FreeU: Free Lunch in Diffusion U-Net

Supplementary Material

In this supplementary file, we provide additional ablation studies in Section 1 and more qualitative results in Section 2. Section 3 shows more generated images from SD 1.4 [46] and SD-XL [43]. In Section 4, we conduct experiments on various diffusion models. Section 5 provides a more detailed overview of related work. We also discuss our limitations in Section 6 and the potential negative societal impacts in Section 7.

1. More Ablation Studies

1.1. The Effects of Backbone Factor

We conduct an evaluation to assess the effects of the backbone factor b. The results, presented in Fig. 17, reveal that as the backbone factor b increases, there is a noticeable enhancement in the quality of generated images. It's important to note that excessively large values of the backbone factor, such as when $b = 1.8$, can lead to oversmoothing issues. This is because increasing the backbone factor b enhances the denoising capability of U-Net, and an overly strong denoising capability compromises the preservation of high-frequency image details.

1.2. The Effects of Skip Factor

To mitigate the issue of oversmoothed textures resulting from enhanced denoising, we introduce the skip factor denoted as s. This factor is employed to selectively reduce low-frequency components within the skip features. In our evaluation, as shown in Fig. 18, we observe that as the skip factor s decreases, the generated images exhibit more detailed backgrounds, and the oversmoothing issues are mitigated. These findings demonstrate that diminishing low-frequency components within the skip features can effectively ameliorate the oversmoothing problem caused by the backbone factor. Therefore, this highlights the effectiveness of the comprehensive FreeU strategy in achieving a balance between features and alleviating issues related to texture smoothing, ultimately leading to the generation of more realistic images.

1.3. The Effects of Channel Selection

We conducted an evaluation to investigate the impact of channel selection in the backbone scaling operation. Fig.19 presents the results. In Fig.19(a), we show images generated using the standard SD approach, which serves as our baseline. Fig. 19(b), (c), (d), and (e) display images generated with the backbone scaling operation. We can observe that employing the backbone scaling operation contributes significantly to the improvement in image quality. However, as shown in Fig. 19(b), applying scaling to all channels leads to oversmoothing issues, as the enhanced U-Net compromises high-frequency image details during denoising. In contrast, as demonstrated in Fig. 19(c), (d), and (e), when we select only half of the channels for the backbone scaling operation using different methods, we observe improvements in mitigating the oversmoothing problem while enhancing image quality and preserving fine-grained image details. Importantly, these results highlight that the specific channel selection method employed has a relatively minor impact on the generated results, as all of them contribute to the enhancement of detail generation.

Table 3. Quantitative evaluation of text-to-image generation.

Method	MUSIQ-AVA ↑	LAION-Aes ↑
SD 1.4 [46]	5.231	5.365
SD 1.4 + FreeU	**5.563**	**5.532**
SD 2.1 [46]	5.432	5.503
SD 2.1 + FreeU	**5.686**	**5.612**
SD-XL [43]	5.675	5.538
SD-XL + FreeU	**5.994**	**5.776**

Table 4. Quantitative Results of FID and CLIP-score.

Method	FID↓	CLIP-sc.↑
SD-XL [43]	43.82	0.31
SD-XL + FreeU	**40.79**	**0.33**
LCM [36]	62.03	0.30
LCM + FreeU	**50.88**	**0.32**

Table 5. Quantitative evaluation of text-to-video generation.

Method	MUSIQ-AVA ↑	LAION-Aes ↑
ModelScope [37]	4.115	4.469
ModelScope + FreeU	**4.338**	**4.602**

2. More Qualitative Results

Text-to-Image. In our evaluation of FreeU, we employ provide FID [17] and CLIP-score [44], and we also follow VBench [22] to use MUSIQ image quality predictors [30] and the LAION aesthetic predictor [34] (LAION-Aes) for quantitative assessments. MUSIQ image quality predictors have been trained on KonIQ [20], AVA [40] datasets, encompassing two evaluation metrics: MUSIQ-KonIQ and

