Appendix to "FreCaS: Efficient Higher-Resolution Image Generation via Frequency-aware Cascaded Sampling"

In this appendix, we provide the following materials:

A Details of timestep shifting in the transition process (referring to Sec. 3.2 in the main paper);

B The detailed settings of FreCaS on $\times 4$ and $\times 16$ generation for SD2.1, SDXL and SD3 (referring to Sec. 4.1 in the main paper);

C Results of user studies and non-reference image quality assessment (NR-IQA) (referring to Sec. 4.1 in the main paper);

D Comparison with training-based methods and super-resolution methods (referring to Sec. 4.2 in the main paper);

E More visual results and visual comparisons (referring to Sec. 4.2 in the main paper);

F Experimental results of generation of SD3 (referring to Sec. 4.3 in the main paper);

G Ablation studies on individual components of FreCaS and inference schedule (referring to Sec. 4.4 in the main paper).

A Shifting Timestep in the Transition Process

As mentioned in Sec. 3.2 of the main paper, FreCaS employs a five-step transition process to transform the last latent in the current stage $\boldsymbol{z}_L^{s_{i-1}}$ to the first latent in the next stage $\boldsymbol{z}_F^{s_i}$. In addition to changing the resolution, we adjust the timestep from L to F to ensure that the signal-to-noise ratio (SNR) (Kingma et al., 2021) could be a constant in the transition process. Given a state $\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\epsilon$ at timestep t, the SNR is defined as $\text{SNR}(\boldsymbol{z}_t) = \frac{\alpha_t}{1-\alpha_t}$, where $\alpha_1, \ldots, \alpha_T$ represent the noise schedule, and ϵ is Gaussian noise. It has been found (Hoogeboom et al., 2023; Chen, 2023) that the SNR maintains a consistent ratio across resolutions for diffusion models using the same noise schedule:

$$\text{SNR}(\boldsymbol{z}_t^{s}) = \text{SNR}(\boldsymbol{z}_t^{\hat{s}}) \cdot \left(\frac{s}{\hat{s}}\right)^{\gamma},$$

where s and $\hat{s}$ denote different resolutions. The value of γ is typically set to 2.

Teng et al. (2024) and Gu et al. (2023) proposed to redesign the noise schedule to keep SNR consistent when changing the resolutions of intermediate states. Since the pre-trained diffusion models have fixed noise schedules, in this paper we adjust the timestep, instead of the noise schedule, to ensure consistent SNR between $\boldsymbol{z}_L^{s_{i-1}}$ and $\boldsymbol{z}_F^{s_i}$:

$$\text{SNR}(\boldsymbol{z}_L^{s_{i-1}}) = \text{SNR}(\boldsymbol{z}_F^{s_i}) \Rightarrow F = \alpha^{-1}\left(\frac{\left(\frac{s_{i-1}}{s_i}\right)^{\gamma} \cdot \alpha_L}{1 + \left(\left(\frac{s_{i-1}}{s_i}\right)^{\gamma} - 1\right) \cdot \alpha_L}\right), \tag{1}$$

where α^{-1} is the inverse function of α_t. Proper adjustment of γ can yield additional improvements.

Besides, SD3 (Esser et al., 2024) employs a similar formula to shift the timestep when varying resolutions:

$$F = \frac{\sqrt{\frac{s_i}{s_{i-1}}} \cdot L}{1 + (\sqrt{\frac{s_i}{s_{i-1}}} - 1) \cdot L}. \tag{2}$$

B Experimental Setting Details

The experimental setting details of our FreCaS are listed in Table 1.

C Results of User Studies and NR-IQA Metrics

We have (a) conducted user studies and (b) employed non-reference image quality assessment (NR-IQA) metrics to further assess the performance of FreCaS and its competing methods.

Table 1: Detailed settings of FreCaS on the experiments. N denotes the count of additional stages. "Steps" presents the sampling steps in each stage. L presents the timestep of last latent in each stage except for the final one. γ denotes the SNR ratio in the transition process. w_l, w_h and w_c are the hyper-parameters of the proposed FA-CFG and CA-maps re-utilization.

		$N+1$	Steps	L	γ	w_l	w_h	w_c
SD2.1	$\times 4$	2	40,10	100	3.0	7.5	45.0	0.6
	$\times 16$	3	30,10,10	200,200	3.0	7.5	35.0	0.4
SDXL	$\times 4$	2	40,10	200	1.5	7.5	35.0	0.6
	$\times 16$	3	30,5,15	400,200	2.0	7.5	35.0	0.6
SD3	$\times 4$	2	20,8	50	-	7.0	35.0	0.5

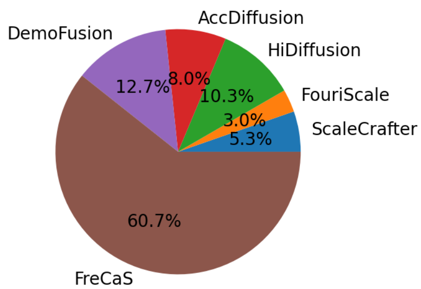

Figure 1: User study results on $\times 4$ generation of SDXL.

