

Appendix

A	How Does TFSA/Attention Guidance Work?	15
A.1	TFSA Clusters Semantically Related Tokens	15
A.2	TFSA Adjusts the Amplitude of High- and Low-frequency Components	16
A.3	Visualization of Attention Maps in TFSA	17
B	Supplementary Qualitative Comparison of §4.3	18
C	Supplementary Ablation Experiments of §5	18
C.1	Further Qualitative Analysis of Attention Guidance	18
C.2	Ablation on the hyper-parameters of Attention Guidance	18
C.3	Ablation on Progressive Scheduler Value	18
D	Ablation on the Attention Guidance Components	20
D.1	Ablation on the Guidance Scale Decay Strategy	20
D.2	Ablation on the Attention Calculation Paradigm	21
E	Further Model Efficiency Analysis	21
F	RepLDM Algorithm	23
G	Robustness Analysis	24
H	Comparative and Ablation Analysis Based on StableDiffusion 2.1	25
H.1	Comparison Experiments	25
H.2	Ablation Study on Attention Guidance	25
I	Attention Guidance Also Works in Other Generation Frameworks	26
I.1	Quantitative Ablation	26
I.2	Qualitative Ablation	26
J	Super-Resolved Images Tend to Lack High-Resolution Details	27
K	Memory Usage Analysis	28

A How Does TFSA/Attention Guidance Work?

In this section, we further elaborate on the working mechanism of attention guidance. Our attention guidance enhances the structural consistency of the latent representation by integrating the output of TFSA. Therefore, we conduct a detailed analysis of TFSA. Specifically, the functionality of TFSA 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.1 TFSA Clusters Semantically Related Tokens

Visualization of the clustering effect of TFSA. TFSA reorganizes tokens based on their similarities. Intuitively, this enables TFSA to perform token clustering, which enhances the structural

consistency of latent representations. To demonstrate the clustering effect of TFSA, we calculated the deviation of the tokens' mean (DTM) of the latent representations \tilde{z}_t and z_t . Concretely, assuming $z_t \in \mathbb{R}^{h \times w \times c}$, and $Z_t = \text{Flatten}(z_t) = [y_{t1}, \dots, y_{tN}] \in \mathbb{R}^{N \times c}$, where $N = h \times w$, we calculate DTM as:

$$\text{DTM} = [\text{mean}(y_{ti}) - \text{mean}(Z_t) \text{ for } i = 1, \dots, N] \quad (5)$$

To provide an intuitive illustration of the clustering effect of TFSA, we visualize the DTM based on token indices (*i.e.*, $i = 1, \dots, N$) when t is relatively large. As shown in columns (A) and (B) of Fig. 13, compared to the DTM of z_t (blue points), the DTM of \tilde{z}_t (red points) becomes more dispersed and exhibits distinct stripe patterns, indicating that TFSA 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. 13, it is evident that TFSA clusters semantically related tokens.

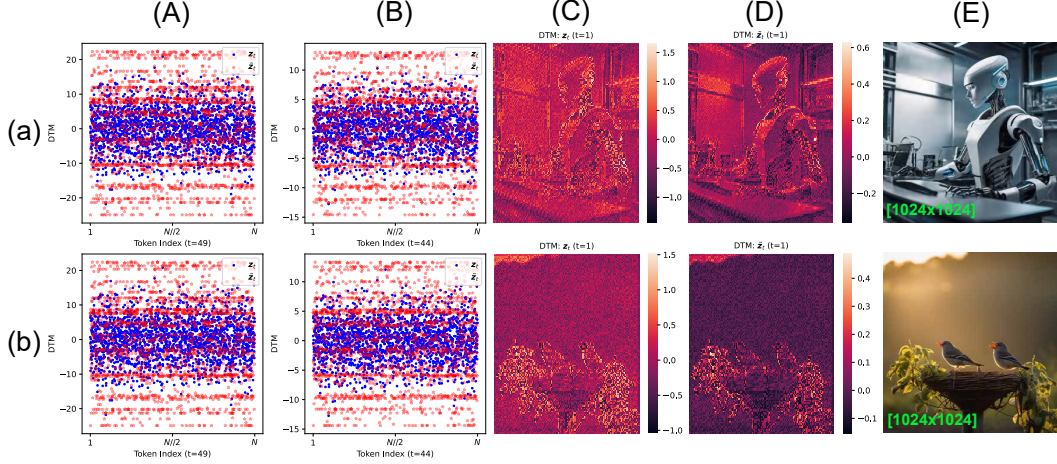


Figure 13: **The clustering effect of TFSA.** Columns (A), (B), (C), and (D) show the DTM of latent representations, while column (E) presents the corresponding generated RGB images.

The clustering effect of TFSA leads to accelerated structural denoising. Fig. 13 shows that the clustering effect of TFSA clarifies the semantic structures of objects, enabling the model to complete the denoising of low-frequency structures earlier. This early revelation of the overall image layout provides a stronger prior for subsequent fine-detail generation. To illustrate this, Fig. 14 presents the denoising process for the ablation of attention guidance. Note the regions highlighted by red boxes. With the incorporation of attention guidance, these areas exhibit clearer structures, which facilitates the generation of more affluent details and more vivid colors in subsequent steps.

To quantitatively demonstrate that TFSA accelerates structural emergence, we calculate the SSIM between z_t and z_0 , where $t \in 1, 2, \dots, T-1$, and $T = 50$. As shown in Fig. 15, compared to the naive denoising process, attention guidance consistently drives the latent representations closer to their final states at each step, indicating the structural foreseeability of TFSA.

A.2 TFSA Adjusts the Amplitude of High- and Low-frequency Components

The aim of this experiment is to explain: (i) why appropriately delaying attention guidance can resolve structural deformation issues (as shown in Fig. 9); (ii) why attention guidance enhances the details and colors of the image (as shown in Fig. 6 and 8); and (iii) why applying attention guidance in the later stages of denoising does not enhance the image details and colors (as shown in Fig. 10).

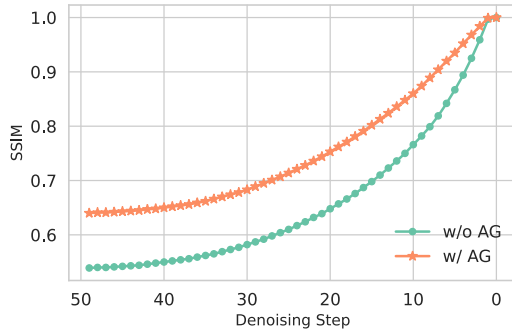


Figure 15: **Quantitatively analysis on the clustering effect of TFSA.** We calculate the SSIM between noised latents z_t ($1 \leq t \leq 49$) and their corresponding clean latent z_0 .

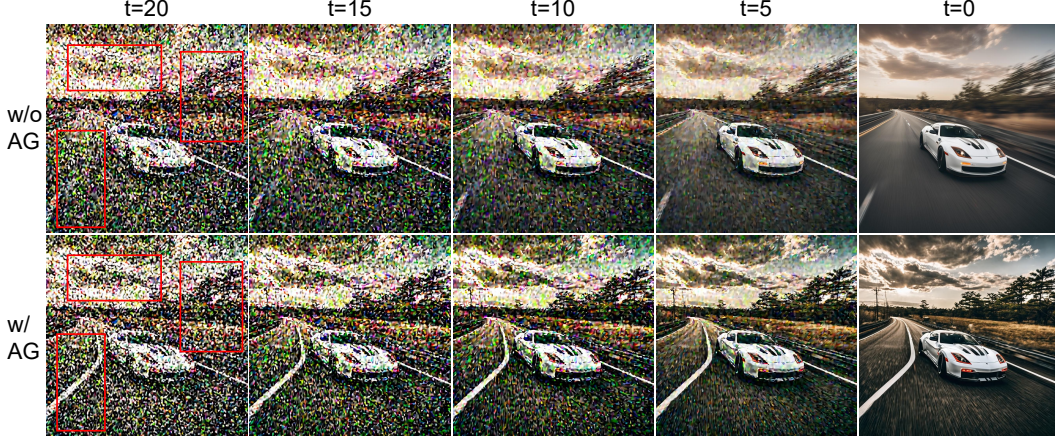


Figure 14: **Denoising visualization for the ablation of attention guidance.** As indicated by the red boxes, the clustering effect of TFSA prompts earlier structural emergence, delivering better prior for subsequent fine-detail generation. Resolution: 1024×1024 .

