Supplementary Material for "RectifiedHR: Enable Efficient High-Resolution Synthesis via Energy Rectification"

1. Supplementary

1.1. Implementation details

Although a limited number of samples may lead to lower values for metrics such as FID [5], we follow prior protocols and randomly select 1,000 prompts from LAION-5B [12] for text-to-image generation. Evaluations are conducted using 50 inference steps, empty negative prompts, and fixed random seeds.

We employ four widely used quantitative metrics: Fréchet Inception Distance (FID) [5], Kernel Inception Distance (KID) [1], Inception Score (IS) [11], and CLIP Score [9]. FID and KID are computed using `pytorch-fid`, while CLIP Score and IS are computed using `torchmetrics`. The subscript r refers to resizing high-resolution images to 299×299 before evaluation, whereas the subscript c indicates that 10 patches of size 1024×1024 are randomly cropped from each generated high-resolution image and then resized to 299×299 for evaluation. Specifically, FID_r, KID_r, and IS_r require resizing images to 299×299. However, such an evaluation is not ideal for high-resolution image generation. Following prior works [3, 8], we randomly crop 10 patches of size 1024×1024 from each generated high-resolution image to compute FID_s, KID_c, and IS_c.

1.2. User study details

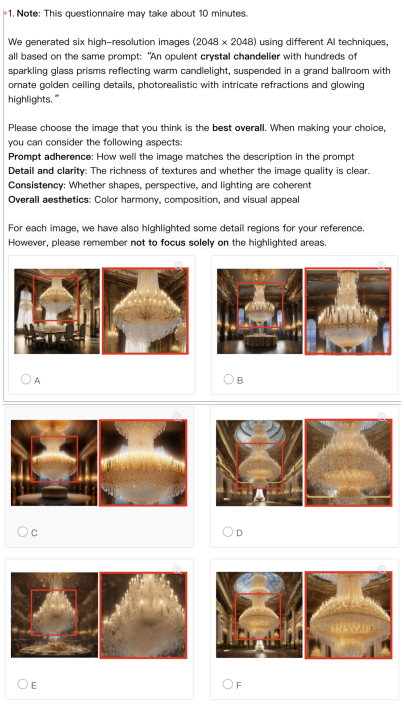

Figure 1. The interface of one question in the user study

We conducted a user study to further demonstrate the effectiveness of our method. We selected 15 images in total, evenly distributed across three resolutions: 2048×2048, 4096×4096, and 2048×4096 (five images per resolution). 30 participants were involved in the study, where they were asked to evaluate the images provided and identify the best. The questionnaire is designed on the https://www.wjx.cn/ platform. The interface of the questionnaire is shown in Fig. 1.

The baselines in this study are consistent with those in Sec. 1.9, except for direct inference and DemoFusion. Direct inference was excluded because most of its generated images exhibited severe global distortions. The outputs of AccDiffusion and DemoFusion are highly similar under a fixed random seed. As [8] has quantitatively demonstrated the superiority of AccDiffusion, we retained AccDiffusion solely for conciseness in this study.

Fig. 2 shows the results of the user study. Our method (RectifiedHR) received 32.2% of the total votes, significantly exceeding the other competing methods. The second most selected method, FreCaS, accounted for only 16.2%, which is approximately half of RectifiedHR's proportion. The remaining methods, including AccDiffusion (13.8%), ScaleCrafter (13.6%), HiDiffusion (12.7%), and FouriScale (11.5%), received relatively lower proportions of the total votes. These results demonstrate that more users are inclined to identify RectifiedHR as the best compared to existing approaches, validating the effectiveness of our method in subjective evaluation.

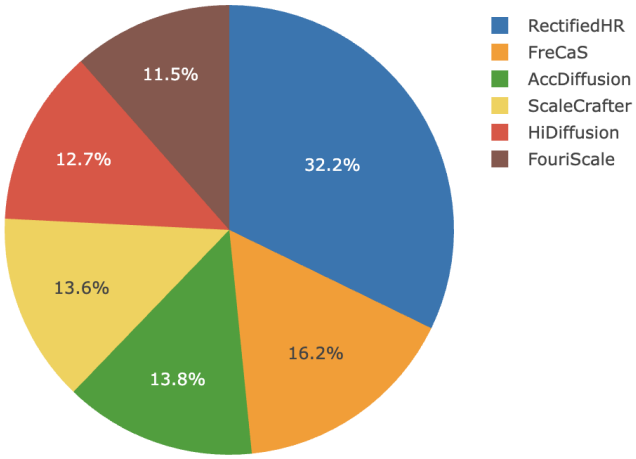

Figure 2. The results of the user study

1.3. Quantitative Analysis of "Predicted x_0"

To quantitatively validate this observation, as shown in Fig.3, we conduct additional experiments on the generation of $p^t_{x_0}$ using 100 random prompts sampled from LAION-5B [12], and analyze the CLIP Score [4] and Mean Squared Error (MSE). From Fig. 3a, we observe that after 30 denoising steps, the MSE between $p^t_{x_0}$ and $p^{t-1}_{x_0}$ exhibits minimal change. In Fig. 3b, we find that the CLIP score between $p^t_{x_0}$ and the corresponding prompt increases slowly beyond 30 denoising steps.