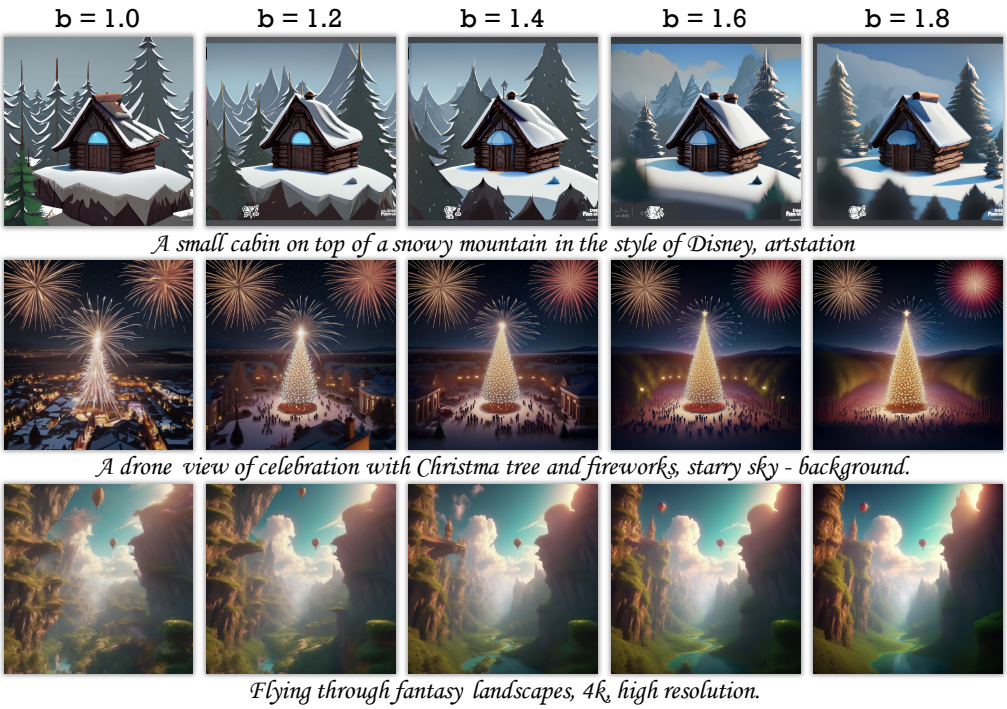

Figure 17. **The ablation study of backbone scaling factor** b**.** As the backbone factor b increases, there is a noticeable enhancement in the quality of generated images. It is noteworthy, though, that excessively large values of the backbone factor may introduce oversmoothing issues.

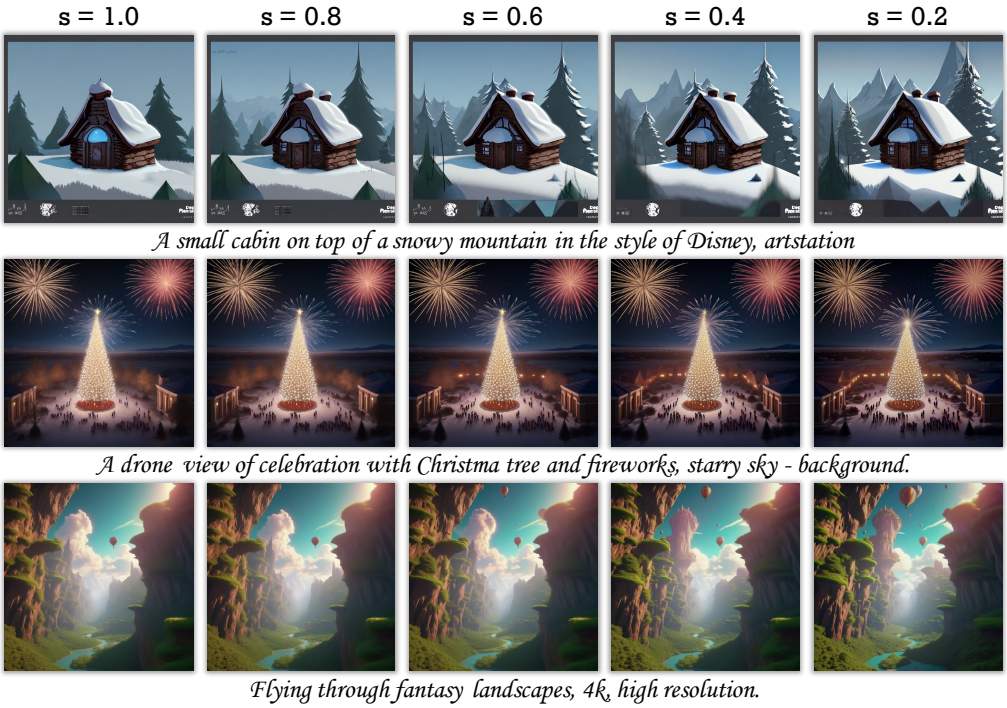

Figure 18. **The ablation study of skip scaling factor** s**.** As the skip factor s decreases, the generated images exhibit more detailed backgrounds and the oversmoothing issues are mitigated.

Figure 19. **The ablation study of channel selection for backbone scaling operation.** (a) The generated images of SD. (b) Scaling applied to all channels. (c) Scaling applied to the first half of the channels. (d) Scaling applied to the second half of the channels. (e) Scaling applied to a uniformly selected half of the channels.

MUSIQ-AVA. We follow prior work and evaluate text-to-image generation on the MS-COCO [35] validation set. Table 3 presents the results comparing FreeU with SD1.4 [46], SD2.1 [46], and SD-XL [43]. We also provide FID [17] and CLIP-score [44] results in Table 4. Notably, FreeU demonstrates improvements over these powerful models. These results highlight the effectiveness of FreeU in enhancing image quality.

Text-to-Video. We further conduct a quantitative evaluation of FreeU for the text-to-video task following a similar way as in the text-to-image task. The results, as presented in Table 5, consistently indicate that FreeU enhances the original generation ability of ModelScope [37] across both metrics.