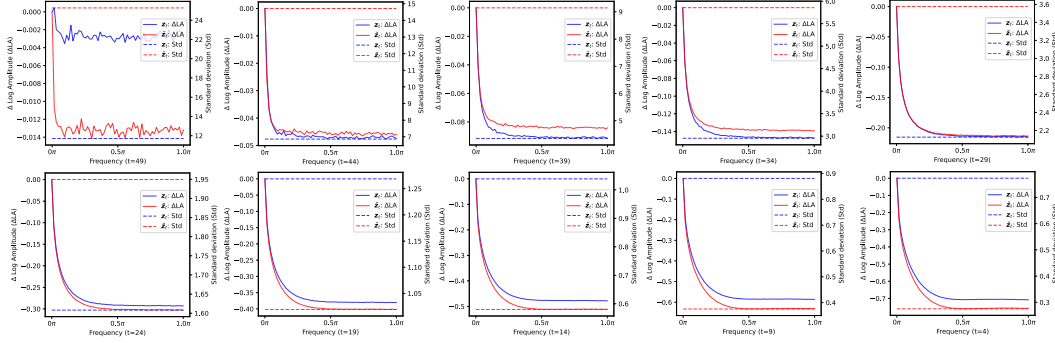


Figure 16: **The Fourier transform of the latent representation and the mean of the standard deviations across all channels.** z_t is represented in blue, while \tilde{z}_t is represented in red; the Fourier transforms are shown as solid lines, and the standard deviations are shown as dashed lines. The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

To explain the aforementioned three points, as shown in Fig. 16, we calculate the Fourier transforms of z_t (blue solid line) and \tilde{z}_t (red solid line), along with the mean of the standard deviations for all their channels (dashed line). It can be observed that TFSA 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 $t = 29$), TFSA increases the amplitudes of high-frequency components in the latent representations, which explains why attention guidance leads to richer details and colors. In the later stages of denoising (from $t = 28$ to $t = 0$), TFSA slightly suppresses the high-frequency components of the latent representations while almost leaving the low-frequency components unchanged. This explains why applying attention guidance in the later stages of denoising cannot enrich details and colors of the generated images.

Additionally, Fig. 16 shows that TFSA increases the standard deviation of \tilde{z}_t during the early and middle stages of denoising, while decreasing it in the later stages. The trend of the standard deviation 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.3 Visualization of Attention Maps in TFSA

To further demonstrate the clustering effect of TFSA on related tokens, we visualize its attention maps. As shown in Fig. 17, without using projection matrices, the correlations between tokens are determined jointly by their represented colors and semantics. For example, in Fig. 17(a), the key

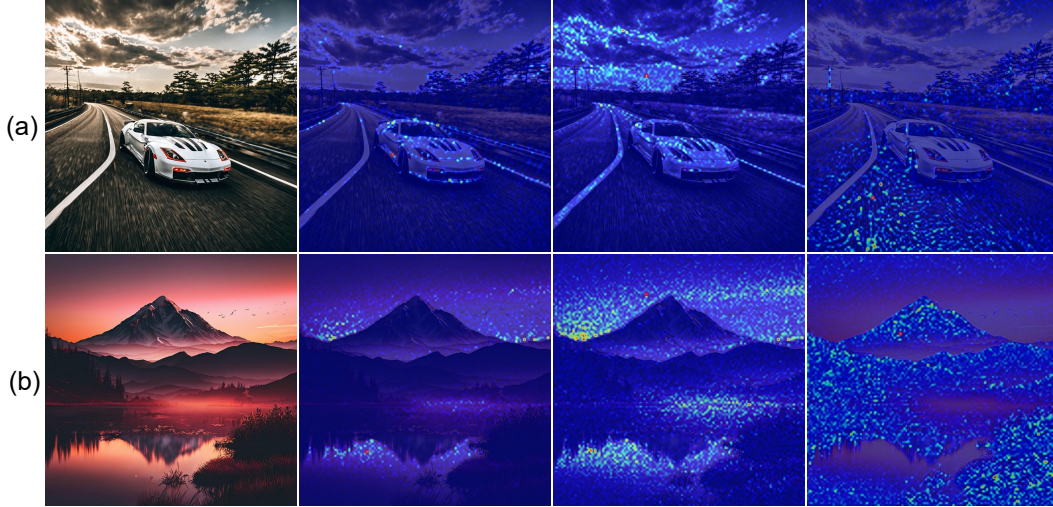


Figure 17: **Visualization of attention maps in TFSA.** The query tokens are highlighted with red boxes, and the heatmap color ranges from blue to red, indicating increasing correlation strength between the key tokens and the query tokens. Resolution: 1024×1024 . Zoom-in for a better view.

tokens correlated with the query token at the selected car location are related not only to the car itself (*i.e.*, the concept of the car) but also to its color. TFSA leverages such correlations to fuse token information, thereby accelerating the formation of the overall image layout.

B Supplementary Qualitative Comparison of §4.3

Fig. 18 presents additional qualitative comparison results. MultiDiffusion continues to struggle with maintaining global consistency; as indicated by the red boxes, DemoFusion tends to produce repetitive content, a problem somewhat alleviated in AccDiffusion but not fully resolved. As highlighted by the black boxes, another issue with AccDiffusion is the presence of noticeable streak artifacts in the images.

C Supplementary Ablation Experiments of §5

C.1 Further Qualitative Analysis of Attention Guidance

Fig. 19 provides additional qualitative ablation results on attention guidance. Individual preferences for contrast, color vividness, and detail richness may vary. attention guidance allows users to adjust parameters such as the guidance scale to synthesize images according to their preferences.

C.2 Ablation on the hyper-parameters of Attention Guidance

Quantitative analysis of guidance scale. We sampled 1k prompts, fixed $\eta_1 = 0.06$, $\eta_2 = [0.2]$ and performed ablation studies for guidance scale γ . The quantitative results are shown in Table 5. Considering all metrics, we find that $\gamma = 0.004$ achieved better quantitative results.

Quantitative analysis of delay rate. We sampled 1k prompts, fixed $\gamma = 0.004$, $\eta_2 = [0.2]$ and performed ablation studies for delay rate η_1 . Table 6 presents the experimental results, indicating that better results can be achieved when $\eta_1 = 0.06$. This means that appropriately delaying the effect of attention guidance can further enhance the quality of the generated images.

C.3 Ablation on Progressive Scheduler Value

This section presents the results of quantitative ablation analysis on the progressive scheduler η_2 in the second stage of RepLDM. We fixed $\gamma = 0$, $\eta_1 = 0$, sampled 500 prompts, and generated 1k images to investigate the optimal value of the progressive scheduler. Table 7 presents the quantitative results, indicating that using an excessively large progressive scheduler may lead to a decline in image quality.

