Supplementary Material
Make a Cheap Scaling: A Self-Cascade Diffusion Model for Higher-Resolution Adaptation

Lanqing Guo[1,2†], Yingqing He[2,3†], Haoxin Chen[2], Menghan Xia[2], Xiaodong Cun[2], Yufei Wang[1], Siyu Huang[4], Yong Zhang[2*], Xintao Wang[2], Qifeng Chen[3], Ying Shan[2], and Bihan Wen[1*]

[1] Nanyang Technological University
[2] Tencent AI Lab
[3] The Hong Kong University of Science and Technology
[4] Clemson University

Project page: https://guolanqing.github.io/Self-Cascade/

In this supplementary material, we include more implementation details of the proposed Self-Cascade Diffusion Model (Section 1), more experiments on 16× image scale adaptation (Section 2), and more visual comparisons and examples on text-to-image and text-to-video synthesis on different base models, *i.e.*, SD 2.1, SD XL 1.0, and LVDM [3] (Section 3). The code will be released.

1 Implementation Details

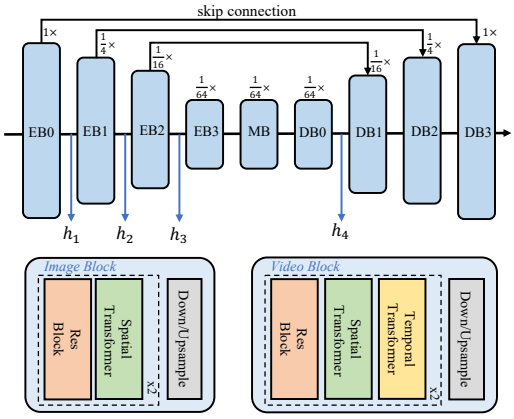

Fig. 1: (a) The detailed architecture of the U-Net denoiser ϵ_θ in stable diffusion model, consisting of various EB: encoder block; MB: middle block; DB: decoder block. We select the multi-scale features $\{h_1, h_2, h_3, h_4\}$ to be re-scaled by the plugged time-aware feature upsampler. (b) The detailed architecture of blocks in the U-Net of text-to-image base model. (c) The detailed architecture of blocks in the U-Net of text-to-video base model.

Methods	Random Select				High-Resolution Select			
	$FID_r\downarrow$	$KID_r\downarrow$	$FID_b\downarrow$	$KID_b\downarrow$	$FID_r\downarrow$	$KID_r\downarrow$	$FID_b\downarrow$	$KID_b\downarrow$
Original	104.70	0.043	104.10	0.040	96.85	0.040	105.04	0.045
Attn-SF [4]	104.34	0.043	103.61	0.041	99.56	0.042	108.67	0.048
ScaleCrafter [2]	59.40	0.021	57.26	0.018	36.64	0.010	37.61	**0.011**
Ours-TF	**38.99**	**0.015**	**34.73**	**0.013**	**27.94**	**0.009**	**29.56**	**0.011**
Full Fine-tuning (20k)	43.55	0.014	41.58	0.012	22.20	0.004	25.43	0.007
LORA-R4 (20k)	50.72	0.020	51.99	0.019	38.17	0.012	44.33	0.017
Ours-T (10k)	**18.46**	**0.005**	**8.99**	**0.001**	**13.87**	**0.003**	**8.32**	**0.001**

Table 1: Quantitative results of different methods on the dataset of *Laion-5B* with $16\times$ image scale adaptation on 2048^2 resolution.

1.1 Details of Model Layers

The detailed architectures of the plugged time-aware feature upsampler are the same for all experiments. Each upsampler consists of one bilinear upsampling operation followed by two residual blocks. The U-Net of Stable Diffusion (SD) v2.1, and SD XL v1.0 share the similar convolution layer layout. Note that the base model of text-to-video experiments, *i.e.*, LVDM [3], applied the SD v2.1 as the backbone. We illustrate the detailed architecture of U-Net denoiser $\epsilon_\theta(\cdot)$ in Figure 1 and explain which feature to select for the plugged upsamplers. We set $N = 4$ for all image and video experiments with $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ feature upsampler set. The detailed locations of corresponding selected multi-scale features in feature group $h = \{h_1, h_2, h_3, h_4\}$ are illustrated in Figure 1(a), which are the feature maps after the downsampling operation of different scales. We also illustrate the detailed architectures of the adopted blocks in U-Net for text-to-image and text-to-video base models in Figure 1(b)&(c).

