare AP-LDM with additional baseline models. Specifically, we include recently proposed HiDiffusion (Zhang et al., 2025) and a super-resolution model (SDXL+BSRGAN, *i.e.*, the outputs of SDXL are upsampled using BSRGAN (Zhang et al., 2021)). Since HiDiffusion experiments are conducted using professional-grade V100 GPUs without optimization for ultra-high-resolution images, it is not feasible to generate images with resolutions above 2048×4096 on consumer-grade GPUs such as the 3090. Due to device limits, we compare its performance only at the resolution of 2048×2048. The experimental setup remains the same as described in §4.

A.6.1 Quantitative Comparison

Comparison of generated image quality. Table 8 presents the extended quantitative comparison results of generated image quality, which further demonstrate our effectiveness on HR image generation. We observe that models employing progressive upsampling generation (*e.g.*, AP-LDM, DemoFusion, and AccDiffusion) achieved relatively better results, showing the robustness of the progressive upsampling generation paradigm.

Table 8: **Quantitative comparison results**. The best results are marked in **bold**, and the second best results are marked by <u>underline</u>.

Method	2048 × 2048					2048 × 4096					4096 × 2048					4096 × 4096				
	FID	IS	FID$_c$	IS$_c$	CLIP	FID	IS	FID$_c$	IS$_c$	CLIP	FID	IS	FID$_c$	IS$_c$	CLIP	FID	IS	FID$_c$	IS$_c$	CLIP
SDXL	99.9	14.2	80.0	16.9	25.0	149.9	9.5	106.3	12.0	24.4	173.1	9.1	108.5	11.5	23.9	191.4	8.3	114.1	12.4	22.9
MultiDiff.	98.8	14.5	67.9	17.1	24.6	125.8	9.6	71.9	15.7	24.6	149.0	9.0	70.5	14.4	<u>24.4</u>	168.4	6.5	76.6	14.4	23.1
ScaleCrafter	98.2	14.2	89.7	13.3	25.4	161.9	10.0	154.3	7.5	23.3	175.1	9.7	167.3	8.0	21.6	164.5	9.4	170.1	7.3	22.3
UG	82.2	17.6	65.8	14.6	<u>25.5</u>	155.7	8.2	165.0	6.6	21.7	185.3	6.8	175.7	6.2	20.5	187.3	7.0	197.6	6.3	21.8
DemoFusion	72.3	**21.6**	53.5	**19.1**	25.2	96.3	17.7	62.3	<u>15.0</u>	<u>25.0</u>	99.6	16.4	61.9	<u>14.7</u>	<u>24.4</u>	101.4	20.7	63.5	13.5	<u>24.7</u>
AccDiff.	71.6	21.0	52.7	17.0	25.1	95.5	16.4	62.9	11.1	24.5	102.2	15.2	65.4	11.5	24.2	103.2	20.1	65.9	13.3	24.6
SDXL+BSR.	<u>66.2</u>	<u>21.1</u>	<u>47.5</u>	16.6	**25.7**	**80.7**	<u>19.8</u>	**50.2**	12.3	**25.1**	**92.7**	<u>17.6</u>	<u>57.9</u>	12.1	**24.9**	**90.0**	<u>20.9</u>	**56.0**	<u>13.8</u>	**25.2**
HiDiff.	81.0	16.8	64.1	14.2	24.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AP-LDM	**66.0**	21.0	**47.4**	<u>17.5</u>	25.1	<u>89.0</u>	**20.3**	<u>56.0</u>	**19.0**	<u>25.0</u>	<u>93.2</u>	**19.5**	**56.9**	**16.5**	**24.9**	<u>90.6</u>	**21.1**	<u>59.0</u>	**14.8**	24.6

In contrast, HiDiffusion fell short compared to methods using progressive upsampling. We speculate that its suboptimal performance is due to two factors: (**i**) the forced resizing of deep feature maps during the generation process, which causes significant distribution shifts; and (**ii**) the use of MSW-MSA (a sparse attention mechanism similar to SwinTransformer (Liu et al., 2021)), which forcibly alters the attention's receptive field and sequence length, leading to severe shifts in the entropy of attention weights (Jin et al., 2024). The aforementioned two issues prevent HiDiffusion from fully addressing the problem of repeated object structures and result in severe artifacts and deformations in the generated images (as shown in Fig. 17).