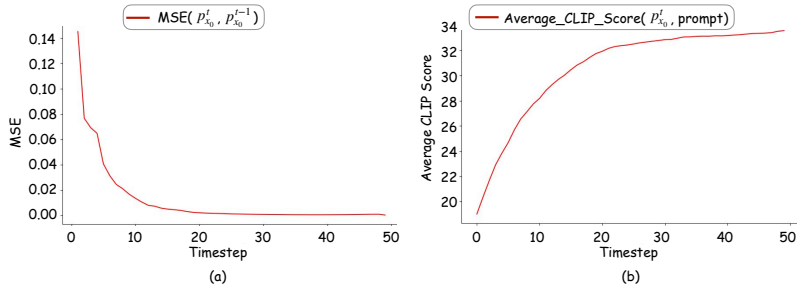

Figure 3. The trend of the "predicted x_0" at different timesteps t, denoted as $p^t_{x_0}$, evaluated on 100 random prompts. (a) The average MSE between $p^t_{x_0}$ and $p^{t-1}_{x_0}$. The x-axis represents the sampling timestep, and the y-axis denotes the average MSE. It can be observed that after approximately 30 steps, the rate of change in $p^t_{x_0}$ slows significantly. (b) The trend of the average CLIP Score between $p^t_{x_0}$ and the prompt across different timesteps. The x-axis represents the sampling timestep, and the y-axis denotes the average CLIP Score.

1.4. The connection between energy rectification and Signal-to-Noise Ratio (SNR) correction

In the proof presented in this section, all symbols follow the definitions provided in the Method section of the main text. Any additional symbols not previously defined will be explicitly specified. This proof analyzes energy variation using the DDIM sampler as an example. The sampling formulation of DDIM is given as follows:

$$\begin{aligned} x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}\left(x_t, t\right)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\cdot\tilde{\epsilon}\left(x_t, t\right) \\ &= \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t + \left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\tilde{\varepsilon}\left(x_t, t\right). \end{aligned} \tag{1}$$

To simplify the derivation, we assume that all quantities in the equation are scalar values. Based on the definition of average latent energy in Eq.8 of the main text, the average latent energy during the DDIM sampling process can be expressed as follows:

$$\begin{aligned} \mathbb{E}[x_{t-1}^2] = \mathbb{E}\left[\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t\right]^2 + \mathbb{E}\left[\left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\tilde{\varepsilon}\left(x_t, t\right)\right]^2 \\ +2\times\mathbb{E}\left[\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t\right]\times\mathbb{E}\left[\left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\tilde{\varepsilon}\left(x_t, t\right)\right]. \end{aligned} \tag{2}$$

We assume that the predicted noise $\tilde{\epsilon}$ follows a standard normal distribution, such that $\mathbb{E}\left[\tilde{\epsilon}\left(x_t, t\right)\right] = 0$. Under this assumption, the average latent energy of the DDIM sampler can be simplified as:

$$\mathbb{E}[x_{t-1}^2] = \mathbb{E}\left[\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t\right]^2 + \mathbb{E}\left[\left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\tilde{\varepsilon}\left(x_t, t\right)\right]^2. \tag{3}$$

Several previous works [6, 7, 14, 18] define the Signal-to-Noise Ratio (SNR) at a given timestep of a diffusion model as follows:

$$SNR_t = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}. \tag{4}$$

Several works [6, 7, 14, 18] have observed that the SNR must be adjusted during the generation process at different resolutions. Suppose the diffusion model is originally designed for a resolution of $H \times W$, and we aim to extend it to generate images at a higher resolution of $H' \times W'$, where $H' > H$ and $W' > W$. According to the derivations in [14, 18], the adjusted formulation of α_t is given as follows:

$$\bar{\alpha}'_t = \frac{\bar{\alpha}_t}{\gamma - (\gamma-1)\bar{\alpha}_t}. \tag{5}$$

Here, the value of γ is typically defined as $(H'/H \cdot W'/W)^2$. By substituting the modified $\bar{\alpha}'_t$ into Eq. 1, we obtain the SNR-corrected sampling formulation as follows:

$$\begin{aligned} \mathbb{E}[x_{t-1}] &= \sqrt{\frac{\bar{\alpha}'_{t-1}}{\bar{\alpha}'_t}}\mathbb{E}[x_t] + \left(\sqrt{1-\bar{\alpha}'_{t-1}} - \frac{\sqrt{\bar{\alpha}'_{t-1}}\sqrt{1-\bar{\alpha}'_t}}{\sqrt{\bar{\alpha}'_t}}\right)\mathbb{E}[\tilde{\epsilon}\left(x_t, t\right)] \\ &= \sqrt{\frac{\frac{\bar{\alpha}_{t-1}}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}}{\frac{\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_t}}}\mathbb{E}[x_t] + \left(\sqrt{1-\frac{\bar{\alpha}_{t-1}}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}} - \sqrt{\frac{\frac{\bar{\alpha}_{t-1}}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}\left(1-\frac{\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_t}\right)}{\frac{\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_t}}}\right)\mathbb{E}[\tilde{\epsilon}\left(x_t, t\right)] \\ &= \sqrt{\frac{\gamma-(\gamma-1)\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}}\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\mathbb{E}[x_t] + \sqrt{\frac{\gamma}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}}\left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\mathbb{E}[\tilde{\varepsilon}\left(x_t, t\right)]. \end{aligned} \tag{6}$$