Table 2: NR-IQA metrics on $\times 4$ and $\times 16$ generation of SDXL.

Methods	$\times 4$			$\times 16$		
	clipiqa↑	niqe↓	musiq↑	clipiqa↑	niqe↓	musiq↑
DirectInference	0.522	4.167	53.98	0.469	4.370	29.00
AttnEntropy	0.547	4.210	54.87	0.528	4.614	27.98
ScaleCrafter	0.664	3.577	61.12	0.618	3.783	36.00
FouriScale	0.662	3.580	60.77	0.612	3.791	35.52
HiDiffusion	**0.690**	4.049	61.69	0.574	7.348	36.71
AccDiffusion	0.627	3.641	57.02	0.626	3.587	31.83
DemoFusion	0.651	3.410	58.98	0.637	3.376	33.46
Ours	0.668	**3.391**	**63.10**	**0.646**	**3.367**	**37.33**

C.1 User Studies

For the user studies, we compare FreCaS with ScaleCrafter, FouriScale, HiDiffusion, DemoFusion, and AccDiffusion on 2048×2048 image generation using SDXL. We randomly selected 30 prompts and generated one image per method for each prompt, creating 30 sets of images. Ten volunteers participated in the test, and they were asked to select the image with the best details and reasonable semantic layout from each set. The results are shown in Figure 1. We can see that FreCaS significantly outperforms other methods, with 60% votes as the best method. DemoFusion, AccDiffusion, and HiDiffusion perform similarly, with each having about 10% of the votes. In contrast, FouriScale and ScaleCrafter have the fewest votes, about 5% each.

C.2 NR-IQA Metrics

For the NR-IQA metrics, we employ CLIPIQA (Wang et al., 2023), NIQE (Mittal et al., 2012), and MUSIQ (Ke et al., 2021) on $\times 4$ and $\times 16$ image generations with SDXL. The results are presented in