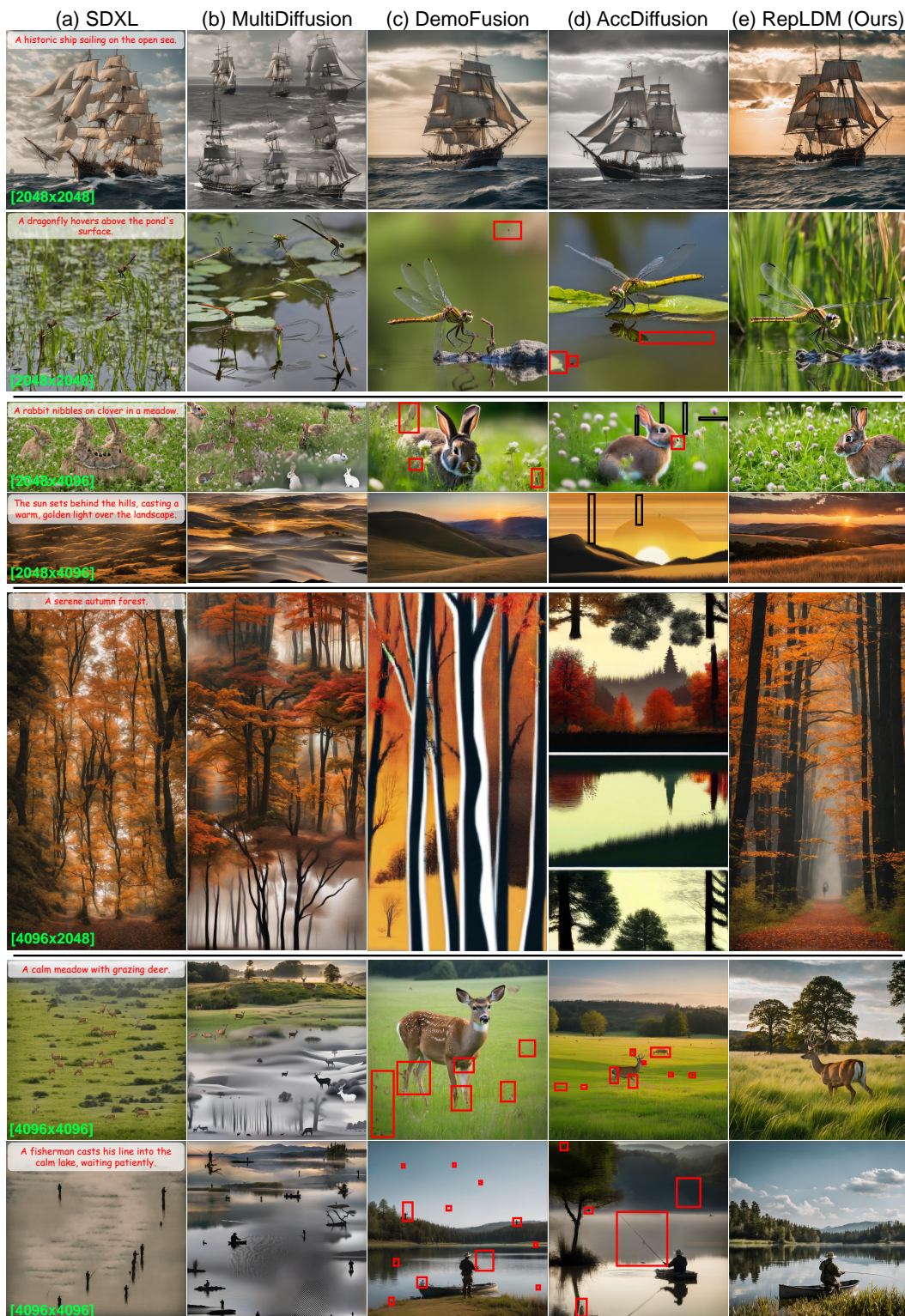


Figure 18: **Qualitative comparison with other baselines.** Zoom-in for a better view.



Figure 19: **Further qualitative analysis of attention guidance (AG).** Using attention guidance significantly enhances image quality. The details were enriched, for example: the clouds in the sky, ripples on the water, reflections on the lake, and even the expressions in a person’s eyes. Resolution: 2048×2048 . Best viewed **ZOOMED-IN**.

Table 5: **Quantitative ablation experiments on the guidance scale γ .** The best results are marked in **bold**, and the second best results are marked by underline.

Method	1024×1024					1600×1600					2048×2048				
	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑
$\gamma = 0.000$	90.85	58.18	21.21	17.69	25.09	90.91	54.74	21.45	15.41	24.93	91.78	59.08	21.57	17.36	24.86
$\gamma = 0.001$	90.50	58.04	21.34	16.76	25.08	91.17	54.31	21.19	<u>15.47</u>	24.93	91.40	58.75	21.87	15.85	24.86
$\gamma = 0.002$	89.82	57.54	<u>21.28</u>	<u>17.04</u>	25.08	90.39	53.71	21.26	15.00	24.97	90.81	<u>58.34</u>	21.45	17.16	24.90
$\gamma = 0.003$	90.10	<u>57.08</u>	20.80	16.61	25.08	90.56	<u>53.95</u>	<u>21.35</u>	15.46	24.98	90.87	58.40	21.47	17.60	24.92
$\gamma = 0.004$	89.40	56.64	20.96	16.63	25.09	89.91	54.23	20.91	15.54	25.01	90.11	58.11	21.18	16.78	24.94s
$\gamma = 0.005$	90.17	57.50	20.89	16.34	25.12	<u>90.24</u>	55.19	20.67	15.21	25.02	90.46	58.91	20.79	16.87	24.97
$\gamma = 0.006$	<u>89.79</u>	58.18	20.33	15.93	25.16	90.36	56.71	20.33	14.59	25.06	<u>90.32</u>	59.86	20.37	16.12	25.00
$\gamma = 0.007$	90.42	60.29	20.07	16.20	25.21	90.91	59.35	20.36	14.16	25.12	90.86	61.81	20.14	15.70	25.06
$\gamma = 0.008$	91.64	63.63	19.66	14.25	25.25	91.98	63.93	19.13	13.71	<u>25.13</u>	92.16	64.82	19.59	14.24	<u>25.08</u>
$\gamma = 0.009$	94.29	67.87	19.15	13.00	<u>25.25</u>	94.38	70.21	19.45	12.12	25.16	94.39	68.84	19.22	13.63	25.12

Table 6: **Quantitative ablation experiments on the delay rate η_1 .** The best results are marked in **bold**, and the second best results are marked by underline.

Method	1024×1024					1600×1600					2048×2048				
	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑
$\eta_1 = 0.00$	89.98	58.29	20.74	16.48	25.06	90.89	55.54	21.00	14.42	24.98	90.75	59.41	20.54	16.99	24.91
$\eta_1 = 0.02$	89.96	57.67	20.99	16.87	25.05	90.76	54.77	21.08	15.35	24.95	91.78	59.08	21.57	18.16	24.86
$\eta_1 = 0.04$	89.47	57.28	20.98	16.63	25.07	90.22	54.14	20.86	15.43	24.98	90.52	58.47	20.76	17.02	24.91
$\eta_1 = 0.06$	<u>89.44</u>	56.64	20.92	16.58	25.11	<u>89.91</u>	54.23	20.91	15.54	25.01	90.11	58.11	21.18	16.78	24.94
$\eta_1 = 0.08$	89.95	56.97	21.05	16.76	25.09	89.87	54.10	21.22	<u>15.65</u>	24.98	90.74	58.45	20.99	17.06	24.92
$\eta_1 = 0.10$	89.29	<u>56.88</u>	21.11	<u>16.84</u>	25.09	89.97	53.99	21.04	15.37	<u>24.99</u>	90.41	58.45	20.99	17.12	<u>24.92</u>
$\eta_1 = 0.12$	89.84	57.32	21.05	16.58	25.08	90.00	53.85	21.24	15.81	24.93	<u>90.24</u>	58.45	<u>21.24</u>	<u>17.36</u>	24.90
$\eta_1 = 0.14$	89.85	57.12	20.91	16.40	<u>25.09</u>	90.06	<u>53.83</u>	<u>21.33</u>	15.62	24.99	90.69	<u>58.25</u>	21.17	16.74	24.91
$\eta_1 = 0.16$	90.06	57.28	<u>21.10</u>	16.53	25.09	90.91	54.74	21.45	15.41	24.93	90.76	58.37	20.97	16.87	24.91
$\eta_1 = 0.18$	90.16	57.29	20.88	15.10	25.08	90.26	53.79	21.06	15.07	24.97	90.78	58.33	21.05	17.21	24.90