The super-resolution model (SDXL + BSRGAN) demonstrated strong performance in quantitative experiments, a phenomenon also observed in the DemoFusion's experiments. This is because super-resolution models can at least preserve the low-frequency structures of images without significant errors. However, as discussed in DemoFusion (Du et al., 2024) and AccDiffusion (Lin et al., 2024), super-resolution models fail to add finer details to high-resolution images (as shown in Fig. 18).

Comparison of resource consumption. We also compare the inference time and GPU memory usage required by the models. Specifically, we test the minimum GPU memory requirements during model inference based on the model's open-source code. Table 9 shows the resource consumption of different models when generating images at various resolutions. SDXL+BSRGAN, unlike DMs, does not require iterative inference, allowing it to achieve the fastest generation speed. However, the super-resolution model fails to generate the level of detail expected in high-resolution images, which has limited its widespread adoption.

Table 9: **Model resource consumption**. The best results are marked in **bold**, and the second best results are marked by <u>underline</u>. Time unit: minute. Storage unit: GB.

Method	2048×2048		2048×4096		4096×4096	
	time cost	storage cost	time cost	storage cost	time cost	storage cost
SDXL	1.0	15.9	3.0	<u>16.1</u>	8.0	**16.6**
MultiDiff.	3.0	22.0	6.0	16.8	15.0	<u>16.8</u>
ScaleCrafter	1.0	17.4	6.0	17.6	19.0	19.1
UG	1.8	23.9	4.0	16.5	11.1	18.0
DemoFusion	3.0	<u>15.2</u>	11.0	18.4	25.0	<u>16.8</u>
AccDiff.	3.0	22.1	12.7	23.0	26.0	22.1
SDXL+BSR.	<u>1.0</u>	**14.6**	**1.0**	**11.1**	**1.0**	21.1
HiDiff.	0.8	23.9	-	-	-	-
AP-LDM	**0.6**	16.0	<u>2.0</u>	21.1	<u>5.7</u>	23.8

It is worth noting that for high-resolution image generation tasks, the memory bottleneck lies in the encoding and decoding of the VAE rather than interpolating the image in pixel space. To address the challenges of encoding and decoding high-resolution images, researchers typically employ tiled encoders and tiled decoders. In this work, we also utilize a tiled-encoder and decoder when generating ultra-high-resolution images, allowing us to generate images with resolutions up to 4096×7280 or higher on a 24GB VRAM NVIDIA 3090 GPU (as shown in Fig. 1).

A.6.2 QUALITATIVE COMPARISON

Qualitative Comparison with HiDiffusion. We conduct extensive qualitative comparison experiments between AP-LDM and HiDiffusion, with the results shown in Fig. 17. From the figure, it can be observed that AP-LDM consistently generates high-quality, high-resolution images. Although capable of generating some good results, HiDiffusion suffers from significant distribution shifts in the UNet features due to forced feature scaling and the use of window attention, which alters the sequence length during attention computation. This often causes the generated images to collapse, as illustrated in Fig. 17 (a)–(e). Even when HiDiffusion avoids image collapse, it frequently produces noticeable artifacts and distortions, as shown in Fig. 17 (f)–(h). In Fig. 17 (i) and (j), HiDiffusion still exhibits severe structural repetition in the generated outputs, indicating that merely resizing the deep features of the UNet is insufficient to completely eliminate low-frequency structural errors.

Qualitative Comparison with SDXL+BSRGAN. We conducted extensive qualitative comparisons between AP-LDM and SDXL+BSRGAN. Specifically, we compared their performance at resolutions of 2048×2048 (Fig. 18 (a)-(d)) and 4096×4096 (Fig. 18 (e)-(h)). As we can see, compared to AP-LDM, SDXL+BSRGAN, while maintaining decent image structure, fails to generate the level of detail expected from HR images. The absence of these details sometimes leads to the model's inability to simulate realistic scenes. For example, in Fig. 18 (c), SDXL+BSRGAN fails to generate realistic shadows. At higher resolutions (*e.g.*, 4096×4096), SDXL+BSRGAN may introduce artifacts, as shown in Fig. 18 (e) and (g).