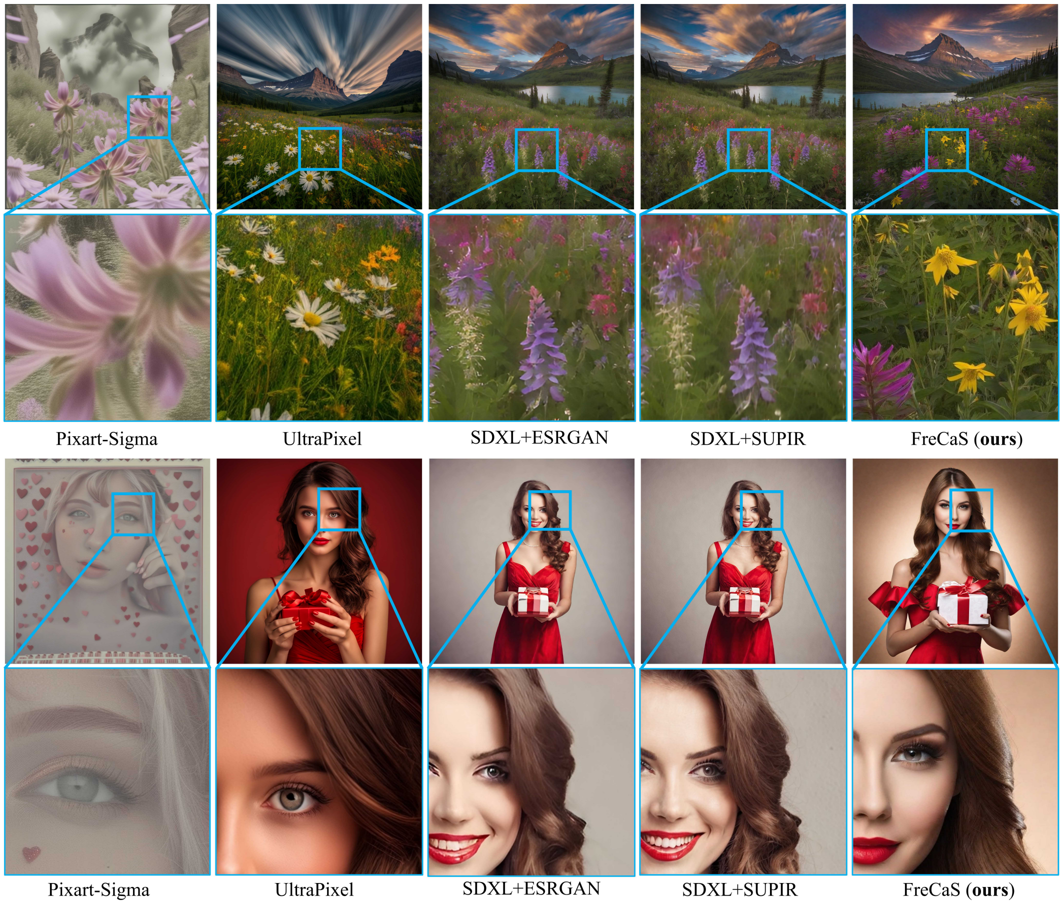

Figure 2: Visual comparison with training-based methods and super-resolution methods on $\times 4$ generation of SDXL.

Table 2. Our FreCaS consistently outperforms all the other methods. For example, on $\times 4$ generation, FreCaS achieves a CLIPIQA score of 0.668, a NIQE score of 3.391, and a MUSIQ score of 63.10, compared to 0.651, 3.410, and 58.98 for DemoFusion. On $\times 16$ generation, FreCaS achieved a CLIPIQA score of 0.646, a NIQE score of 3.367, and a MUSIQ score of 37.33, compared to 0.626, 3.587, and 31.83 for AccuDiffusion. Notably, FreCaS only lags behind HiDiffusion on the CLIPIQA metric in $\times 4$ image generation.

D Comparison with training-based methods and super-resolution methods

We conducted additional experiments comparing FreCaS with training-based methods (Pixart-Sigma (Chen et al., 2024) and UltraPixel (Ren et al., 2024)) and super-resolution methods (ESRGAN (Wang et al., 2021) and SUPIR (Yu et al., 2024)). To ensure fair comparisons, we set the model precision to fp16 (bf16 for UltraPixel, as recommended by the authors) and use the DDIM sampler for diffusion-based methods. For Pixart-Sigma, we can only report its results for 2048×2048 image generation since its 4K model is not available. The quantitative results are summarized in Table 3.

From Table 3, we can see that FreCaS outperforms Pixart-Sigma and UltraPixel in most metrics. For example, FreCaS achieves an FID score of 16.48 and an IS score of 17.18, compared to 26.1 and 14.44 of Pixart-Sigma, and 25.56 and 17.11 of UltraPixel on the $\times 4$ image generation task. This is because Pixart-Sigma, as acknowledged by the authors, heavily relies on the advanced samplers (see https://github.com/PixArt-alpha/PixArt-sigma/issues/65) so that the results are not very stable.

Table 3: Comparison with training-based methods and super-resolution methods on $\times 4$ and $\times 16$ generation of SDXL.

	Methods	FID↓	FID_p↓	IS↑	IS_p↑	CLIP SCORE↑	Latency(s)↓
$\times 4$	Pixart-Sigma	26.11	38.58	14.44	14.45	28.10	71.45
	UltraPixel	25.56	19.95	17.11	17.10	33.17	41.70
	SDXL+ESRGAN	13.03	18.10	17.30	16.58	34.13	6.36
	SDXL+SUPIR	12.08	17.31	17.57	17.12	34.16	105.5
	Ours	16.48	17.91	17.18	17.31	33.28	13.84
$\times 16$	UltraPixel	51.43	45.88	12.48	13.73	33.07	162.4
	SDXL+ESRGAN	45.86	43.10	12.94	13.48	33.44	7.25
	SDXL+SUPIR	43.94	39.35	13.22	14.37	33.49	512.4
	Ours	42.75	39.82	12.68	14.16	33.03	85.87

UltraPixel, while achieving comparable performance to DemoFusion, still lags behind FreCaS in most metrics. Besides, the two methods are much slower than our FreCaS.

For SR-based methods, FreCaS may have lower FID, IS, and CLIP scores than SDXL+ESRGAN. This is because SR methods are designed to strictly adhere to low-resolution inputs, while these metrics (FID, IS, and CLIP) evaluate images by downsampling them to low resolution, which cannot well reflect the quality of generated high-resolution images. However, FreCaS significantly outperforms SDXL+ESRGAN in FID_p and IS_p. Specifically, FreCaS achieves an FID_p score of 39.82 and an IS_p score of 14.16, compared to 43.10 and 13.48 of SDXL+ESRGAN on $\times 16$ image generation. This indicates its superior ability to generate high-resolution local details. This observation is consistent with the findings in the DemoFusion paper. Additionally, SDXL+SUPIR outperforms FreCaS in FID_p and IS_p, but at the cost of much longer inference latency (85.87 seconds for FreCaS vs. 512.4 seconds for SDXL+SUPIR on $\times 16$ image generations).