D Ablation on the Attention Guidance Components

D.1 Ablation on the Guidance Scale Decay Strategy

To investigate the impact of different guidance scale decay strategies, we conduct ablation studies using two additional schemes—linear decay and exponential decay—and analyze their quantitative and qualitative performance. For quantitative ablation, we generate 2k samples at a resolution of 2048×2048 using each strategy and calculate the criteria on the SAM benchmark. Table 8 shows that different strategies yield similar results, indicating that RepLDM is not sensitive to a specific decay strategy. Fig. 20 illustrates the qualitative results. Qualitatively, these decay strategies also produce similar visual experience.

Table 7: **Quantitative ablation study of the progressive scheduler Value.** The best results are marked in **bold**, and the second best results are marked by underline.

Method	1600 × 1600					2048 × 2048				
	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑
SDXL	101.56	25.78	73.67	21.23	26.87	112.64	18.44	79.03	20.61	26.55
$\eta_2 = [0.9]$	94.59	27.04	67.60	23.01	26.97	97.14	24.48	64.34	22.14	26.59
$\eta_2 = [0.8]$	93.13	28.80	65.67	24.83	26.99	93.93	26.75	60.84	23.27	26.77
$\eta_2 = [0.7]$	92.05	29.44	65.35	24.97	27.07	92.50	28.17	57.34	24.05	26.93
$\eta_2 = [0.6]$	92.94	30.79	64.57	24.29	27.11	91.86	30.45	55.38	24.96	26.98
$\eta_2 = [0.5]$	<u>92.73</u>	30.65	63.43	24.26	27.13	<u>91.80</u>	<u>31.18</u>	54.32	24.48	27.02
$\eta_2 = [0.4]$	93.04	<u>30.96</u>	63.33	24.77	27.14	91.71	32.47	53.72	25.16	27.03
$\eta_2 = [0.3]$	92.93	30.91	63.09	24.84	27.15	92.39	30.72	<u>53.32</u>	26.63	27.07
$\eta_2 = [0.2]$	93.09	31.17	<u>63.23</u>	25.71	<u>27.17</u>	92.71	30.45	53.19	<u>26.19</u>	<u>27.12</u>
$\eta_2 = [0.1]$	93.44	30.69	63.75	<u>25.18</u>	27.22	92.94	30.69	53.77	24.71	27.18

Table 8: **Ablation on the guidance scale decay strategies.** The best results are marked in **bold**, and the second best results are marked by underline.

Strategies	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑
Linear	<u>66.2</u>	<u>21.5</u>	<u>47.2</u>	20.3	25.4
Exponential	66.8	21.8	47.0	16.3	<u>25.3</u>
Cosine (default)	66.0	21.0	47.4	<u>17.5</u>	25.1



Figure 20: **Qualitative ablation on guidance scale decay strategies.**

D.2 Ablation on the Attention Calculation Paradigm

For TFSA, our objective is to remove the learnable parameters from the Self-Attention mechanism, while maintaining its computational paradigm as unchanged as possible. In TFSA, Q , K , and V are identical. Therefore, TFSA is a totally symmetric formula. As analyzed before, this paradigm encourages the clustering of semantically related tokens, and finally leads to finer details and richer colors. An interesting question arises: if we spatially downsample Q , K , or V before applying TFSA and reformulate it into an asymmetric paradigm (denoted as TFSA-A), would TFSA-A encourage the model to attend more explicitly from fine details to coarse structures?

To answer this question, we design an asymmetric variants, TFSA-A. Specifically, TFSA-A performs a 2×2 pooling operation to downsample the K and V matrices before the attention calculation operation, ensuring that the output of $\text{Softmax}(QK^T/\sqrt{d})V$ remains the of shape $(hw) \times c$. Table 9 shows that TFSA-A produces comparable quantitative results. In Fig. 21, we observe that although TFSA-A achieves quantitative results comparable to those of TFSA, its visual quality is significantly inferior. In fact, TFSA-A tends to reduce image details. This aligns with our hypothesis: the 2×2 pooling acts as a low-pass filter, causing the loss of fine-grained information in the latent representations and leading the model to focus more on low-frequency structures.

E Further Model Efficiency Analysis

Computational complexity analysis of TFSA. Note that attention guidance is only applied during the first stage of generation. Assume we have a HR image x_0 with a resolution of $H \times W \times C$. we

Table 9: **Ablation on the attention calculation paradigm.** The best results are marked in **bold**, and the second best results are marked by underline.

Paradigm	FID ↓	IS _c ↑	FID _c ↓	IS _c ↑	CLIP ↑
w/o guidance	<u>66.8</u>	<u>21.6</u>	<u>47.5</u>	17.4	25.3
w/ TFSA-A	67.4	22.6	47.9	20.4	25.3
w/ TFSA	66.0	21.0	47.4	<u>17.5</u>	<u>25.1</u>

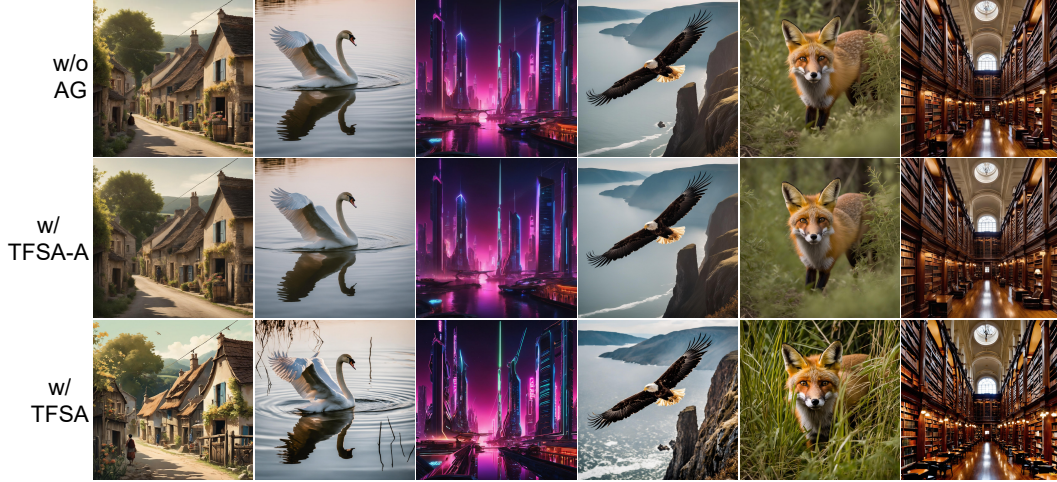


Figure 21: **Ablation on the attention calculation paradigm.** Resolution: 2048×2048 .

encode the image x_0 into latent space and obtain latent representation $z_0 \in \mathbb{R}^{h \times w \times c}$. Before feeding z_0 into TFSA, we reshape it to a $(hw) \times c$ matrix. The computation of TFSA follows a formulation similar to that of self-attention: $\text{Softmax}(z_0 z_0^T / \sqrt{c})z$. Thus, the computational complexity of TFSA is $O((hw)^2 c)$. Taking SDXL as an example, the training resolution is $H = 1024$, $W = 1024$. After VAE encoding, $c = 4$, $h = H/8 = 128$, $w = W/8 = 128$. For each denoising step, the FLOPs of TFSA is approximately $2 \times (h \times w)^2 \times c$, which is around 2.15 GFLOPs—negligible compared to the FLOPs of the denoising network (several TFLOPs per step).