3. More Results of Stable Diffusion

Fig. 20 and Fig. 21 show the results of SD 1.4 [46] and SD-XL [43]. The incorporation of FreeU into SD [43, 46] yields improvements in both entity portrayal and fine-grained details. For instance, as shown in Fig. 20 when provided with the prompt *"a blue car is being filmed"*, FreeU refines the image, eliminating rooftop irregularities and enhancing the textural intricacies of the surrounding structures. These compelling results prove that FreeU introduces substantial enhancements for the SD 1.4 [46] and SD-XL [43].

4. More Generative Models

To further evaluate the proposed FreeU as a foundational method, we conduct experiments on various diffusion-based methods, *e.g.* ScaleCrafter [16], Animatediff [14], ControlNet [65], LCM [36], Dreambooth [47], ReVersion [23], Rerender [61].

ScaleCrafter [16] is a training-free method designed to adapt a pre-trained diffusion model to generate images of much higher resolution than the image size used during training. As shown in Fig. 22, when FreeU is combined with ScaleCrafter [16], the resulting combination can generate 4K images using SD-XL [43]. These images exhibit superior fine-grained details and texture quality compared to those produced solely by ScaleCrafter [16]. Consequently, FreeU serves as an effective tool for enhancing the capability of ScaleCrafter [16] in generating high-quality, higher-resolution images.

Animatediff [14] represents a framework designed to convert static Text-to-Image models into generators of animated videos. In Fig. 23, a comparison is made between the videos generated by Animatediff [14], with and without the incorporation of FreeU. The results demonstrate that

FreeU significantly enhances the quality of each frame and ensures a higher level of consistency in appearance throughout the generated videos. For instance, when provided with a prompt "best quality, masterpiece, 1girl, looking at viewer, blurry background, upper body, contemporary, dress", the version augmented with FreeU generates frames with a more consistent appearance of the dress across all frames.

ControlNet [65] is a framework designed to introduce conditional controls to pre-trained text-to-image diffusion models. In our work, we have integrated FreeU into ControlNet [65]. Fig. 24 presents a comparison of the results. We observe a notable improvement in image quality and the presence of more detailed features in both the background and foreground when FreeU is employed alongside ControlNet [65]. These enhancements are particularly impressive given that the conditional image itself already contains a substantial amount of detailed information. This proves the effectiveness of FreeU in further enhancing the generative capabilities of ControlNet [65] in a conditional image synthesis setting.

LCM [36] is a highly efficient one-stage guided distillation method that enables few-step or even one-step inference on pre-trained Latent Diffusion models. The integration of the FreeU into the LCM [36] has yielded significant advancements in image generation. The comparative analysis, shown in Fig. 25, reveals that the use of FreeU alongside LCM [36] not only enhances image quality but also improves the details generations.

Dreambooth [47] is a diffusion model specialized in personalized text-to-image tasks. The enhancements are evident, as demonstrated in Fig. 26, the synthesized images present marked improvements in realism. For instance, while the base DreamBooth [47] model struggles to synthesize the appearance of the action figure's legs from the prompt *"a photo of action figure riding a motorcycle"*, the FreeU-augmented version deftly overcomes this hurdle. Similarly, for the prompt *"A toy on a beach"*, the initial output exhibited body shape anomalies. FreeU's integration refines these imperfections, providing a more accurate representation and improving color fidelity.

ReVersion [23] is a Stable Diffusion based relation inversion method, enhancing its quality as shown in Fig. 27. For example, when the relation "back to back" is to be expressed between two children, FreeU enhances ReVersion's ability to accurately represent this relationship. For the "inside" relation, when a *dog* is supposed to be placed inside of a *basket*, ReVersion sometimes generates a dog with artifacts, and introducing FreeU helps eliminate these artifacts. While ReVersion effectively captures relational concepts, Stable Diffusion might occasionally struggle to synthesize the relation concept due to excessive high-frequency noises in the U-Net skip features. Adding FreeU allows better entity and relation synthesis quality by using exactly the same relation prompt learned by ReVersion.

Rerender [61] is a diffusion model tailored for zero-shot text-guided video-to-video translations. Fig. 28 depicts the results: clear improvements in the detail and realism of synthesized videos. For instance, when provided with the prompt *"A dog wearing sunglasses"* and an input video, Rerender [61] initially produces a dog video with artifacts related to the *"sunglasses"*. However, the incorporation of FreeU successfully eliminates such artifacts, resulting in a refined output.