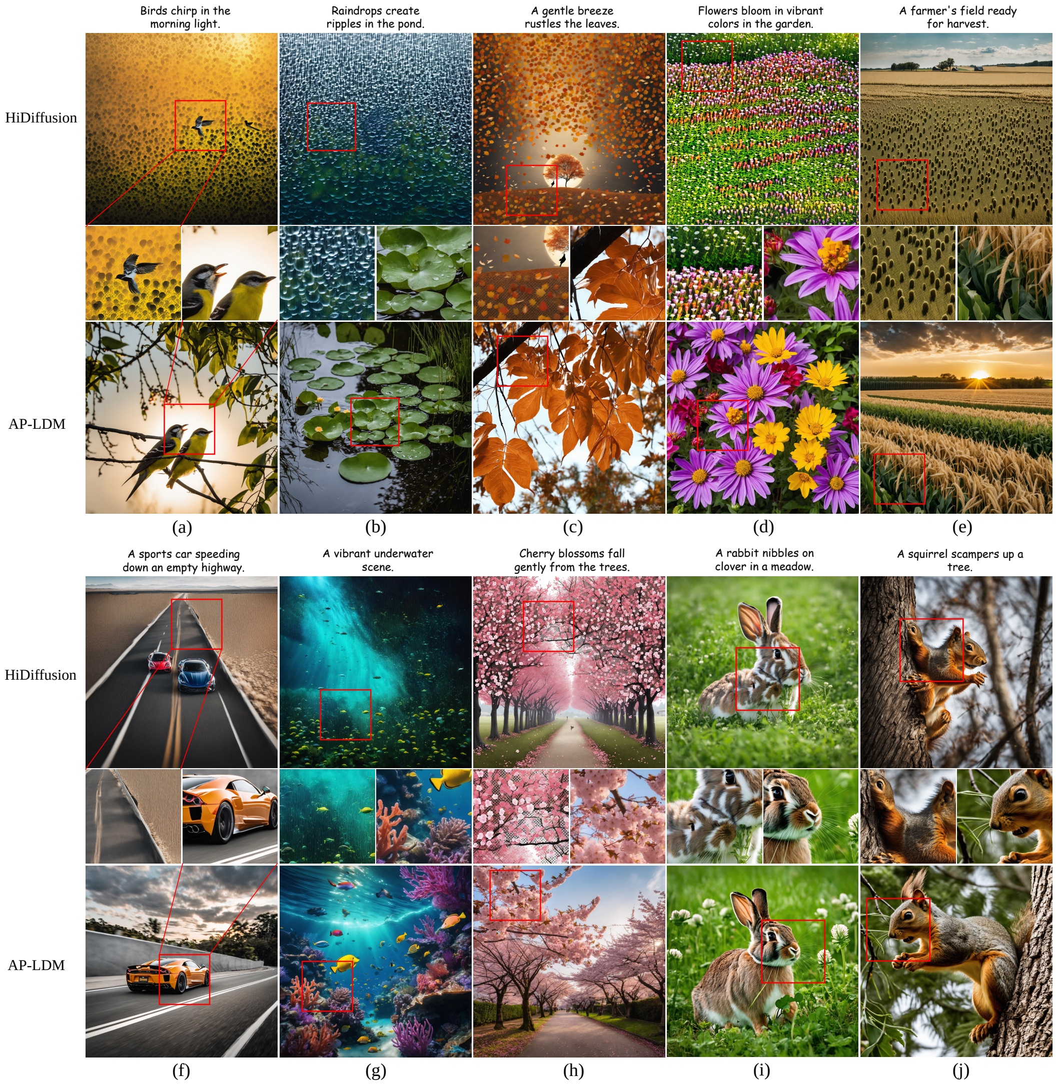

Figure 17: **Quantitative comparison with HiDiffusion**, where all images have a resolution of 2048×2048. The prompts for the generated images are provided above the figures.

A.7 Attentive Guidance Also Works in Other Generation Frameworks

In this section, we apply attentive guidance to other generative frameworks to demonstrate its generalization capability. Specifically, we apply attentive guidance to the generative frameworks of HiDiffusion and DemoFusion, and conduct both quantitative and qualitative ablation studies.

A.7.1 Quantitative Ablation in Other Generative Frameworks

In this section, considering the long inference time of DemoFusion, we perform quantitative ablation studies on attentive guidance using the HiDiffusion generation frameworks at a resolution of 2048×2048. All experimental settings are consistent with those in §4.

Table 10: **Quantitative ablation of attentive guidance using HiDiffusion frameworks**. The best results are marked in bold. AG: attentive guidance.