We have provided some visual comparisons in Figure 2. One can see that FreCaS demonstrates better visual quality than either training-based or SR-based methods in high-resolution image generation, such as the more vivid and clearer flowers, hairs and the more natural color of lips.

E More Visual Results

E.1 More Visual Results

Figure 3 illustrates more visual results of FreCaS, including those with varying aspect ratios. From top to bottom, and left to right, the prompts used in examples are: 1. “Beautiful winter wallpapers.” 2. “A regal queen adorned with jewels.” 3. “A majestic phoenix, wings ablaze, rising from ashes, the flames casting a warm glow.” 4. “Lady in Red oil portrait painting won the John Singer Sargent People’s award.” 5. “Star of the day – Actress Evelyn Laye - 1917.” 6. “Photograph - Clouds Over Daicey Pond by Rick Berk.” 7. “little-boy-with-large-bulldog-in-a-garden-france.” 8. “03-Brussels-Maja-Wronska-Travels-Architecture-Paintings.”, 9. “Red Fox Pup Print by William H. Mullins.” 10. “Lovely Illustrations Of Cityscapes Inspired By Southeast Asia Malaysian digital illustrator Chong Fei Giap’s illustrations of cityscapes are lovely and inspiring. Fantasy Landscape, Landscape Art, Illustrator, Japon Tokyo, Animation Background, Art Background, Matte Painting, Anime Scenery, Jolie Photo.” 11. “A plate with creamy chicken and vegetables, a side of onion rings, a cup of coffee and a slice of cheesecake.” 12. “Hyper-Realistic Portrait of Redhead Girl Drawn with Bic Pens.”

To further validate the performance of FreCaS in real-world application scenarios, we have provided additional visual results in three categories:

- **Simple scenes.** These images typically contain a single object in a realistic style. We display images of people, animals, landscapes, buildings, and other common objects. The visual results for this group are presented in Figure 4.

Figure 3: Visual results of FreCaS on SDXL. Please zoom-in for better view.

- **Various styles.** This group showcases images in different artistic styles, including oil painting, pencil sketch, ink wash, watercolor, and poster art. The results are shown in the first two rows of Figure 5.