5. Related Work

Diffusion Probabilistic Models. Diffusion models have achieved great success in generation tasks [7, 8, 11, 13, 18, 24, 29, 33, 41, 45, 46, 49]. Distinct from other classes of generative models [4, 9, 12, 25–28, 32, 39, 53, 55] such as Variational Autoencoder (VAE) [32], Generative Adversarial Networks (GANs) [4, 12, 25–28, 39], and vector-quantized approaches [9, 53], diffusion models introduce a novel generative paradigm. These models employ a fixed Markov chain to map the latent space, facilitating intricate mappings that capture latent structural complexities within a dataset. Recently, its impressive generative capabilities, ranging from the high level of details to the diversity of the generated examples, have fueled groundbreaking advancements in a variety of computer vision applications such as image synthesis [18, 46, 49], image editing [1, 6, 21, 38], image-to-image translation [6, 48, 56], and text-to-video generation [3, 15, 19, 37, 52, 57, 58, 64]. Though successful, these studies mainly focus on utilizing pre-trained diffusion models for downstream applications, while the internal properties of the diffusion models remain largely underexplored. In this paper, we conduct a pioneering exploration of the potential of diffusion models.

Frequency Analysis Frequency analysis is commonly used to understand and enhance the performance of deep neural networks [2, 42, 51, 54, 59, 60, 63]. Recent studies such as [5, 10, 31, 50] have closely examined the frequency biases present in GANs models. Furthermore, the frequency characteristics of trained small diffusion models have been studied in [62]. In this study, we explore the denoising process in the Fourier domain for diffusion models and pioneer an investigation into the denoising potential of diffusion U-Net.

6. Limitations

FreeU significantly improves the generative capacity of the diffusion U-Net through adjusting two scaling factors. One obvious limitation of the proposed FreeU is that it requires manual configuration of scaling factors for each generative model. To address this limitation, an automated parameter

search mechanism for FreeU would be a viable and effective solution.

7. Potential Negative Societal Impacts

FreeU is a fundamental research project aimed at improving the generation quality of existing diffusion models without direct societal implications. However, when combined with other generative models, FreeU could potentially be used maliciously to create fake content or manipulate real human figures.