1.2 Details of Scale Adaptation Settings

As mentioned in the main paper, we conduct evaluation experiments on text-to-image models, specifically focusing on Stable Diffusion (SD). We examine two widely-used versions: SD 2.1 [1] and SD XL 1.0 [6], as they adapt to two unseen higher-resolution domains. The original SD 2.1 is trained with 512^2 images, and its inference resolutions are 1024^2 and 2048^2, corresponding to $4\times$ and $16\times$ more pixels than the training, respectively. Similarly, the original SD XL 1.0 is trained with 1024^2 images, and its inference resolutions are 2048^2 and 4096^2, also corresponding to $4\times$ and $16\times$ more pixels than the training, respectively. Additionally, we conduct evaluation experiments on text-to-video models, selecting the LVDM [3] as the base model. This model is trained with 16×256^2 videos (16 frames), and its inference resolutions are 16×512^2, which is $4\times$ more pixels than the training.

2 More experiments on 16× Scale Adaptation

Apart from the tuning-free comparison included in the main paper, we also provide the comparison of tuning methods, including the full fine-tuning, LORA-R4, and Ours-T. Following previous work [2], we randomly sample $10k$ images as metric evaluation for this setting, denoted as Random Select. To ensure consistency in image pre-processing steps, we use the clean-fid implementation [5]. Besides, to further evaluate the detailed structures in high-resolution results, we selected a high-resolution real set, including $10k$ higher than 2048^2 resolution images, denoted as High-Resolution Select. Table 1 shows the quantitative results on Laion-5B [7] over 16× more pixels compared to base model SD 2.1. We show random samples from our method on adapted higher resolutions 2048^2 in Figure 2 and Figure 3.

3 More Visual Examples

3.1 Visual Examples on SD XL 1.0

Figure 4 and Figure 5 illustrate the visual examples of our proposed method on varied resolutions, *i.e.*, 2048^2 and 4096^2 for 4× and 16× scale adaptations.

3.2 Comparison on SD 2.1

Figure 6 and Figure 7 illustrate the visual comparison between our proposed method and the competing methods, including both tuning-free and tuning settings.

3.3 Visual Examples on Video Synthesis

Please view the video example in the **attached webpage**. It is recommended to use the Chrome browser to open it. Click on the **'supp.html'** file and choose to open it with the **Chrome browser**.

References

1. Diffusion, S.: Stable diffusion 2-1 base. https://huggingface.co/stabilityai/stable-diffusion-2-1-base/blob/main/v2-1_512-ema-pruned.ckpt (2022) 2
2. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. arXiv preprint arXiv:2310.07702 (2023) 2, 3, 8, 9
3. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022) 1, 2
4. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. arXiv preprint arXiv:2306.08645 (2023) 2, 8, 9

Fig. 2: Visual examples of the higher-resolution adaptation to various higher resolutions, *e.g.*, 1024^2 and 2048^2 for $4\times$ and $16\times$ scale adaptation, with the pre-trained SD 2.1 trained with 512^2 images. Please zoom in for more details.

5. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022) 3
6. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2
7. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022) 3

Fig. 3: Visual examples of the higher-resolution adaptation to various higher resolutions, *e.g.*, 1024^2 and 2048^2 for $4\times$ and $16\times$ scale adaptation, with the pre-trained SD 2.1 trained with 512^2 images. Please zoom in for more details.