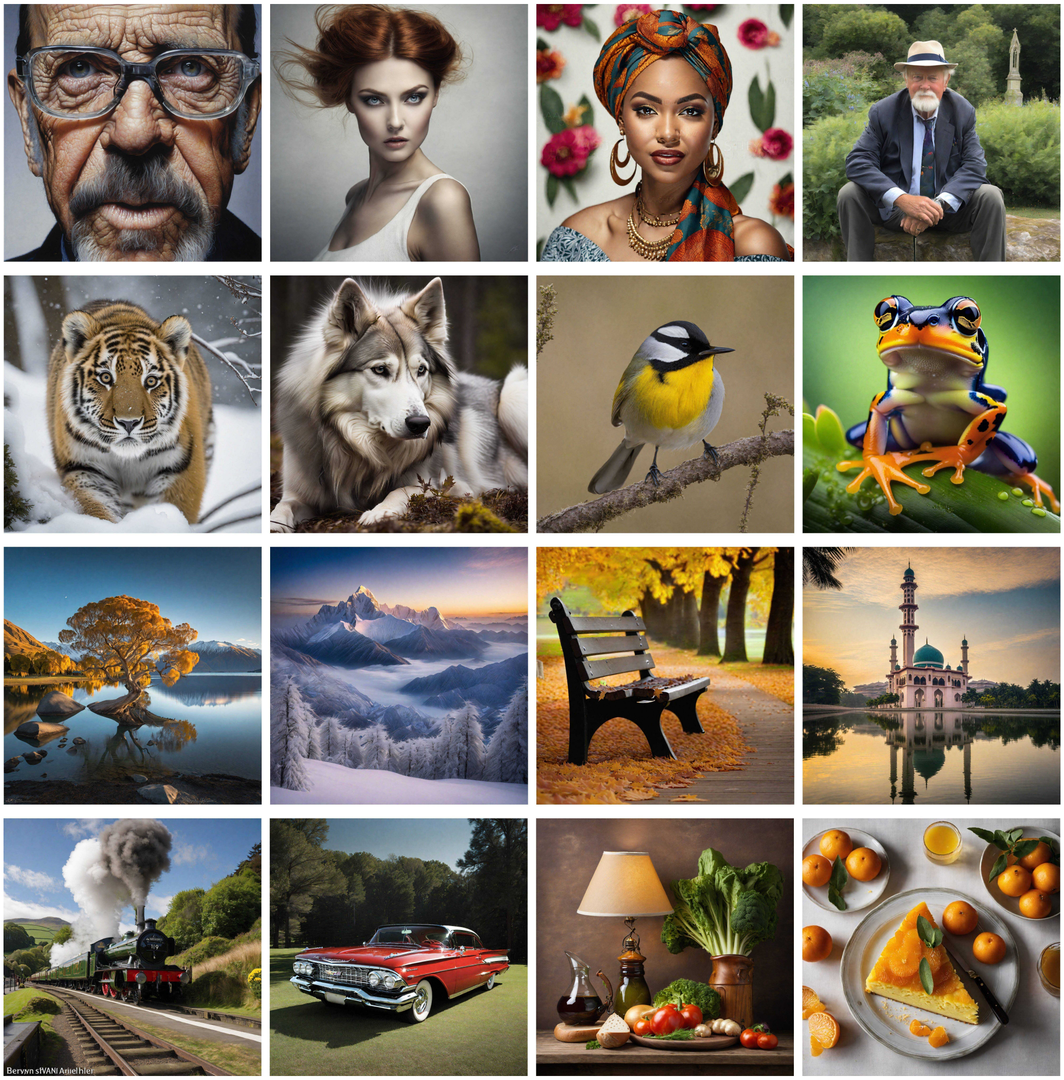

Figure 4: More visual results on simple scenes.

- **Complex scenes.** These images contain multiple objects or have intricate textures. The results are displayed in the bottom two rows of are presented in Figure 5.

From these visual results, it is evident that FreCaS consistently generates high-quality images across various styles and contents, demonstrating the capability of FreCaS in real-world applications.

z

E.2 More Visual Comparisons

We show more visual comparisons in Figure 6. From top to bottom, the prompts used in the four groups of examples are: 1. "A small den with a couch near the window." 2. "A painting of a candlestick holder with a candle, several pieces of fruit and a vase, with a gold frame around the painting." 3. "A noble knight, riding a white horse, the castle gates opening." 4. "Mystical Landscape Digital Art - Lonely Tree Idyllic Winterlandscape by Melanie Viola."

We have provided more 4K visual comparisons under realistic scenes in Figure 7. As can be seen, our FreCaS consistently delivers better results in both image layout and semantic details.

Figure 5: More visual results of various styles (top two rows) and complex scenes (bottom two rows).