How does pixel space upsampling accelerate generation? To answer this question, we analyze the time consumption of each component in DemoFusion and RepLDM when generating images at the resolution of 4096×4096 .

Table 10: **The time consumption of DemoFusion when generating 4096×4096 resolution images.**

Metric	Denoise 1024	Denoise 2048	Denoise 3072	Denoise 4096	Decode 4096	Total
number of steps	50	50	50	50	-	200
Time (s)	12	185	480	901	106	1684

Table 11: **The time consumption of RepLDM when generating 4096×4096 resolution images.** The intermediate encoding/decoding operations are highlighted in underline.

Metric	Denoise 1024	Decode 1024	Encode 3304	Denoise 3304	Decode 3304	Encode 4096	Denoise 4096	Decode 4096	Total
number of steps	50	-	-	5	-	-	10	-	65
Time (s)	12	<u>0</u>	<u>12</u>	20	<u>64</u>	<u>11</u>	118	106	343

Table 10 shows that denoising at high resolutions is a time-consuming process. DemoFusion requires substantial generation time because it performs the full denoising process at high resolutions. Note that, compared with the cost of the denoising process at high resolutions, the costs of encoding and decoding are negligible. Table 11 shows that RepLDM significantly accelerates generation by substantially reducing the number of denoising steps at high resolutions. This is because RepLDM performs pixel space upsampling through multiple rounds of encoding and decoding, producing high-quality low-resolution images that serve as better initialization. As a result, RepLDM can significantly reduce the number of sampling steps required for HR generation, thereby accelerating the process. Moreover, Table 11 shows that the additional overhead from multiple intermediate encoding and decoding operations is also relatively minor compared to the total generation cost.

Further efficiency comparison across different models. To provide a more comprehensive assessment of model efficiency, we further report the NFE and FLOPs of different models when generating a single image at resolutions of 2048×2048 and 4096×4096 . Tables 12 and 13 show

that RepLDM significantly reduces the NFE and FLOPs required for inference by decreasing the number of denoising steps at high resolutions, thereby substantially reducing the time needed to generate HR images.

Table 12: Inference cost of generating a 2048×2048 Image for different models.

Model	SDXL [36]	MultiDiff. [1]	ScaleCrafter [14]	HiDiff. [56]	UG [21]	DemoFusion [6]	AccDiff. [28]	RepLDM
NFE	50	50	50	50	80	100	100	60
TFLOPs	3010	5420	2437	1857	3608	9015	8597	1140
Time (min)	1.0	3.0	1.0	0.8	1.8	3.0	3.0	0.6

Table 13: Inference cost of generating a 4096×4096 Image for different models.

Model	SDXL [36]	MultiDiff. [1]	ScaleCrafter [14]	HiDiff. [56]	UG [21]	DemoFusion [6]	AccDiff. [28]	RepLDM
NFE	50	50	50	50	80	200	200	65
TFLOPs	12026	29566	9759	5211	12624	72167	74225	7140
Time (min)	8.0	15.0	19.0	3.4	11.1	25.0	26.0	5.7

Qualitative analysis on the progressive upsampling stage. To clearly illustrate the progressive upsampling process of RepLDM, we set $\eta_2 = [0.2, 0.2, 0.2]$ to generate 4096×4096 images. As shown in Fig. 22, the images generated at different sub-stages of RepLDM 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. As discussed in Table 10 and Table 11, 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 [6, 28]. 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 [44, 53], RepLDM leverages the prior knowledge of low-frequency structures in low-resolution images, thereby effectively accelerating the generation process.

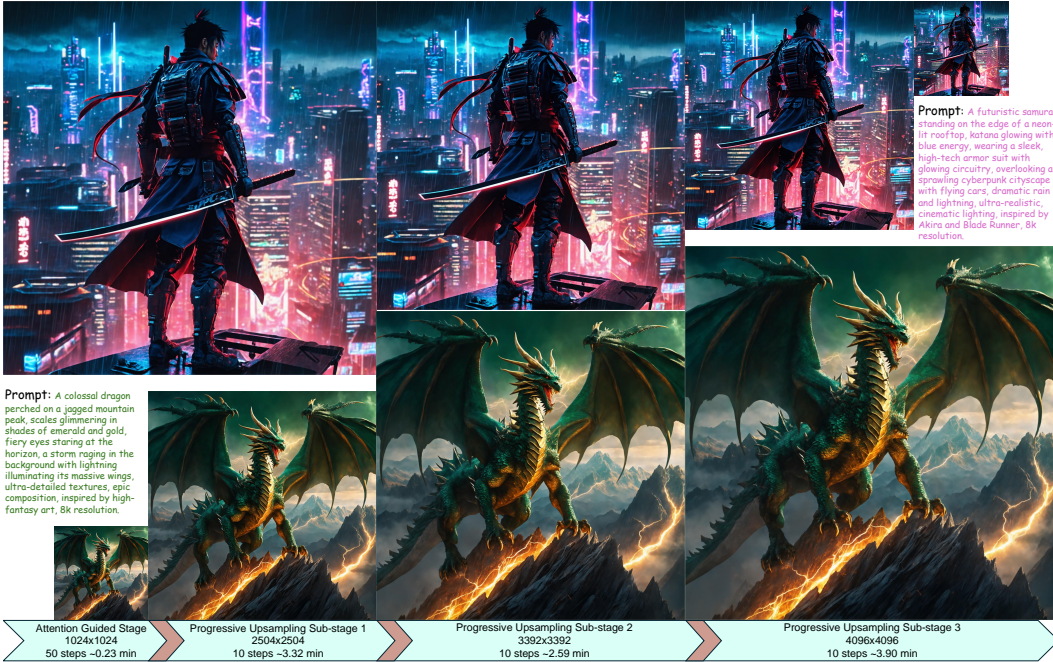


Figure 22: Illustration of the progressive upsampling generation process. The inference speed is evaluated on a single 3090 GPU.

F RepLDM Algorithm

The implementation details of RepLDM can be found in Algorithm 1, and further information is available in our code repository.