The average latent energy under SNR correction can be derived as follows:

$$\begin{aligned} \mathbb{E}[x_{t-1}^2] &= \mathbb{E}\left[\sqrt{\frac{\bar{\alpha}'_{t-1}}{\bar{\alpha}'_t}}x_t\right]^2 + \mathbb{E}\left[\left(\sqrt{1-\bar{\alpha}'_{t-1}} - \frac{\sqrt{\bar{\alpha}'_{t-1}}\sqrt{1-\bar{\alpha}'_t}}{\sqrt{\bar{\alpha}'_t}}\right)\tilde{\epsilon}\left(x_t, t\right)\right]^2 \\ &= \frac{\gamma-(\gamma-1)\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}\mathbb{E}\left[\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}x_t\right]^2 + \frac{\gamma}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}\mathbb{E}\left[\left(\sqrt{1-\bar{\alpha}_{t-1}} - \frac{\sqrt{\bar{\alpha}_{t-1}}\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\right)\tilde{\epsilon}\left(x_t, t\right)\right]^2. \end{aligned} \tag{7}$$

Compared to the original energy formulation Eqa. 3, two additional coefficients appear: $\frac{\gamma-(\gamma-1)\bar{\alpha}_t}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}$ and $\frac{\gamma}{\gamma-(\gamma-1)\bar{\alpha}_{t-1}}$. Since $\bar{\alpha}_{t-1}$ and $\bar{\alpha}_t$ are very close, the first coefficient is approximately equal to 1. In the DDIM sampling formulation, $\bar{\alpha}_t$ is within the range [0, 1], which implies that the second coefficient falls within [1, γ]. As a result, after the SNR correction, the average latent energy increases. Therefore, SNR correction essentially serves as a mechanism for energy enhancement. In this sense, both energy rectification and SNR correction aim to increase the average latent energy. However, since our method allows for the flexible selection of hyperparameters, it can achieve superior performance.

1.5. Applying *RectifiedHR* to Stable Diffusion 3

Model:SD3	CLIP-Score↑	DEQA-score↑
Direct-Inference	0.275	3.311
RectifiedHR	**0.289**	**3.621**

Table 1. The quantitative results of SD3.

To validate the effectiveness of our method on a transformer-based diffusion model, we apply it to `stable-diffusion-3-medium` using the `diffusers` library. As shown in Tab. 1, we provide additional quantitative results on SD3 (50 images, 2048×2048), and the test results mainly include CLIP-Score [4] and DEQA-Score [16].

1.6. Ablation results on hyperparameters

In this section, we conduct ablation experiments on the hyperparameters in Eq.7 and Eq.9 of the main text using SDXL. The baseline hyperparameter settings follow those described in the Evaluation Setup section of the main text. We vary one hyperparameter at a time while keeping the others fixed at the two target resolutions to evaluate the impact of each parameter on performance, as defined in Eq.7 and Eq.9 of the main text. The evaluation procedure for FID_c, FID_r, IS_c, and IS_r follows the protocol outlined in Sec. 1.1.

In Eq.7 and Eq.9 of the main text, ω_{min} and T_{max} are fixed and do not require ablation. The value of N in both equations is kept consistent. For the 2048×2048 resolution scene, with N set to 2, variations in M_T and M_ω do not significantly affect the results. Thus, only N, ω_{max}, and T_{min} are ablated. The quantitative ablation results for the 2048×2048 resolution are shown in Fig. 4, Fig. 5, and Fig. 6. For the 4096×4096 resolution scene, N, ω_{max}, T_{min}, M_T, and M_ω are ablated. The corresponding quantitative ablation results for the 4096×4096 resolution are presented in Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11. Based on these results, it can be concluded that the basic numerical settings used in this experiment represent the optimal solution.

In Eq.7 and Eq.9 of the main text, ω_{min} and T_{max} are fixed and thus excluded from ablation. The value of N is kept consistent across both equations. For the 2048×2048 resolution setting, with N set to 2, variations in M_T and M_ω have minimal impact on performance. Therefore, only N, ω_{max}, and T_{min} are subject to ablation. The corresponding quantitative ablation results are shown in Fig. 4, Fig. 5, and Fig. 6. For the 4096×4096 resolution setting, we ablate N, ω_{max}, T_{min}, M_T, and M_ω. The corresponding results are presented in Fig. 7, Fig. 8, Fig.9, Fig.10, and Fig. 11. Based on these findings, we conclude that the default numerical settings used in our experiments yield the optimal performance.