Method	FID	IS	FID_c	IS_c	CLIP
HiDiffusion	81.0	16.8	64.1	14.2	**24.9**
HiDiff.+AG	**79.4**	**17.0**	**62.4**	**14.6**	**24.9**

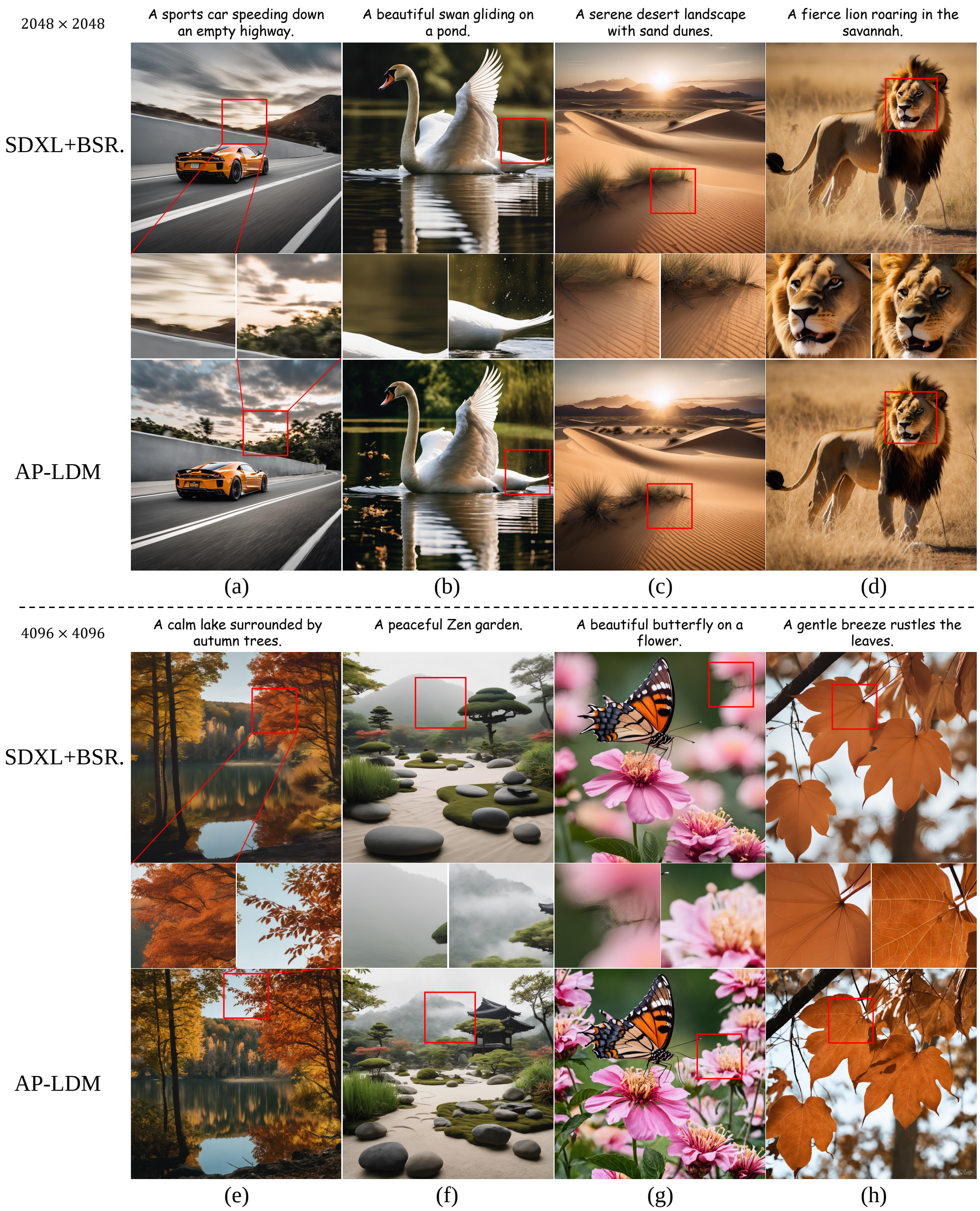

Figure 18: **Qualitative comparison with SDXL+BSRGAN**. Figures (a)-(d) have a resolution of 2048×2048, while Figures (e)-(h) have a resolution of 4096×4096. The prompts for the generated images are provided above the figures.

Table 10 presents the quantitative ablation results using the HiDiffusion framework. It is evident that incorporating attentive guidance improves HiDiffusion across all metrics. This is further corroborated by the qualitative analysis in Fig. 19, which demonstrates that attentive guidance alleviates some of the structural collapses observed in HiDiffusion.