Algorithm 1 RepLDM Inference Pipeline

Require: The number of inference time steps of the first stage T_0 ; progressive scheduler η_2 ; attention guidance scale γ ; attention guidance delay rate η_1 ; the decay factor β ; target image size tuple (H', W') ; the denoising model \mathcal{F} ; denoising model's training resolution tuple (H, W) ; VAE encoder \mathcal{E} ; VAE decoder \mathcal{D} ; noise scheduler's hyper-parameter list $\bar{\alpha}_{1:T_0}$.

```
1: Initialization:
2:  $\mathbf{z}_{T_0}^{(0)} = \epsilon \sim \mathcal{N}(0, \mathbf{I})$  {Sampling from Standard Gaussian Distribution}
3:  $n_{\text{stages}} = \text{length}(\eta_2) + 1$  {Get the total number of denoising stages}
4:  $r' = \frac{H'}{W'}$  {Keep the aspect ratio and number of pixels unchanged}
5:  $H^{(0)} = \text{ceil}(\sqrt{H \times W \times r'})$ 
6:  $W^{(0)} = \text{ceil}(\sqrt{\frac{H \times W}{r'}})$ 
7:  $H^{(n)} = H'$ 
8:  $W^{(n)} = W'$ 
9:  $\text{area}_{\text{list}} = \text{linspace}(H^{(0)} \times W^{(0)}, H^{(n)} \times W^{(n)}, n_{\text{stages}})$  {Upsampling according to the number of pixels}
10:  $H_{\text{list}} = [\text{ceil}(\sqrt{i \times r'}) \text{ for } i \text{ in } \text{area}_{\text{list}}]$  {Get the height and width of each stage}
11:  $W_{\text{list}} = [\text{ceil}(\sqrt{i/r'}) \text{ for } i \text{ in } \text{area}_{\text{list}}]$ 
12:  $k_{\text{denoising}} = [T_0]$  {Get the number of denoising steps for each stage}
13:  $k_{\text{denoising}}.\text{extend}([i \times T_0 \text{ for } i \text{ in } \eta_2])$ 
14:  $k = T_0 \times \eta_1$  {Obtain the number of delay steps}
15:  $\gamma_{\text{list}} = [\gamma(\frac{\cos(\frac{T-k-i}{T-k}\pi)+1}{2})^\beta \text{ for } i = 1, \dots, T-k]$  {Obtain the guidance scale for each step}
16: Denoising:
17: for  $s = 0, \dots, n_{\text{stages}} - 1$  do
18:    $n_{\text{steps}} \leftarrow k_{\text{denoising}}[s]$ 
19:   if  $s \geq 1$  then
20:      $\mathbf{x}^{(s)} \leftarrow \text{upsample}(\mathbf{x}^{(s-1)}, H_{\text{list}}[s], W_{\text{list}}[s])$  {Upsampling in pixel space}
21:      $\mathbf{z}_0^{(s)} \leftarrow \mathcal{E}(\mathbf{x}^{(s)})$ 
22:      $\mathbf{z}_{n_{\text{steps}}}^{(s)} \sim \mathcal{N}(\sqrt{\bar{\alpha}[n_{\text{steps}}]}\mathbf{z}_0^{(s)}, (1 - \bar{\alpha}[n_{\text{steps}}])\mathbf{I})$ 
23:   end if
24:   for  $t = n_{\text{steps}} - 1, \dots, 0$  do
25:      $\mathbf{z}_t^{(s)} \leftarrow \mathcal{F}(\mathbf{z}_{t+1}^{(s)}, t+1)$  {Denoising}
26:     if  $s == 0$  and  $t \leq T - 1 - k$  then
27:        $\mathbf{z}_t^{(s)} \leftarrow \gamma_{\text{list}}[t]\text{PFSA}(\mathbf{z}_t^{(s)}) + (1 - \gamma_{\text{list}}[t])\mathbf{z}_t^{(s)}$  {Attention Guidance}
28:     end if
29:   end for
30:    $\mathbf{x}^{(s)} \leftarrow \mathcal{D}(\mathbf{z}_0^{(s)})$  {Obtain the pixel space image}
31: end for
```

G Robustness Analysis

In this section, we conduct a robustness analysis to complement the experiments in §4.2, providing a more comprehensive evaluation of the models' performance. Our robustness analysis is conducted from two perspectives: (i) we vary the random seeds and repeat each experiment three times to compute the mean and standard deviation of all results; (ii) we randomly sample 20k HR images from the HR subset of LAION-5B dataset [41] to construct a new benchmark for evaluating the models' generalization performance. Since HR generation requires substantial computational resources, we analyze the four best-performing models from Table 1, *i.e.*, HiDiffusion, DemoFusion, AccDiffusion, and RepLDM.

Analysis on the SAM benchmark. We maintain the exact experimental settings as in §4.2 and conduct the analysis at resolutions of 2048×2048 and 4096×4096 . Table 14 shows that RepLDM continues to exhibit superior performance across the repeated experiments.

Table 14: **Robustness analysis on the SAM benchmark.** The best results are marked in **bold**.

Method	2048 × 2048					4096 × 4096				
	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑	FID ↓	IS ↑	FID _c ↓	IS _c ↑	CLIP ↑
HiDiff. [56]	80.29±0.57	17.18±0.40	63.55±0.63	15.26±0.76	24.95±0.04	144.24±0.84	12.71±0.14	146.62±0.32	7.48±0.28	21.18±0.05
DemoF. [6]	71.89±0.60	22.10±0.37	53.58±0.22	19.21±0.27	25.21±0.01	101.83±0.49	20.81±0.11	63.60±0.46	14.92±1.24	24.75±0.03
AccDiff. [28]	71.37±0.48	21.21±0.32	53.04±0.33	19.24±1.72	25.13±0.01	102.41±1.40	19.88±0.24	65.86±0.17	12.73±0.71	24.65±0.02
RepLDM	66.08±0.02	22.13±0.74	47.31±0.11	20.38±2.03	25.30±0.12	91.46±0.61	21.63±0.46	58.93±0.20	15.02±0.16	24.62±0.02

Analysis on the LAION-5B benchmark. Considering that only 1K samples were used for the 4096×4096 resolution in §4.2, which may lead to unstable metric evaluations, we double the number of samples to 2k for this resolution in the current experiment. Regarding evaluation metrics, since IS may lead to high variances beyond ImageNet, we follow some recent studies and adopt Kernel Inception distance (KID) for more accurate evaluation [20, 37]. Table 15 shows that on the LAION benchmark, RepLDM still demonstrates superior performance, surpassing previous SOTA models across all metrics.

Table 15: **Robustness analysis on the LAION-5B benchmark.** The best results are marked in **bold**. Since the magnitude of KID is relatively small, we multiply its mean and standard deviation by 10^3 .

Method	2048 × 2048					4096 × 4096				
	FID ↓	KID ↓	FID _c ↓	KID _c ↓	CLIP ↑	FID ↓	KID ↓	FID _c ↓	KID _c ↓	CLIP ↑
HiDiff. [56]	48.17±0.41	8.06±0.20	36.26±0.37	10.93±0.11	23.16±0.03	92.81±0.78	35.36±0.60	120.26±0.91	103.45±0.27	18.55±0.06
DemoF. [6]	34.15±0.31	4.50±0.05	21.38±0.17	6.80±0.06	25.44±0.02	37.03±0.27	5.71±0.14	30.77±0.36	16.12±0.22	25.12±0.04
AccDiff. [28]	34.49±0.31	4.92±0.08	22.71±0.17	8.57±0.11	24.90±0.02	38.56±0.23	7.21±0.20	38.85±0.29	20.87±0.20	24.46±0.01
RepLDM	34.08±0.25	4.18±0.04	20.30±0.30	4.87±0.13	25.78±0.03	34.01±0.26	4.13±0.05	23.08±0.26	12.08±0.13	25.88±0.04

H Comparative and Ablation Analysis Based on StableDiffusion 2.1

H.1 Comparison Experiments

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

Qualitative comparison. Fig. 23 presents the results of the qualitative comparison. It can be observed that, when generating high-resolution images, SD2.1 also faces issues with repetitive object structures. ScaleCrafter often exhibits structural collapse during denoising with SD2.1, resulting in suboptimal performance. In contrast, RepLDM consistently produces high-quality results across all resolutions, highlighting the generalizability of the RepLDM generation framework.

Quantitative comparison. Since the code for using SD2.1 as the pretrained model in AccDiffusion and DemoFusion is not publicly available, we compare RepLDM with ScaleCrafter. 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 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 16 presents the results of the quantitative comparison, showing that RepLDM maintains strong performance when using SD2.1 as the pre-trained model. In contrast, ScaleCrafter performs suboptimally, as it tends to produce structural collapse in the generated images, a phenomenon that is more apparent in the qualitative analysis.