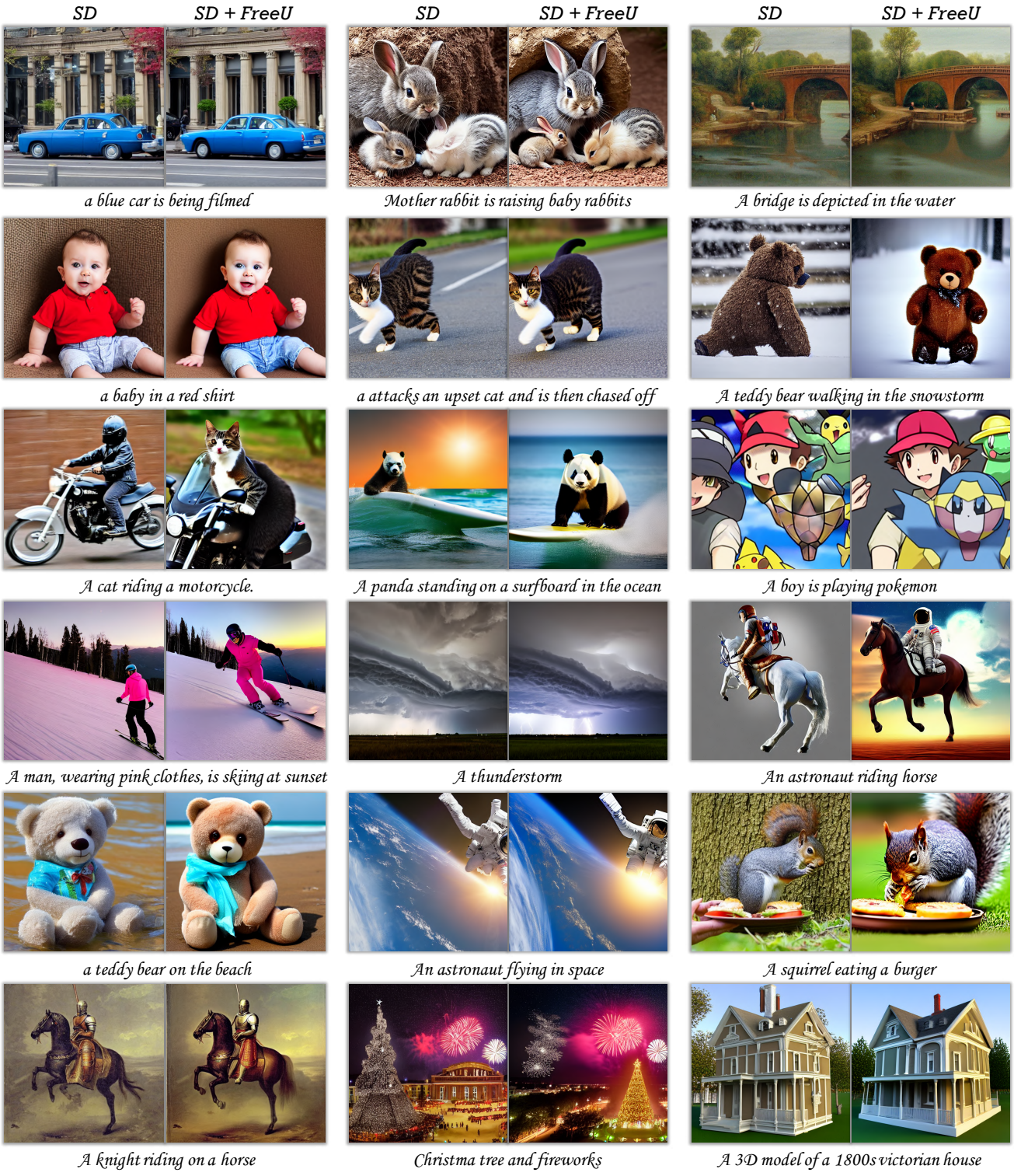

Figure 20. **Generated images from SD 1.4 [46] with and without FreeU enhancement.**

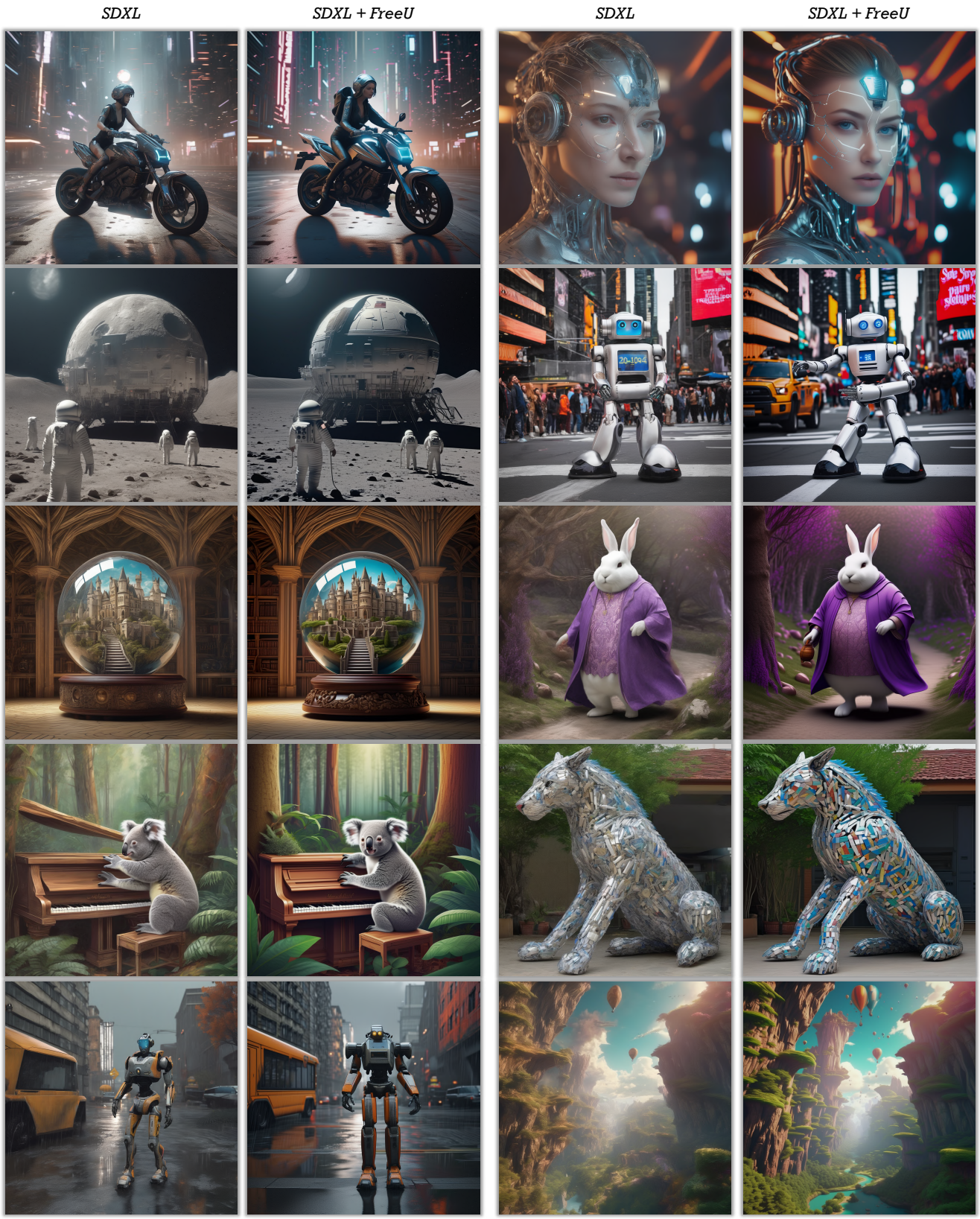

Figure 21. **Generated images from SD-XL [43] with and without FreeU enhancement.**