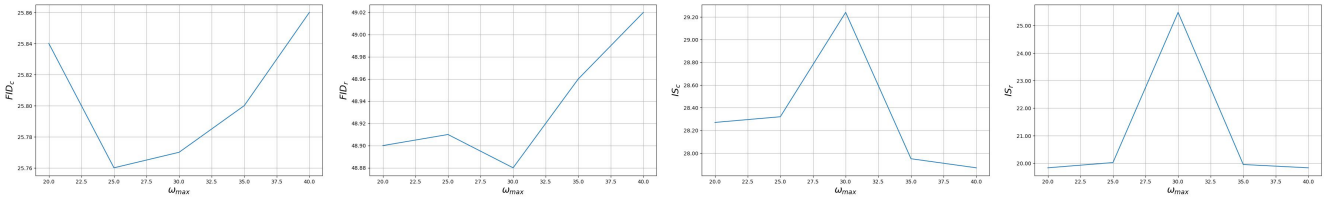

Figure 4. The image illustrates the ablation study of ω_{max} in Eq.9 of the main text for the 2048×2048 resolution setting. The values of ω_{max} range over $20, 25, 30, 35, 40$.

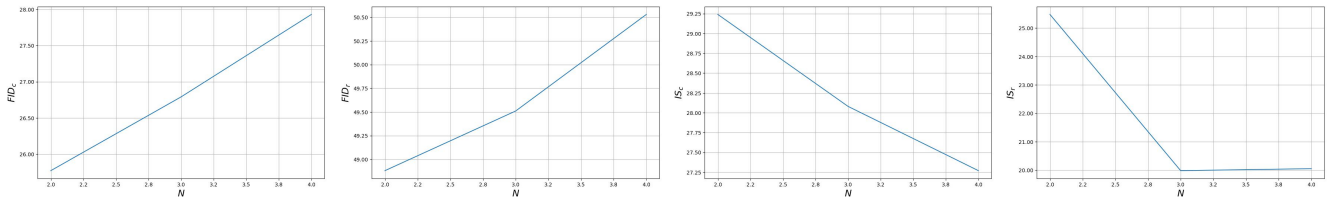

Figure 5. The image illustrates the ablation study of N in Eq.7 and Eq.9 of the main text for the 2048×2048 resolution setting. The values of N range over $2, 3, 4$.

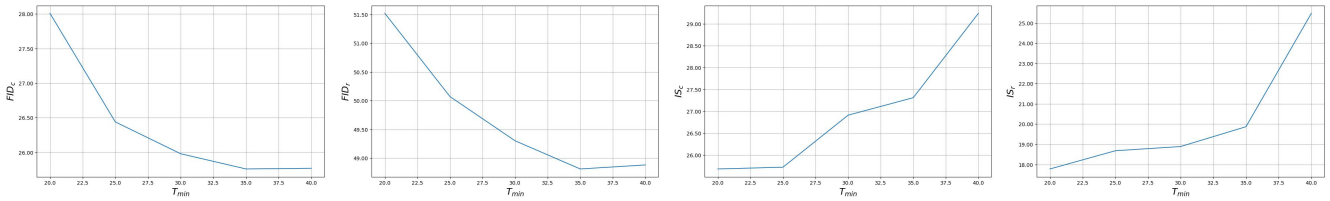

Figure 6. The image illustrates the ablation study of T_{min} in Eq.7 of the main text for the 2048×2048 resolution setting. The values of T_{min} range over $20, 25, 30, 35, 40$.

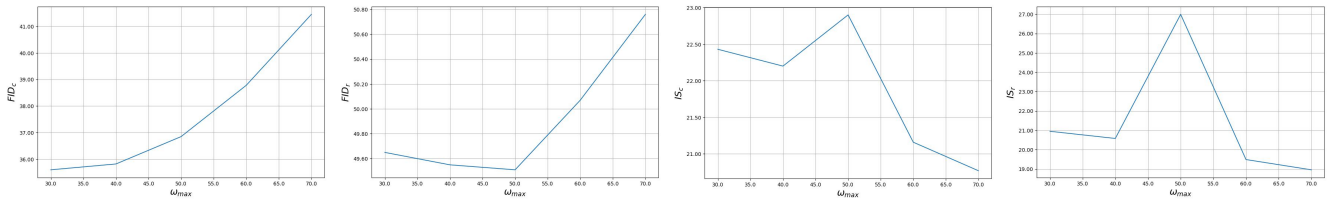

Figure 7. The image illustrates the ablation study of ω_{max} in Eq.9 of the main text for the 4096×4096 resolution setting. The values of ω_{max} range over $30, 40, 50, 60, 70$.

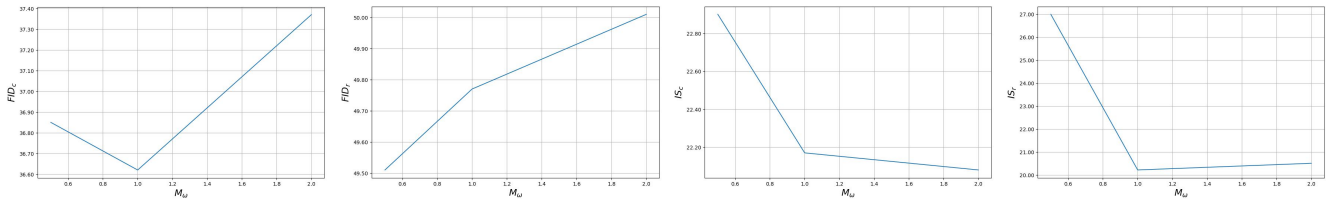

Figure 8. The image illustrates the ablation study of M_ω in Eq.9 of the main text for the 4096×4096 resolution setting. The values of M_ω range over $0.5, 1, 2$.