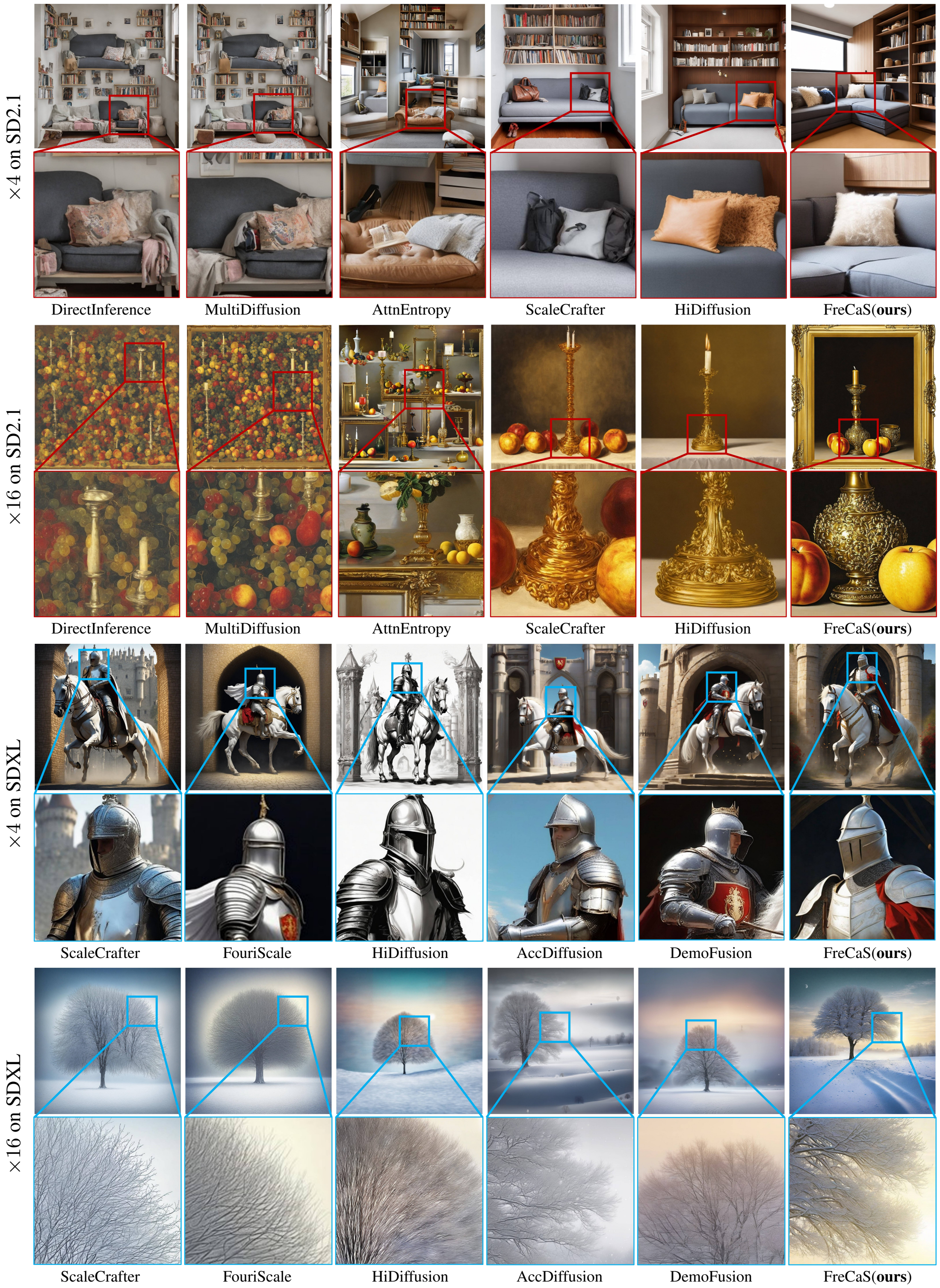

Figure 6: Visual comparisons on $\times 4$ and $\times 16$ experiments of SD2.1 and SDXL. Please zoom-in for better view.

F EXPERIMENTS ON SD3

In this section, we present the results of the $\times 4$ generation experiments on SD3. SD3 employs a transformer-based denoising network. It eliminates all convolutional layers, thereby preventing