H.2 Ablation Study on Attention Guidance

Quantitative ablation. Table 17 shows the results of the quantitative ablation on attention guidance using SD2.1 as the pretrained model. It can be observed that attention guidance leads to improvements in metrics. These improvements are more evident in the qualitative ablation analysis.

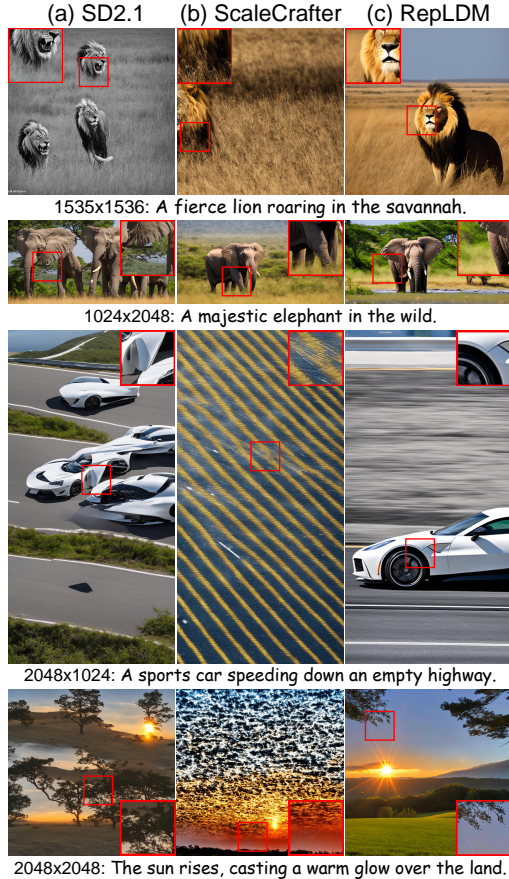


Figure 23: **Qualitative comparison using SD2.1 as the pretrained model.**

Table 16: **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 [39]	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 [14]	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
RepLDM	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 17: **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

Qualitative ablation. Fig. 24 presents the ablation analysis of attention guidance based on SD2.1. From the figure, it can be observed that attention guidance also enhances detail richness and color vibrancy when using SD2.1, further demonstrating its generalization capability.

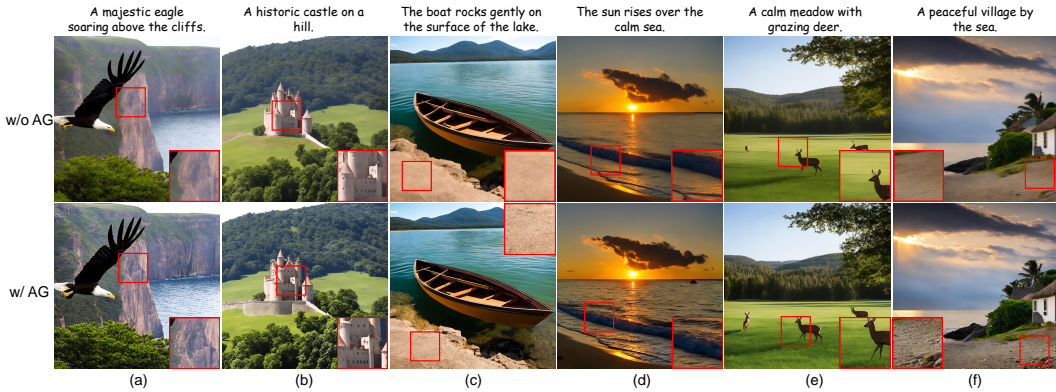


Figure 24: **Ablation study of attention guidance using SD2.1 as the pre-trained model.** Resolution: 2048×2048 .

I Attention Guidance Also Works in Other Generation Frameworks

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

I.1 Quantitative Ablation

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

Table 18 presents the quantitative ablation results using the HiDiffusion framework. It is evident that incorporating attention guidance improves HiDiffusion across all metrics. This is further corroborated by the qualitative analysis in Fig. 25, which demonstrates that attention guidance alleviates some of the structural collapses observed in HiDiffusion.

Table 18: **Quantitative ablation of attention guidance using HiDiffusion frameworks.** The best results are marked in **bold**.

Method	FID	IS	FID _c	IS _c	CLIP
HiDiffusion [56]	81.0	16.8	64.1	14.2	24.9
HiDiffusion+AG	79.4	17.0	62.4	14.6	24.9

I.2 Qualitative Ablation

HiDiffusion+attention guidance. We incorporate attention guidance into the generative framework of HiDiffusion. Fig. 25 (a)-(c) demonstrate that using attention guidance effectively mitigates the issue of structural collapse in synthesized images. Fig. 25 (d)-(f) further show that attention guidance can also address the structural deformation inherent to HiDiffusion, enhance image details, and improve overall image quality.

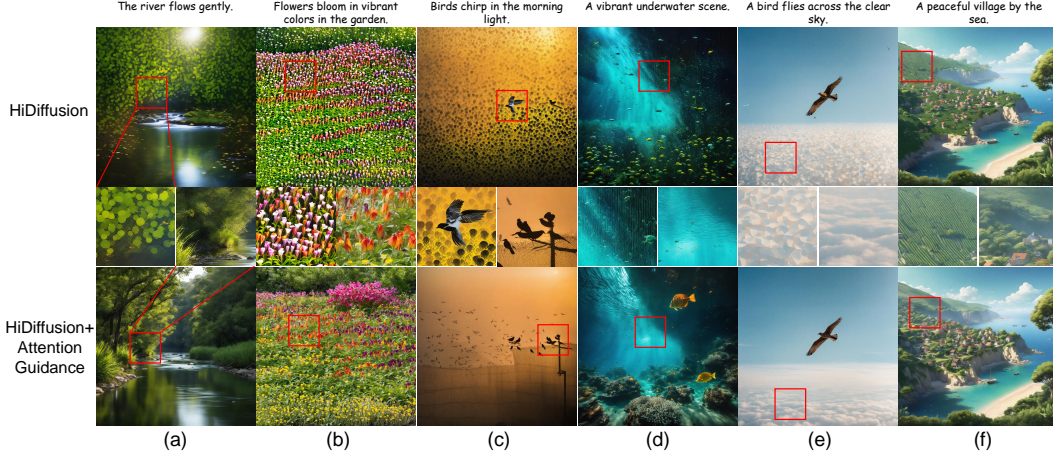


Figure 25: **Qualitative ablation of attention guidance in the HiDiffusion Framework.** All images have a resolution of 2048×2048 . Figures (a)-(c) demonstrate that attention guidance can mitigate the issue of structural collapse in generated images, while Figures (d)-(f) show that attention guidance resolves structural deformation issues and enhances image details.

DemoFusion+attention guidance. We incorporate attention guidance into the generative framework of DemoFusion. As shown in Fig. 26 (a)-(c), attention guidance effectively mitigates the issue of repetitive structures in DemoFusion. Fig. 26 (d)-(f) further illustrate role of attention guidance in enriching image details and enhancing overall image quality.