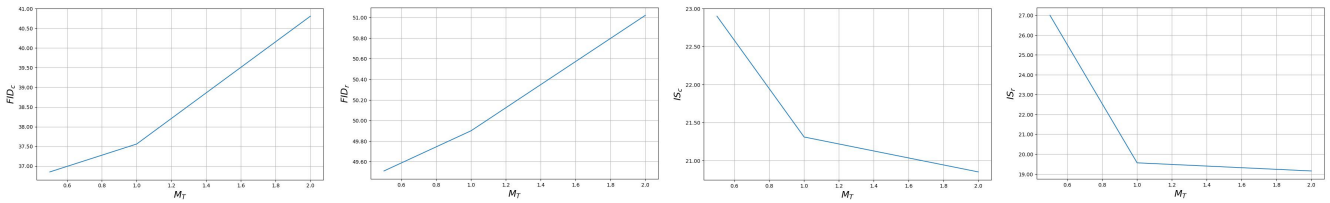

Figure 9. The image illustrates the ablation study of M_T in Eq.7 of the main text for the 4096×4096 resolution setting. The values of M_T range over $0.5, 1, 2$.

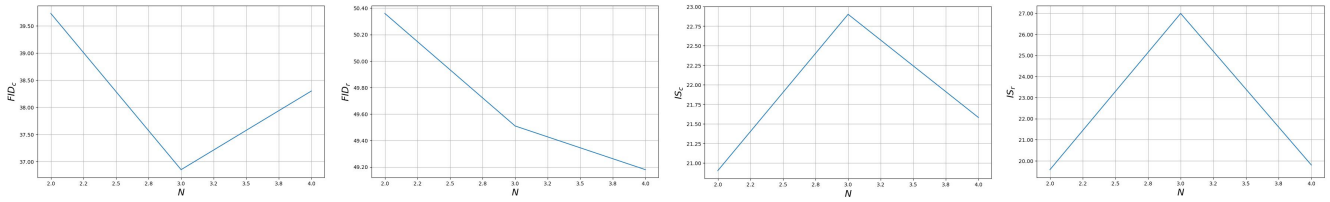

Figure 10. The image illustrates the ablation study of N in Eq.7 and Eq.9 of the main text for the 4096×4096 resolution setting. The values of N range over $2, 3, 4$.

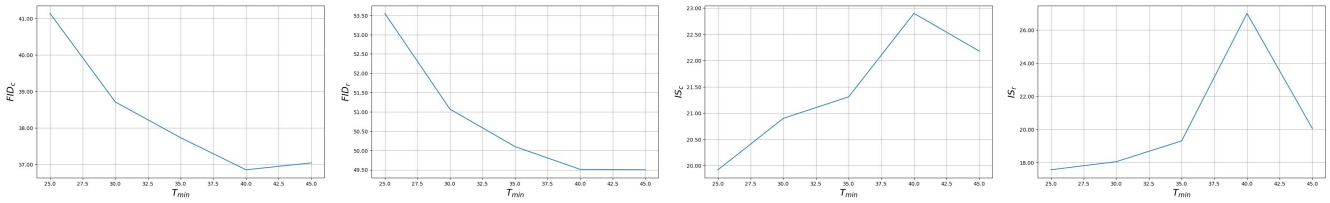

Figure 11. The image illustrates the ablation study of T_{min} in Eq.7 of the main text for the 4096×4096 resolution setting. The values of T_{min} range over $25, 30, 35, 40, 45$.

1.7. Hyperparameter details and quantitative results for applying *RectifiedHR* to applications

The combination of *RectifiedHR* and WAN. *RectifiedHR* can be directly applied to video diffusion models such as WAN [13]. The officially supported maximum resolution for WAN 1.3B is 480×832 over 81 frames. Our goal is to generate videos at 960×1664 resolution using WAN 1.3B. The direct inference baseline refers to generating a 960×1664 resolution video directly using WAN 1.3B. In contrast, *WAN+RectifiedHR* refers to using *RectifiedHR* to generate the same-resolution video. The selected hyperparameters in Eq.7 and Eq.9 of the main text are: $N = 2$, $\omega_{\text{max}} = 10$, $\omega_{\text{min}} = 5$, $T_{\text{min}} = 30$, $T_{\text{max}} = 50$, $M_T = 1$, and $M_\omega = 1$. Our quantitative experimental details follow [2] on 40 videos.