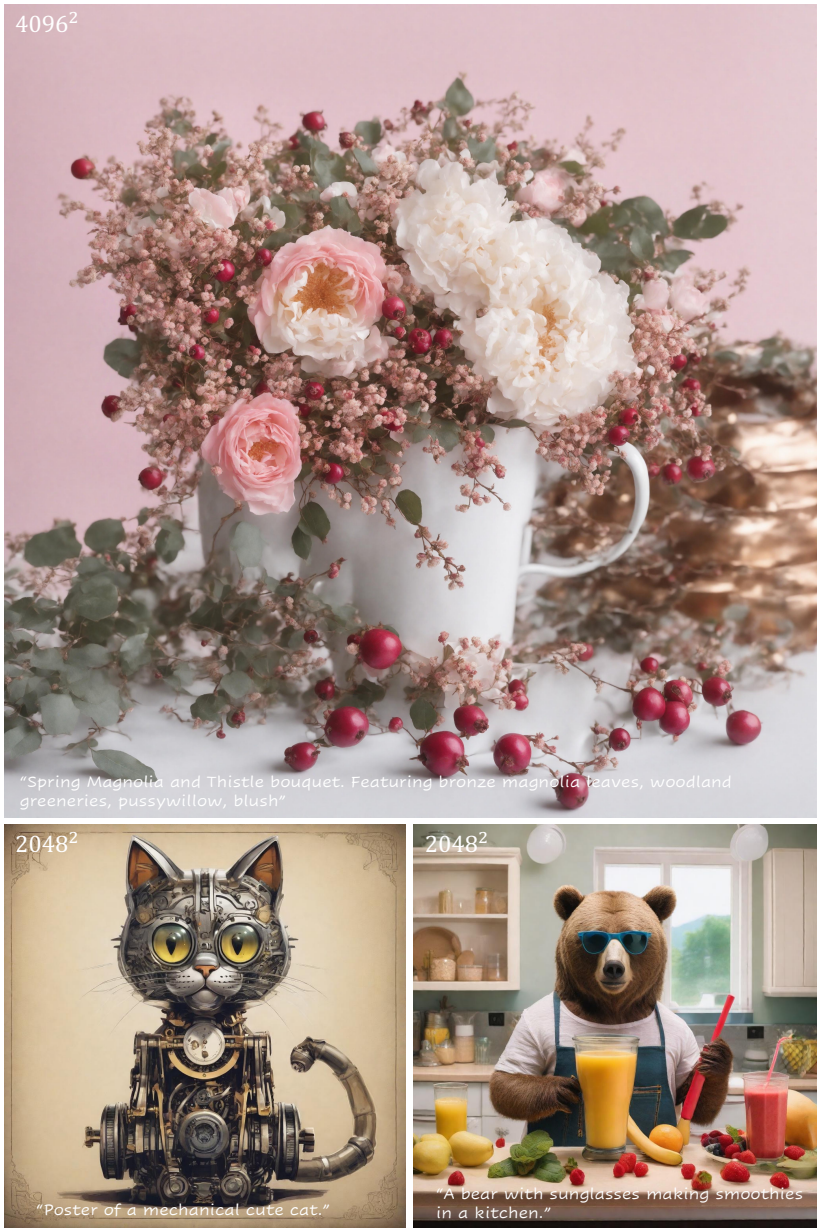

Fig. 4: Visual examples of the higher-resolution adaptation to various higher resolutions, *e.g.*, 2048^2 and 4096^2 for $4\times$ and $16\times$ scale adaptation, with the pre-trained SD XL 1.0 trained with 1024^2 images. Please zoom in for more details.