A.7.2 Qualitative Ablation Studies in Other Generative Frameworks

HiDiffusion+attentive guidance. HiDiffusion enforces scaling of the UNet feature maps during image generation, which often leads to structural collapse and deformations in the generated images (as shown in Fig. 17). Fig. 19 (a)-(f) demonstrate that using attentive guidance effectively mitigates

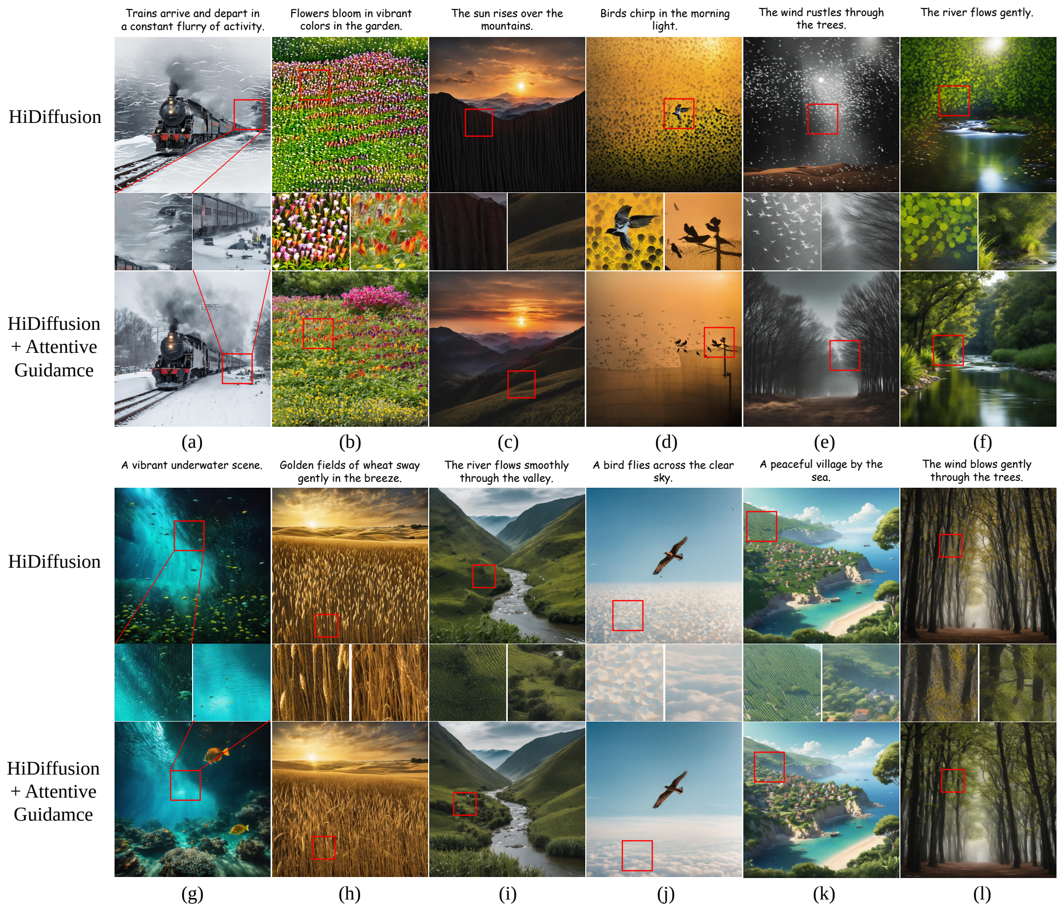

Figure 19: **Qualitative ablation of attentive guidance in the HiDiffusion Framework**. All images have a resolution of 2048×2048. Figures (a)-(f) demonstrate that attentive guidance can mitigate the issue of structural collapse in generated images, while Figures (g)-(l) show that attentive guidance resolves structural deformation issues and enhances image details.

the issue of structural collapse in synthesized images. Fig. 19 (g)-(l) further show that attentive guidance can also address the structural deformation inherent to HiDiffusion, enhance image details, and improve overall image quality.

DemoFusion+attentive guidance. In the analysis presented in §4.3 and §A.3, we observed that DemoFusion tends to produce repetitive structures (as shown in Fig. 5 and 13), a phenomenon also noted in other studies (Lin et al., 2024). We incorporate attentive guidance into the generative framework of DemoFusion. As shown in Fig. 20 (a)-(e), attentive guidance effectively mitigates the issue of repetitive structures in DemoFusion. Fig. 20 (f)-(j) further illustrate role of attentive guidance in enriching image details and enhancing overall image quality.