The combination of *RectifiedHR* and OIR. *RectifiedHR* can also be applied to image editing tasks. We employ SDXL as the base model and randomly select several high-resolution images from the OIR-Bench [15] dataset for qualitative comparison. Specifically, we compare two approaches: (1) direct single-object editing using OIR [15], and (2) OIR combined with *RectifiedHR*. While the OIR baseline directly edits high-resolution images, the combined method first downsamples the input to 1024×1024, performs editing via the OIR pipeline, and then applies *RectifiedHR* during the denoising phase to restore fine-grained image details. For the 2048×2048 resolution setting, the hyperparameters in Eq.7 and Eq.9 of the main text are: $N = 2$, $\omega_{\text{max}} = 30$, $\omega_{\text{min}} = 5$, $T_{\text{min}} = 40$, $T_{\text{max}} = 50$, $M_T = 1$, and $M_\omega = 1$. For the 3072×3072 resolution setting, the hyperparameters are: $N = 3$, $\omega_{\text{max}} = 40$, $\omega_{\text{min}} = 5$, $T_{\text{min}} = 40$, $T_{\text{max}} = 50$, $M_T = 1$, and $M_\omega = 1$.

The combination of *RectifiedHR* and DreamBooth. *RectifiedHR* can be directly adapted to various customization methods, where it is seamlessly integrated into DreamBooth without modifying any of the training logic of DreamBooth [10]. The base model for the experiment is SD1.4, which supports a native resolution of 512×512 and a target resolution of 1536×1536. The hyperparameters selected in Eq.7 and Eq.9 of the main text are as follows: N is 3, ω_{max} is 30, ω_{min} is 5, T_{min} is 40, T_{max} is 50, M_T is 1, and M_ω is 1. Furthermore, as demonstrated in Tab. 2, we conduct a quantitative comparison between the *RectifiedHR* and direct inference, using the DreamBooth dataset for testing. The test metrics and process were fully aligned with the methodology in [10]. It can be observed that *RectifiedHR* outperforms direct inference in terms of quantitative metrics for high-resolution customization generation.

RectifiedHR can be directly adapted to various customization methods and is seamlessly integrated into DreamBooth [10] without modifying any part of its training logic. The base model used in this experiment is SD1.4, which natively supports a resolution of 512×512, with the target resolution set to 1536×1536. The selected hyperparameters in Eq.7 and Eq.9 of the main text are as follows: $N = 3$, $\omega_{\text{max}} = 30$, $\omega_{\text{min}} = 5$, $T_{\text{min}} = 40$, $T_{\text{max}} = 50$, $M_T = 1$, and $M_\omega = 1$. Furthermore, as shown in Tab.2, we conduct a quantitative comparison between *RectifiedHR* and direct inference using the DreamBooth dataset for evaluation. The test metrics and protocol are fully aligned with the methodology described in [10]. The results demonstrate that *RectifiedHR* outperforms direct inference in terms of quantitative metrics for high-resolution customization generation.

Direct Inference	DINO ↑	CLIP-I ↑	CLIP-T ↑
DreamBooth + RectifiedHR	**0.625**	**0.761**	**0.249**
DreamBooth	0.400	0.673	0.220

Table 2. Quantitative comparison results between *RectifiedHR* and direct inference after DreamBooth training. The evaluation is conducted on a scene with a resolution of 1536×1536.

The combination of *RectifiedHR* and ControlNet. Our method can be seamlessly integrated with ControlNet [17] to operate directly during the inference stage, enabling image generation conditioned on various control signals while simultaneously enhancing its ability to produce high-resolution outputs. The base model used is SDXL. The selected hyperparameters in Eq.7 and Eq.9 of the main text are: $N = 3$, $\omega_{\text{max}} = 40$, $\omega_{\text{min}} = 5$, $T_{\text{min}} = 40$, $T_{\text{max}} = 50$, $M_T = 1$, and $M_\omega = 1$.

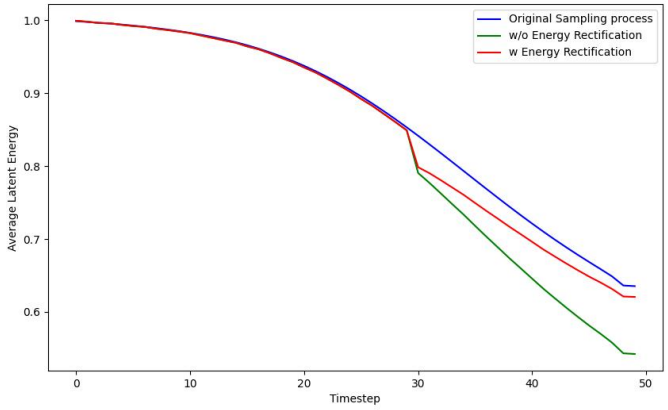

Figure 12. Visualization of the average latent energy curve following energy rectification.

1.8. Visualization of the energy rectification curve

To better visualize the average latent energy during the energy rectification process, we plot the corrected energy curves. We randomly select 100 prompts from LAION-5B for the experiments. As shown in Fig. 12, the blue line represents the energy curve at a resolution of 1024×1024. For the 2048×2048 resolution setting, we use the following hyperparameters: $T_{\text{min}} = 30$, $T_{\text{max}} = 50$, $N = 2$, $\omega_{\text{min}} = 5$, $\omega_{\text{max}} = 30$, $M_T = 1$, and $M_\omega = 1$. The red line corresponds to our method with energy rectification for generating 2048×2048 resolution images, while the green line shows the result of our method without the energy rectification module. It can be observed that energy rectification effectively compensates for energy decay.