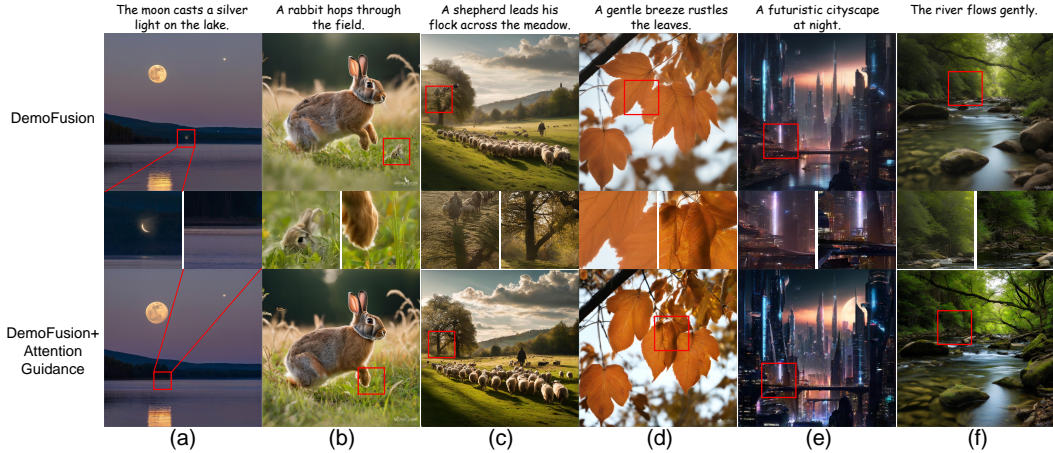


Figure 26: **Qualitative ablation of attention guidance in the DemoFusion Framework.** All images have a resolution of 2048×2048 . Figures (a)-(c) demonstrate that attention guidance effectively mitigates the issue of repetitive structures in images, while Figures (d)-(f) showcase attention guidance’s ability to enrich image details.

J Super-Resolved Images Tend to Lack High-Resolution Details

To explain why using super-resolution models to obtain HR images is sub-optimal, in this section, we conduct both qualitative and quantitative comparisons between RepLDM and the super-resolution results. Specifically, we use BSRGAN [54] to upsample the generated results of SDXL [36] at its training resolution.

Quantitative results. As shown in table. 19, the super-resolution model (SDXL + BSRGAN) demonstrate comparable 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, quantitative metrics such as FID and IS, are widely recognized as insufficient for comprehensively evaluating the performance of model’s generation. As a result, user studies are commonly employed to provide human-level

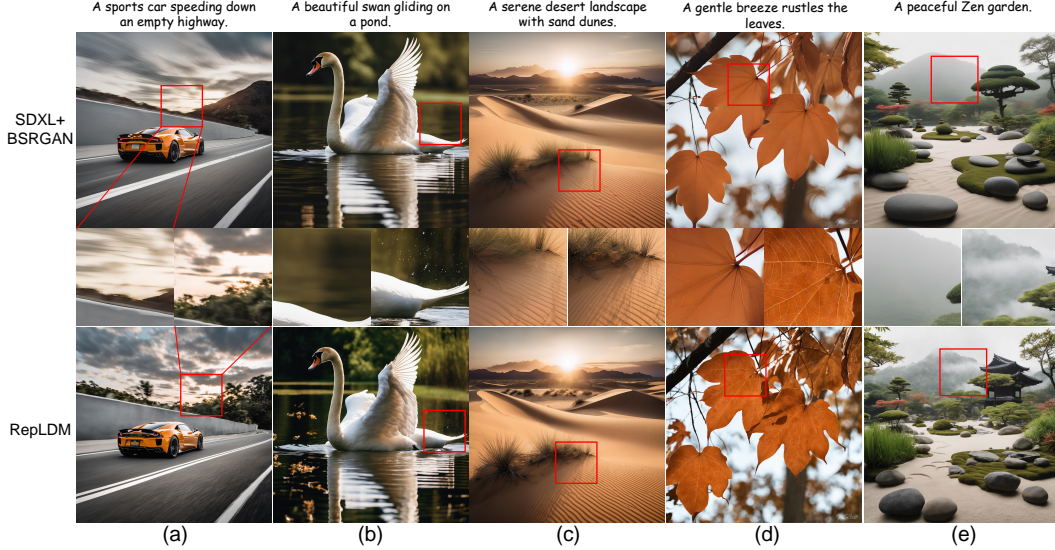


Figure 27: **Qualitative comparison with SDXL+BSRGAN.** The prompts for the generated images are provided above the figures. The resolution of (a) to (c) are 2048×2048 , and the resolution of (d) and (e) are 4096×4096 .

evaluation with more intuition [11, 14, 22, 30, 36, 39, 55]. For example, in ScaleCrafter [14], they conducted both quantitative and user study analyses in comparison with the SD+SR approach. Their results show that, although ScaleCrafter performs worse than SD+SR on quantitative metrics, users significantly prefer the textures and details generated by ScaleCrafter. One important reason is that the goal of the SR model is to produce images consistent with the input, which limits its performance in high-resolution generation – needing more detail for true high-resolution visuals beyond simple smoothing [6, 10, 14, 22, 27, 28, 57].

Table 19: **Quantitative comparison results between RepLDM and SDXL+BSRGAN.** The best results are marked in **bold**.

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+BSRGAN	66.2	21.1	47.5	16.6	25.7	80.7	19.8	50.2	12.3	25.1	92.7	17.6	57.9	12.1	24.9	90.0	20.9	56.0	13.8	25.2
RepLDM	66.0	21.0	47.4	17.5	25.1	89.0	20.3	56.0	19.0	25.0	93.2	19.5	56.9	16.5	24.9	90.6	21.1	59.0	14.8	24.6

Qualitative results. As shown in Fig. 27, compared to RepLDM, 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. 27 (c), SDXL+BSRGAN fails to generate realistic shadows.

K Memory Usage Analysis

We compare the 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 20 shows the resource consumption of different models when generating images at various resolutions.

Table 20: **Model Memory Usage (GB).** The best results are marked in **bold**, and the second best results are marked by underline.

Resolutions	2048 × 2048	2048 × 4096	4096 × 4096
SDXL [36]	15.9	16.1	<u>16.6</u>
MultiDiff. [1]	22.0	16.8	16.8
ScaleCrafter [14]	17.4	17.6	19.1
UG [21]	23.9	16.5	18.0
DemoFusion [6]	15.2	18.4	16.8
AccDiff. [28]	22.1	23.0	22.1
HiDiff. [56]	23.9	<u>16.2</u>	16.2
RepLDM	16.0	21.1	23.8

It is worth noting that for HR 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 HR 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).

It is important to note that different models have undergone varying degrees of additional optimization in their official open-source implementations. Specifically, some open-source codes utilize existing optimization tools, such as accelerate [9] or Flash Attention [3], which provide additional advantages in terms of inference speed and memory usage performance. To ensure a fair comparison, in Table 20, we did not use such additional optimizations in the implementation of RepLDM.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations in §6, analyze their underlying causes, and discuss potential directions for future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide a detailed description of the methodology and its underlying ideas in §3. In addition, we present the full algorithmic pipeline using pseudocode in Appendix F. We also commit to open-sourcing the code of our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In Appendix F, we detail the method using pseudocode. Additionally, upon acceptance, we will clean and release our code base and share it on GitHub. All data used in this paper belongs to existing open source datasets and have been correctly cited to ensure reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The proposed method requires no training. We provide a detailed explanation of the inference and evaluation settings in §4.1. We determine the hyperparameters through ablation studies, with details provided in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We repeated the experiments in Table 1 in Appendix G, computing the mean and standard deviation, and also conducted additional replication experiments on the LAION dataset [41].

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report the resources needed to reproduce the experiments in §4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We conform to the NeurIPS code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: the proposed method builds upon a pretrained generative model to produce higher-resolution images in a training-free manner, and thus does not introduce any additional or specific societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our method does not rely on any specific publicly released model and requires no specialized fine-tuning, and therefore does not necessitate additional safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all used resources such as implementations of baselines and data. We release our work with CC-By 4.0 license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This study does not involve the release of any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: In the user study conducted in this work, we recruited volunteers to evaluate the quality of the generated images. Detailed instructions were provided to the participants. As the participants were volunteers, no compensation was involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: This study only required volunteers to evaluate the quality of generated images, and therefore poses no particular potential risks. Furthermore, to protect the privacy of participants' preferences, all responses were anonymized and randomized, and no personal information was collected.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.