A.8 Comparative and Ablation Analysis Based on StableDiffusion 2.1

To validate the generalization capability of AP-LDM, we conducted extensive quantitative and qualitative analyses using StableDiffusion 2.1 (SD2.1) as the pretrained base model.

A.8.1 Comparison Experiments

Quantitative comparison. Since the code for using SD2.1 as the pretrained model in AccDiffusion and DemoFusion is not publicly available, we compare AP-LDM with ScaleCrafter in this section. We compared the model performance at four resolutions: 1536×1536, 1024×2048, 2048×1024, and 2048×2048. Considering that SD2.1's generation capabilities are weaker than

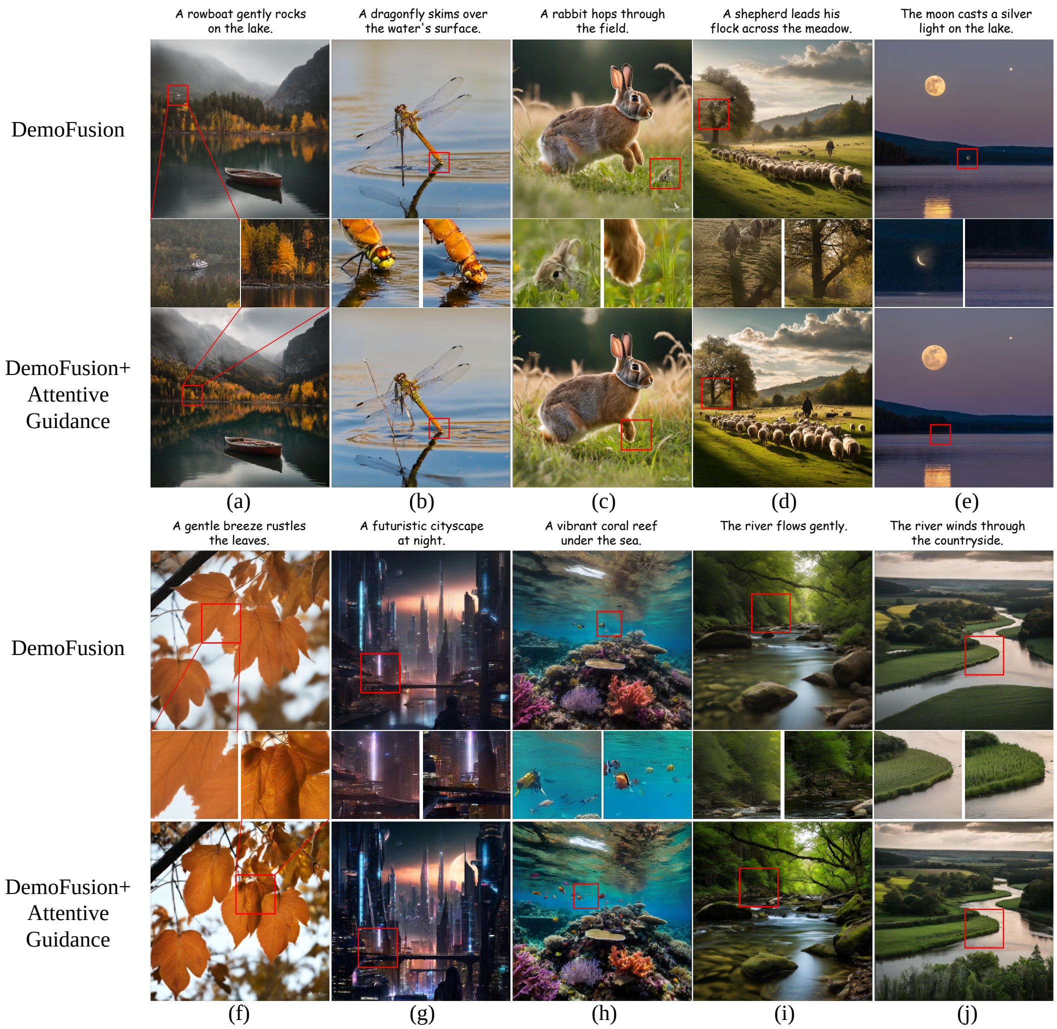

Figure 20: **Qualitative ablation of attentive guidance in the DemoFusion Framework**. All images have a resolution of 2048×2048. Figures (a)-(e) demonstrate that attentive guidance effectively mitigates the issue of repetitive structures in images, while Figures (f)-(j) showcase attentive guidance's ability to enrich image details.