1.9. Qualitative Results

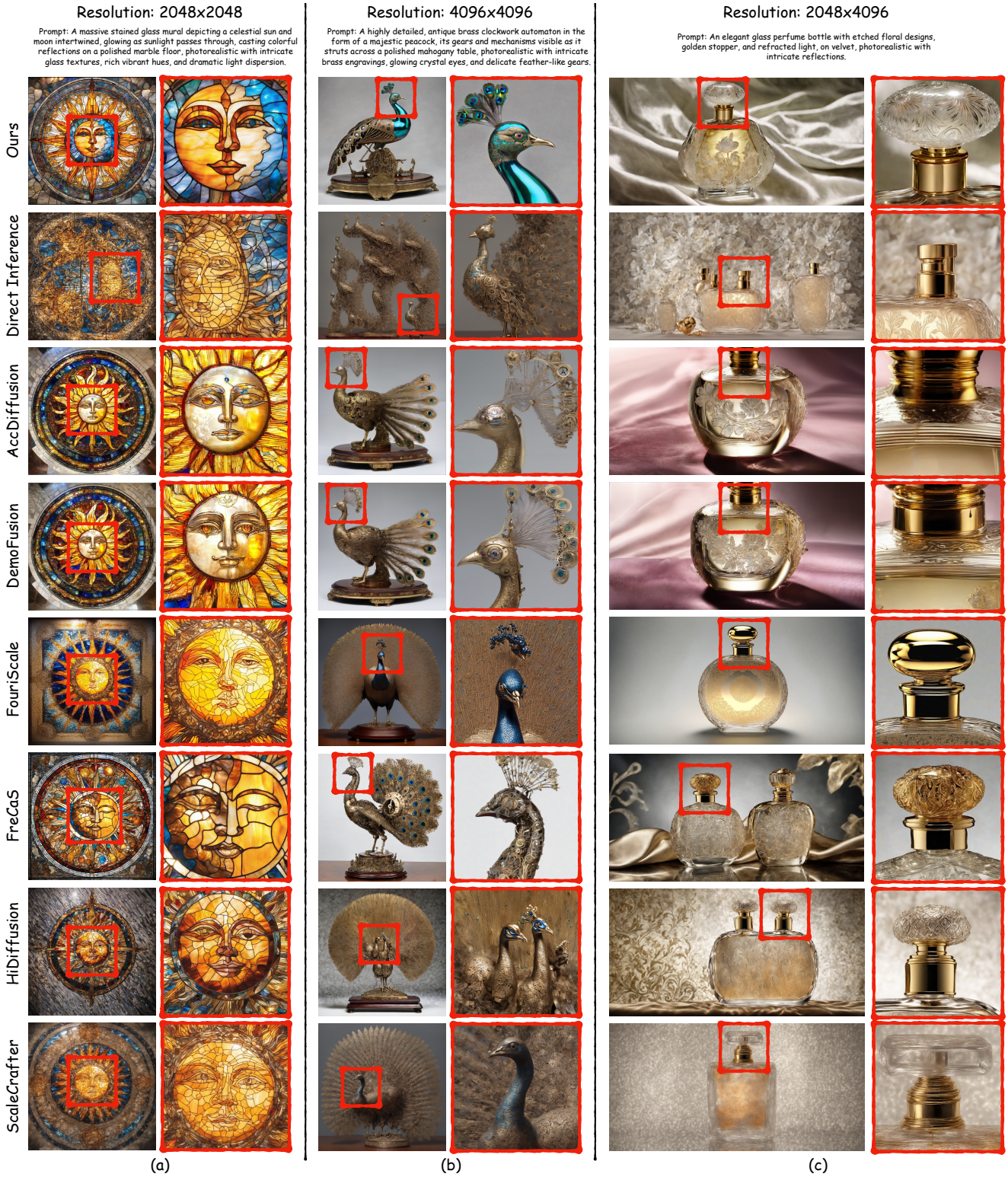

Figure 13. Qualitative comparison across three different resolutions between our method and other training-free methods. The red box indicates an enlarged view of a local region within the high-resolution image.

As shown in Fig. 13, to clearly illustrate the differences between our method and existing baselines, we select a representative prompt for each of the three resolution scenarios and conduct qualitative comparisons against SDXL direct inference, AccDiffusion, DemoFusion, FouriScale, FreCas, HiDiffusion, and ScaleCrafter. AccDiffusion and DemoFusion tend to produce blurry details and lower visual quality, such as the peacock's eyes and feathers in column b, and the bottle stoppers in column c. FouriScale and ScaleCrafter often generate deformed or blurred objects that fail to satisfy the prompt, such as feathers lacking peacock characteristics in column b, and a blurry bottle body missing the velvet element specified in the prompt in column c. HiDiffusion may introduce repetitive patterns, as seen in the duplicate heads in column b and the recurring motifs on the bottles in column c. FreCas can produce distorted details or fail to adhere to the prompt, such as the deformed and incorrect number of bottles in column c. In contrast, our method consistently achieves superior visual quality across all resolutions. In column a, our approach generates the clearest and most refined faces and is the only method that correctly captures the prompt's description of the sun and moon intertwined. In column b, our peacock is the most detailed and visually accurate, with a color distribution and fine-grained features that closely align with the prompt's reference to crystal eyes and delicate feather-like gears. In column c, our method demonstrates the highest fidelity in rendering the bottle stopper and floral patterns, and it uniquely preserves the white velvet background described in the prompt. These qualitative results highlight the effectiveness of our method in generating visually consistent, detailed, and prompt-faithful images across different resolution settings.