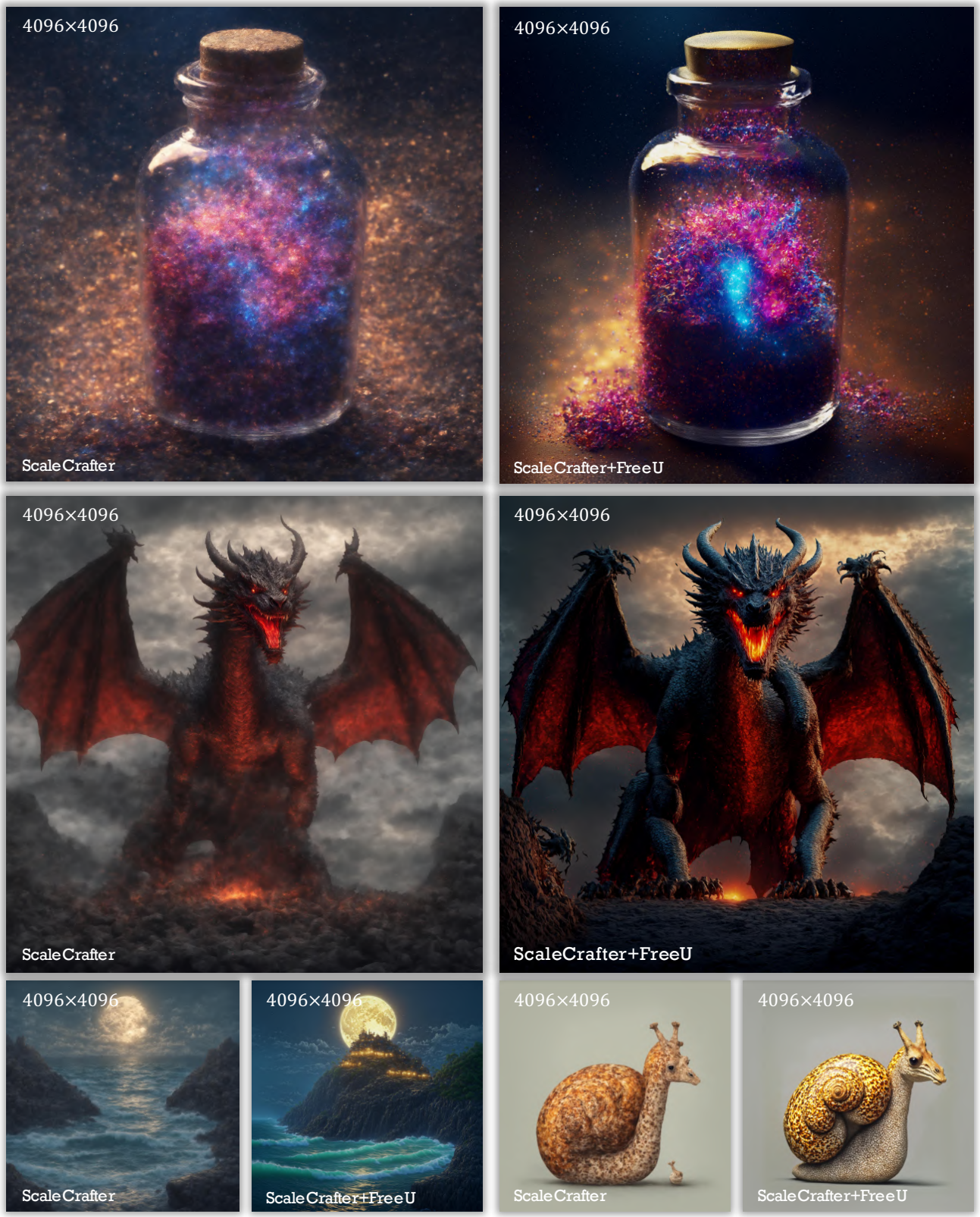

Figure 22. 4096 × 4096 **SD-XL [43] Images generated by ScaleCrafter [16] with or without FreeU.**

night, b&w photo of old house, post apocalypse, forest, storm weather, wind, rocks, 8k uhd, dslr, soft lighting, high quality, film grain

best quality, masterpiece, 1girl, looking at viewer, blurry background, upper body, contemporary, dress

... Begin it where warm waters halt and take it in a canyon down, not far but too far to walk...

Figure 23. **Generated videos from Animatediff [14] with and without FreeU enhancement.**

Figure 24. **Generated images from ControlNet [65] with and without FreeU enhancement.**