SDXL, we set $\eta_2 = [0.2, 0.2, 0.3]$ for the experiments in this section, while keeping other settings consistent with §4.

Table 11: **Quantitative comparison results based on SD2.1**. The best results are marked in **bold**.

Method	1536 × 1536					1024 × 2048					2048 × 1024					2048 × 2048				
	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP
SD2.1	95.4	17.8	83.4	15.8	25.0	85.8	15.9	76.1	16.3	**25.2**	101.8	15.8	79.8	16.8	24.6	121.7	14.4	92.7	14.4	24.5
ScaleCrafter	140.4	10.6	136.4	9.7	21.9	150.0	10.1	139.3	10.1	21.7	149.8	10.4	135.6	11.5	21.8	144.2	10.4	135.2	10.3	23.4
AP-LDM	**60.3**	**21.0**	**50.6**	**18.3**	**25.4**	**61.1**	**19.9**	**54.1**	**18.4**	25.0	**63.7**	**19.2**	**50.4**	**18.2**	**24.7**	**60.5**	**21.5**	**48.8**	**17.2**	**25.3**

Table 11 presents the results of the quantitative comparison, demonstrating that AP-LDM maintains strong performance when using SD2.1 as the pre-trained model. ScaleCrafter, on the other hand, performs suboptimally due to its tendency to produce structural collapse in generated images, a phenomenon more evident in the qualitative analysis.

Qualitative comparison. Fig. 21 presents the results of the qualitative comparison. It can be observed that when generating high-resolution images, SD2.1 also encounters issues with repetitive object structures. ScaleCrafter frequently exhibits structural collapse in generated images during denoising with SD2.1, leading to its suboptimal performance. In contrast, AP-LDM consistently produces high-quality results across all resolutions, demonstrating the generalizability of the AP-LDM generation framework.