1.10. More Video Results

Figure 14. More video results

1.11. More Image Results

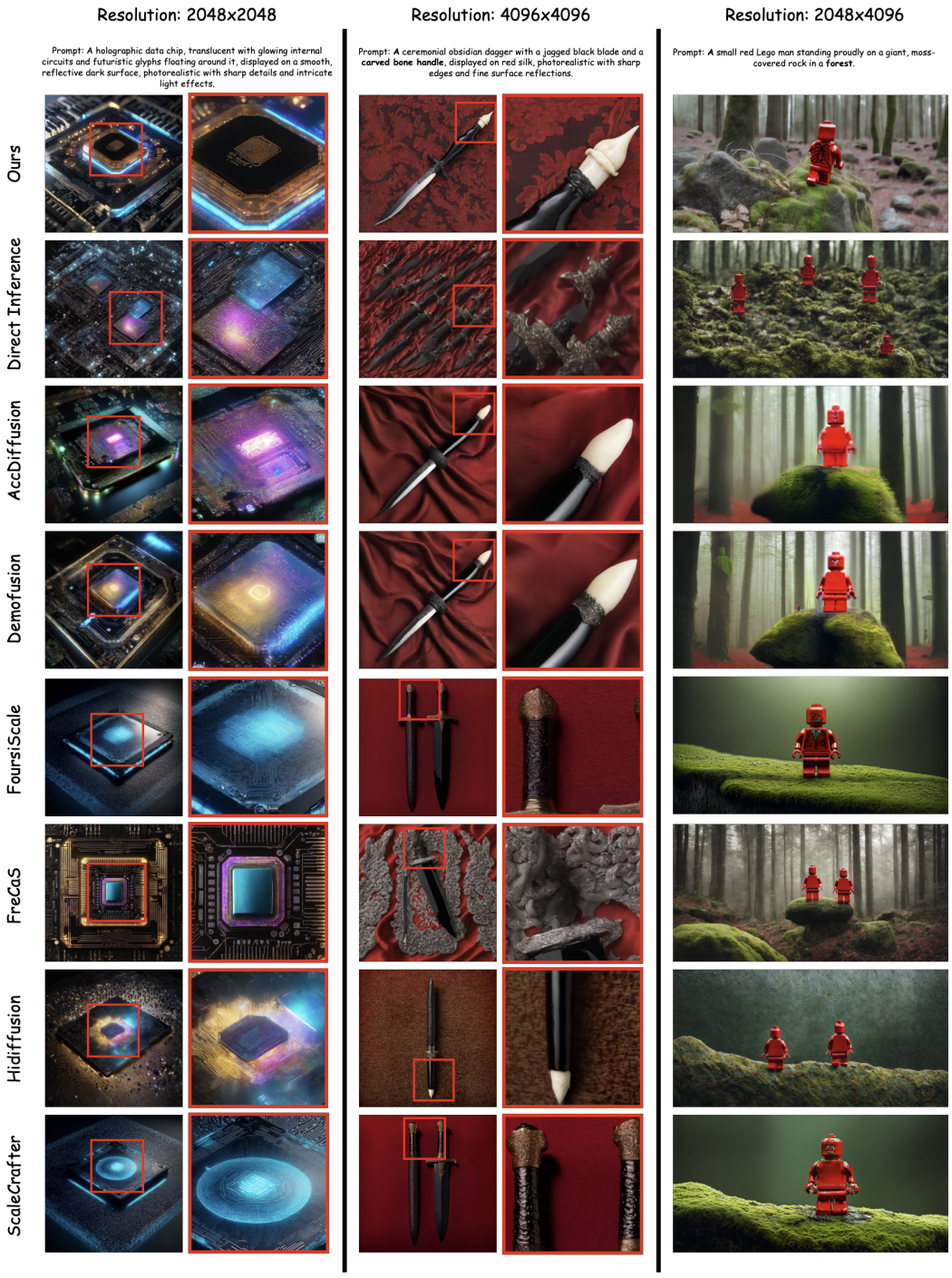

Figure 15. More image results

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 1

[2] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 6

[3] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6159–6168, 2024. 1

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. 2021. 2, 4

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[6] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. 3

[7] Juno Hwang, Yong-Hyun Park, and Junghyo Jo. Upsample guidance: Scale up diffusion models without training. *arXiv preprint arXiv:2404.01709*, 2024. 3

[8] Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*, pages 38–53. Springer, 2025. 1, 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 6

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1, 2

[13] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6

[14] Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. *arXiv preprint arXiv:2408.11001*, 2024. 3

[15] Zhen Yang, Ganggui Ding, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. *arXiv preprint arXiv:2310.12149*, 2023. 6

[16] Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14483–14494, 2025. 4

[17] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 7

[18] Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024. 3