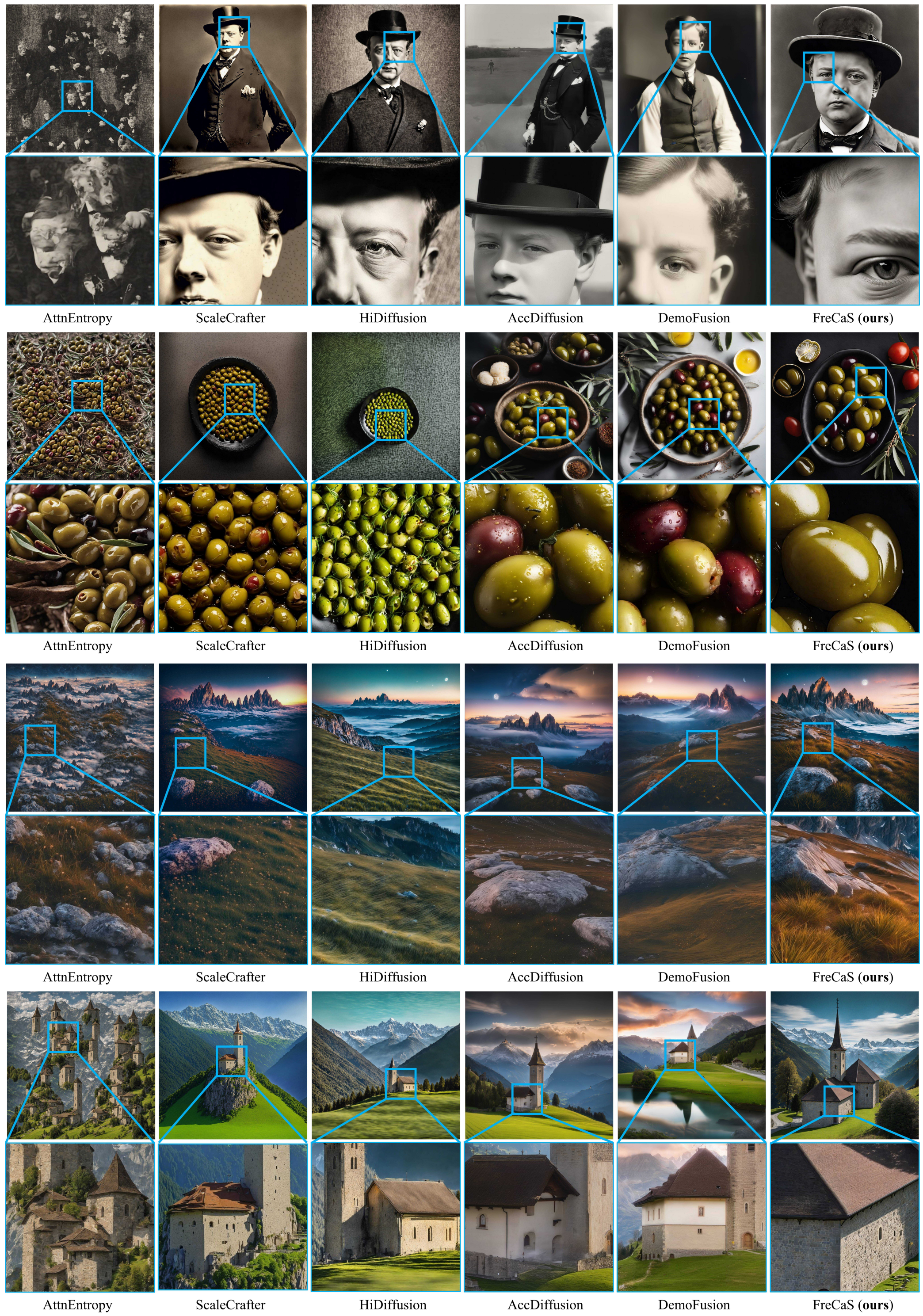

Figure 7: More 4K comparisons in realistic styles. From top to bottom, the prompts are "Young winston churchill.", "Olive food photography.", "Mountains in fog at beautiful night. Dreamy landscape with mountain peaks, stones, grass, blue sky with blurred low clouds, stars and moon. Rocks at dusk." and "Image Church Switzerland towers San Romerio Nature Mountains Scenery Made of stone Tower mountain landscape photography."

Table 4: Experiments on $\times 4$ generation of SD3.

Methods	$\text{FID}_b \downarrow$	$\text{FID}_p \downarrow$	IS$\uparrow$	$\text{IS}_p \uparrow$	CLIP SCORE$\uparrow$	Latency (s)$\downarrow$	SpeedUP$\uparrow$
DirectInference	35.68	45.35	12.52	12.60	31.45	38.53	1$\times$
Demodiffusion	15.19	44.34	17.84	14.99	31.09	63.33	0.61$\times$
Ours	9.76	26.62	17.83	16.72	31.17	15.94	2.42$\times$

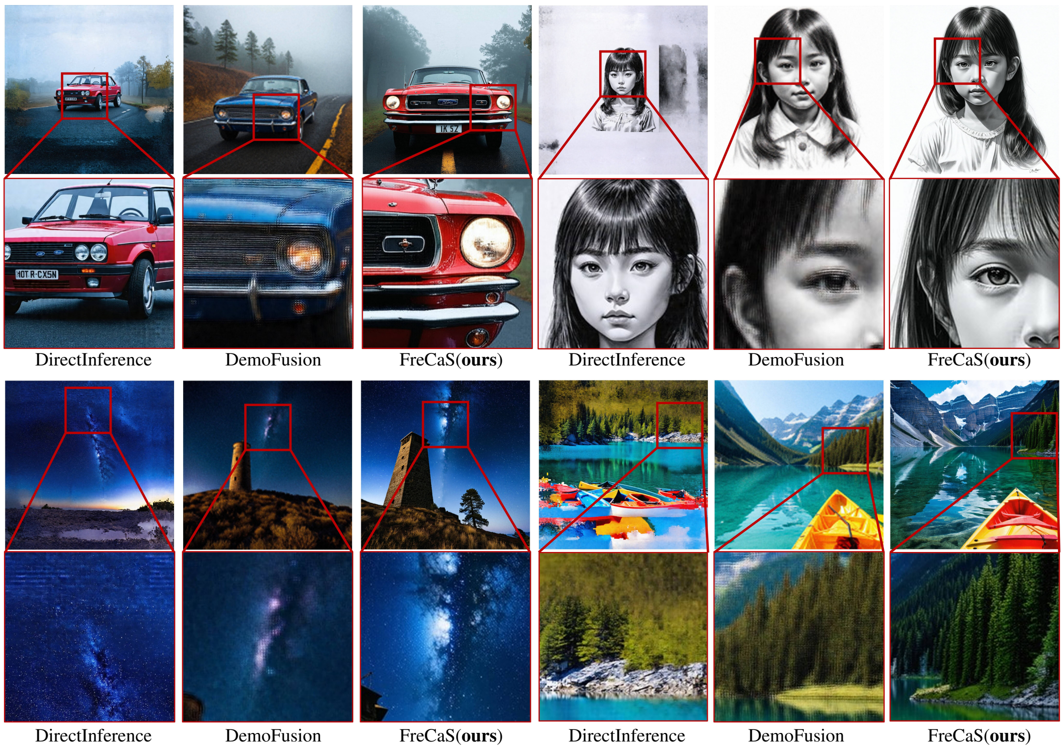

Figure 8: Visual comparison on $\times 4$ experiments of SD3. From top to bottom, from left to right, the prompts used in the four groups of examples are: 1. "Car Photograph - Ford In The Fog by Debra and Dave Vanderlaan." 2. "Rupert Young is Sir Leon in Merlin season 5 copy." 3. "Watchtower, Shooting Star & Milky Way, Gualala, CA." 4. "Colorful Autumn in Mount Fuji, Japan - Lake Kawaguchiko is one of the best places in Japan to enjoy Mount Fuji scenery of maple leaves changing color giving image of those leaves framing Mount Fuji.". Zoom-in for better view.