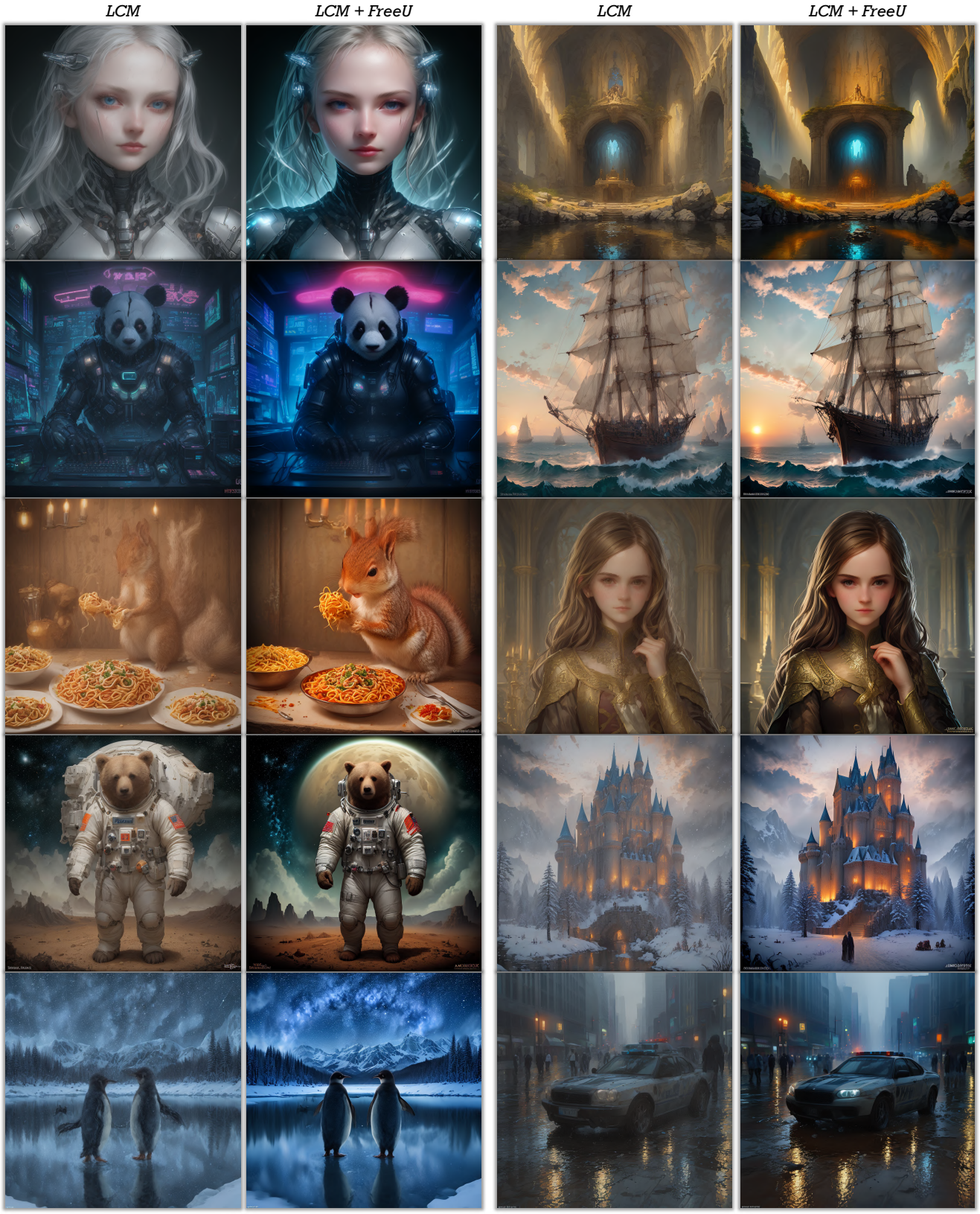

Figure 25. **Generated images from LCM [36] with and without FreeU enhancement.**