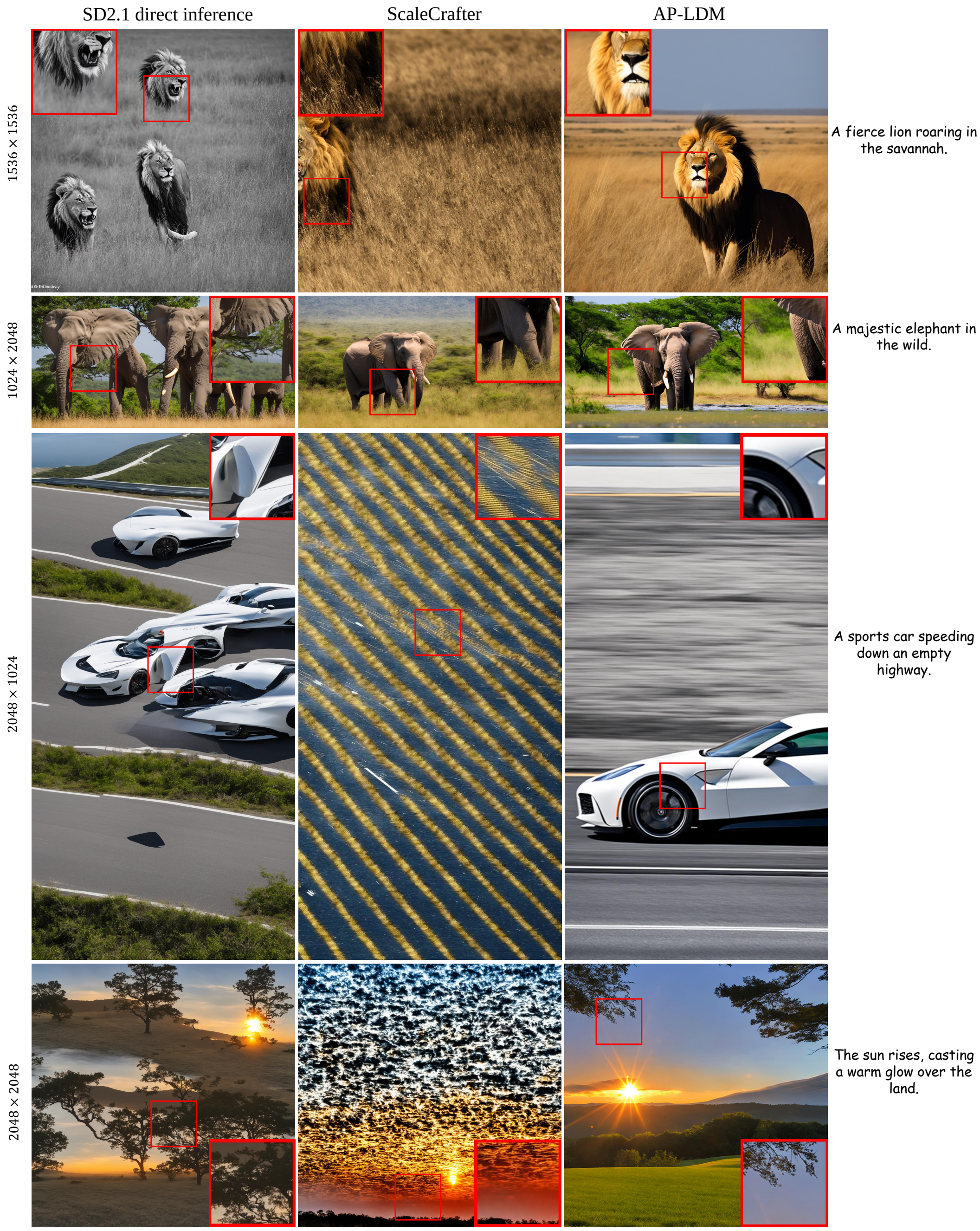

Figure 21: **Qualitative comparison using SD2.1 as the pretrained model**.

A.8.2 Ablation Study on attentive guidance

Quantitative ablation. Table 12 shows the results of the quantitative ablation on attentive guidance using SD2.1 as the pretrained model. It can be observed that attentive guidance leads to improvements in metrics. These improvements are more evident in the qualitative ablation analysis.

Table 12: **Quantitative ablation results based on SD2.1**. The best results are marked in **bold**.

Method	1536 × 1536					1024 × 2048					2048 × 1024					2048 × 2048				
	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP	FID	IS	FID_c	IS_c	CLIP
w/o AG	61.2	20.9	**50.2**	**18.9**	25.2	61.5	19.6	**54.0**	**19.5**	24.9	64.6	**19.6**	**49.2**	17.0	24.6	61.1	21.2	**46.5**	**18.2**	25.2
w/ AG	**60.3**	**21.0**	50.6	18.3	**25.4**	**61.1**	**19.9**	54.1	18.4	**25.0**	**63.7**	19.2	50.4	**18.2**	**24.7**	**60.5**	**21.5**	48.8	17.2	**25.3**

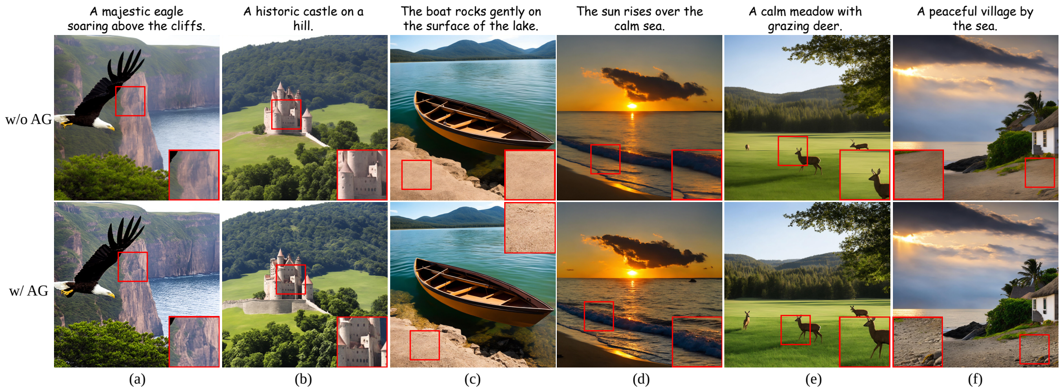

Figure 22: **Ablation study of attentive guidance using SD2.1 as the pre-trained model**. Resolution: 2048×2048.

Qualitative ablation. Fig. 22 presents the ablation analysis of attentive guidance based on SD2.1. From the figure, it can be observed that attentive guidance also enhances detail richness and color vibrancy when using SD2.1, further demonstrating its generalization capability.