the application of many existing methods, such as ScaleCrafter and FouriScale. Besides, SD3 exhibits fine details in the central region but shows corrupted textures in the surrounding regions (see Figure 8). This issue with the image layout also significantly impacts the performance of other methods, such as DemoFusion. Therefore, we only compare our FreCaS with DirectInference and DemoFusion. Table 4 and Figure 8 present the quantitative and qualitative results, respectively.

From Table 4, it is evident that FreCaS achieves superior performance in terms of image quality and inference speed. Specifically, FreCaS achieves the best results on FID_b, FID_p, IS, and IS_p, and only slightly lags behind DirectInference in terms of CLIP score. Moreover, FreCaS generates a 2048×2048 image in about 16 seconds, achieving a speed-up of $2.42\times$ and $3.97\times$ compared to DirectInference and DemoFusion, respectively. Figure 8 illustrates the generated images. Directly employing the pre-trained SD3 model to generate higher-resolution images, DirectInference leads to unreasonable image layout with the surrounding parts being corrupted, such as the road and trees. The results of DemoFusion exhibits strange artifacts, such as the car faces and eyes. In contrast, our FreCaS successfully maintains the natural image structure while obtaining fine details.

Table 5: Ablation studies on 2048×2048 generation of SDXL.

Model	cascaded framework	FA-CFG	CA-reuse	FID↓	$\text{FID}_p\downarrow$	IS↑	$\text{IS}_p\uparrow$	CLIP SCORE↑	Latency (s)
#1				39.14	29.71	11.52	14.60	32.51	34.10
#2	✓			17.62	20.49	17.01	16.54	33.24	13.71
#3	✓	✓		16.62	17.91	17.16	16.82	33.34	13.74
#4	✓	✓	✓	16.48	17.91	17.18	17.31	33.28	13.84

Table 6: Ablation studies on N in FreCaS.

N	resolutions	$\text{FID}_b\downarrow$	$\text{FID}_p\downarrow$
0	2048	43.83	29.71
1	$1024 \to 2048$	12.63	17.91
2	$1024 \to 1536 \to 2048$	41.36	28.68

Table 7: Ablation studies on L in FreCaS.

L	$\text{FID}_b\downarrow$	$\text{FID}_p\downarrow$
0	12.57	18.20
100	12.69	18.10
200	12.63	17.91
300	13.30	18.57
400	13.34	18.62

G Ablation Studies on Individual Components and Inference Schedule

We further conduct ablation studies to verify the effectiveness of each components and the settings of inference schedule of our FreCaS.

G.1 Effectiveness of Each Component

To better verify the effectiveness of each component of FreCaS, we conducted more ablation studies on our proposed cascaded framework, FA-CFG, and CA-reuse strategies. The results are shown in Table 5. One can see that our cascaded framework significantly outperforms the baseline, with a decrease of 22.52 in the FID score and a reduction of 20.39 seconds in latency. This demonstrates the high efficiency of our proposed cascaded framework. Our FA-CFG strategy improves both FID and IS scores and shows substantial improvement in FID_p, demonstrating its effectiveness in generating realistic image details. The CA-reuse strategy further enhances IS_p, indicating its effectiveness in improving semantic appearance. Moreover, these strategies introduce minimal additional latency.

G.2 Experiments on Inference Schedule

In this section, we conduct experiments on the selection of N (number of additional stages) and L (the timestep of last latent in each stage). The two factors are employed to adjust the inference schedule of our FreCaS. We reports the scores of FID_b and FID_p by varying the two factors in Table 6 and Table 7, respectively.

Choice of N. From Table 6, we see that $N = 1$ achieves an FID_b score of 12.63 and an FID_p score of 17.91, significantly better than $N = 0$ and $N = 2$ in the $\times 4$ generation task for SDXL. This could be attributed to the fact that a larger value of N introduces more transition steps, which can lead to much information loss. Conversely, a smaller value of N reduces the effectiveness of FreCaS, degenerating it to the DirectInference method.

Choice of L. From Table 7, we can see that a smaller L improves FID_b score but deteriorates FID_p. This is because the details generated at lower resolutions conflict with those at higher resolutions. Thus, we set L to 200 to avoid generating excessive unwanted details in the early stages.

REFERENCES

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Miguel Ángel Bautista, and Joshua M Susskind. f-dm: A multi-stage diffusion model via progressive signal transformation. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.