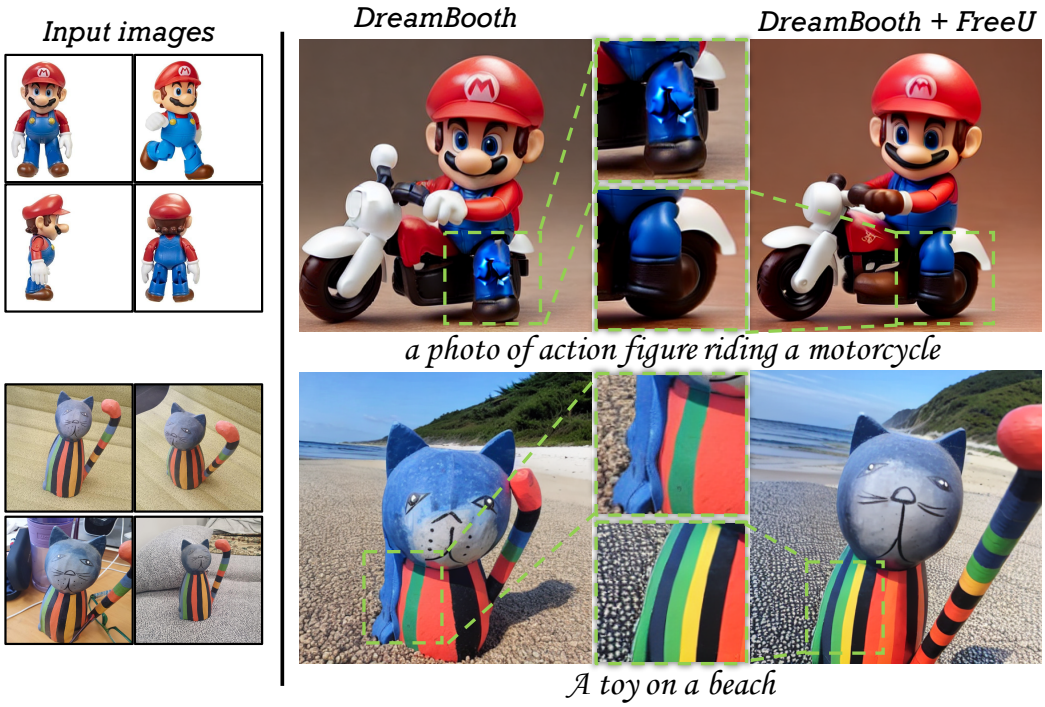

Figure 26. **Generated images from DreamBooth [47] with and without FreeU enhancement.**

Figure 27. **Generated images from ReVersion [23] with and without FreeU enhancement.**

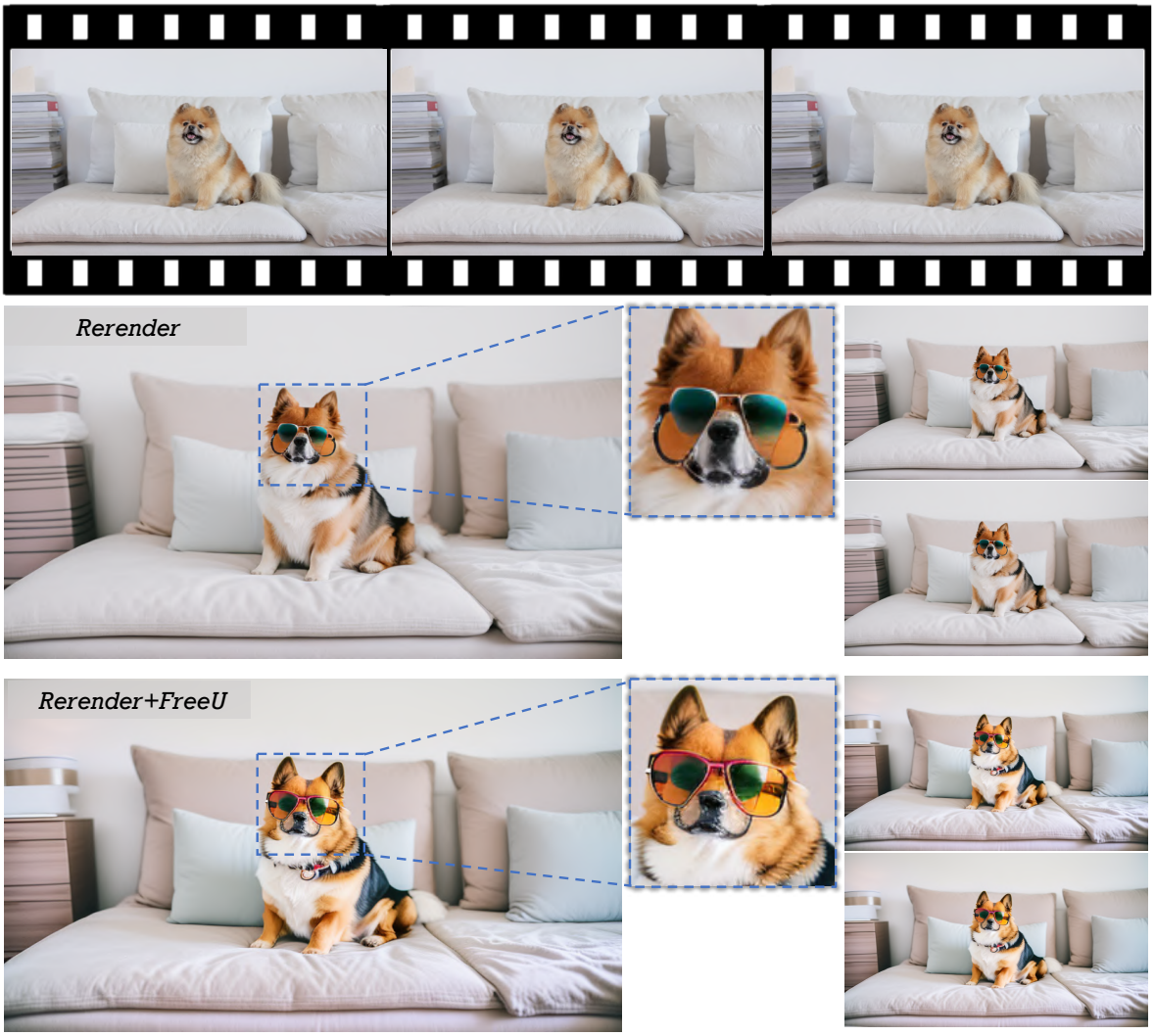

Figure 28. **Generated videos from Rerender [61] with and without FreeU Enhancement.**