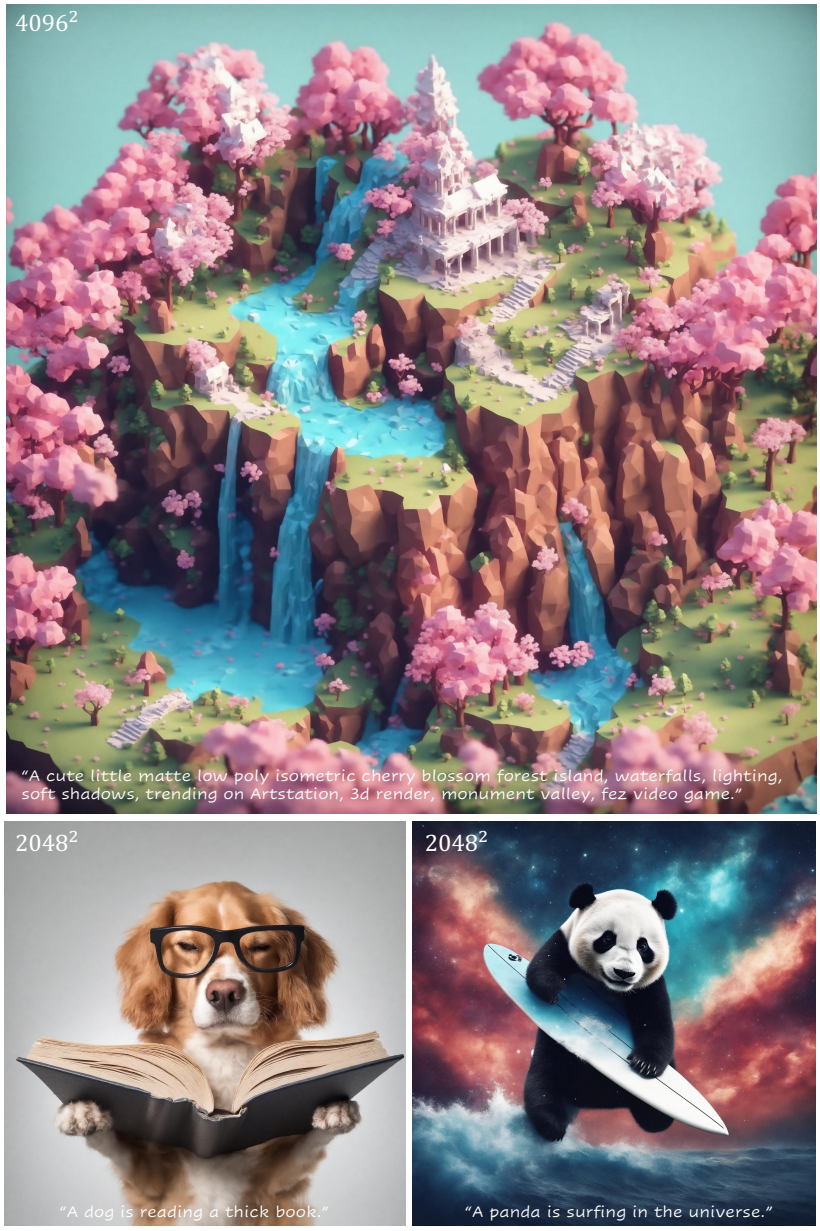

Fig. 5: Visual examples of the higher-resolution adaptation to various higher resolutions, *e.g.*, 2048^2 and 4096^2 for $4\times$ and $16\times$ scale adaptation, with the pre-trained SD XL 1.0 trained with 1024^2 images. Please zoom in for more details.

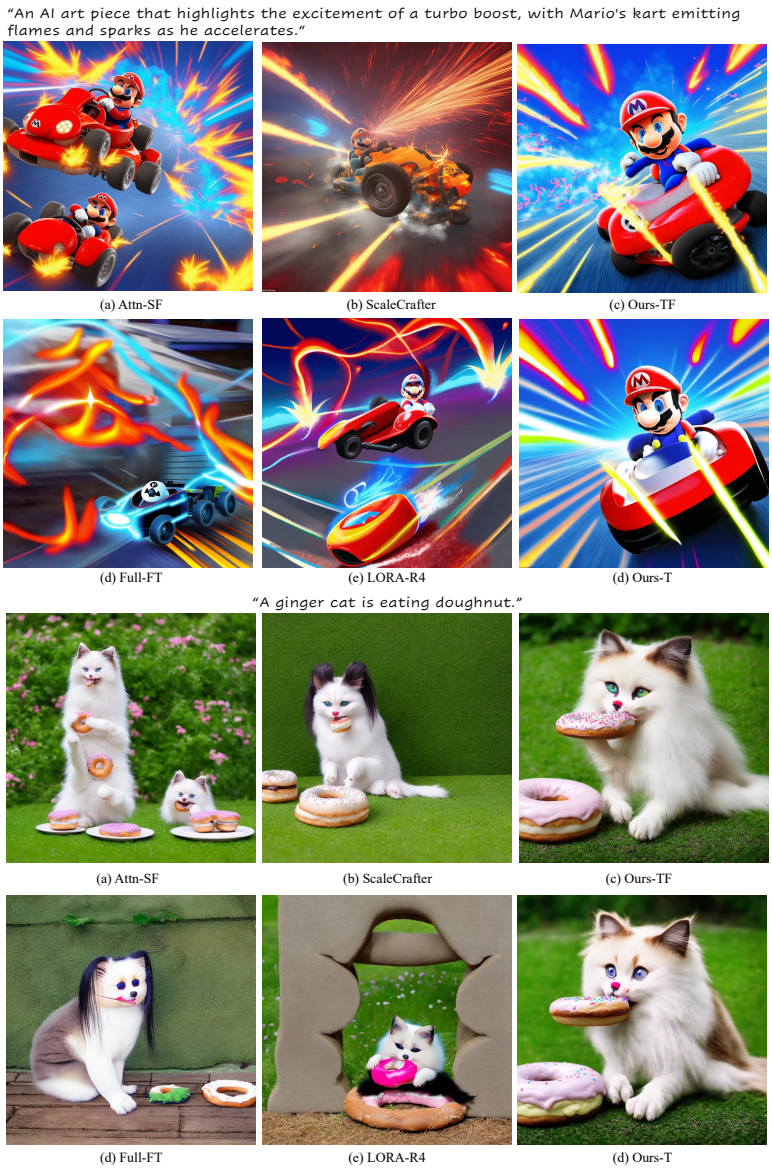

Fig. 6: Visual comparisons between Turing-free methods with 1024^2: (a) Attn-SF [4], (b) ScaleCrafter [2], (c) Ours-TF; and Tuning methods: (d) Full Finetuning (Full-FT) tuning with $18k$ steps, (e) LORA-R4 tuning with $18k$ steps, (d) Ours-T tuning with $4k$ steps, of the higher-resolution adaptation to $4\times$ higher resolutions over SD 2.1 base model. Please zoom in for more details.

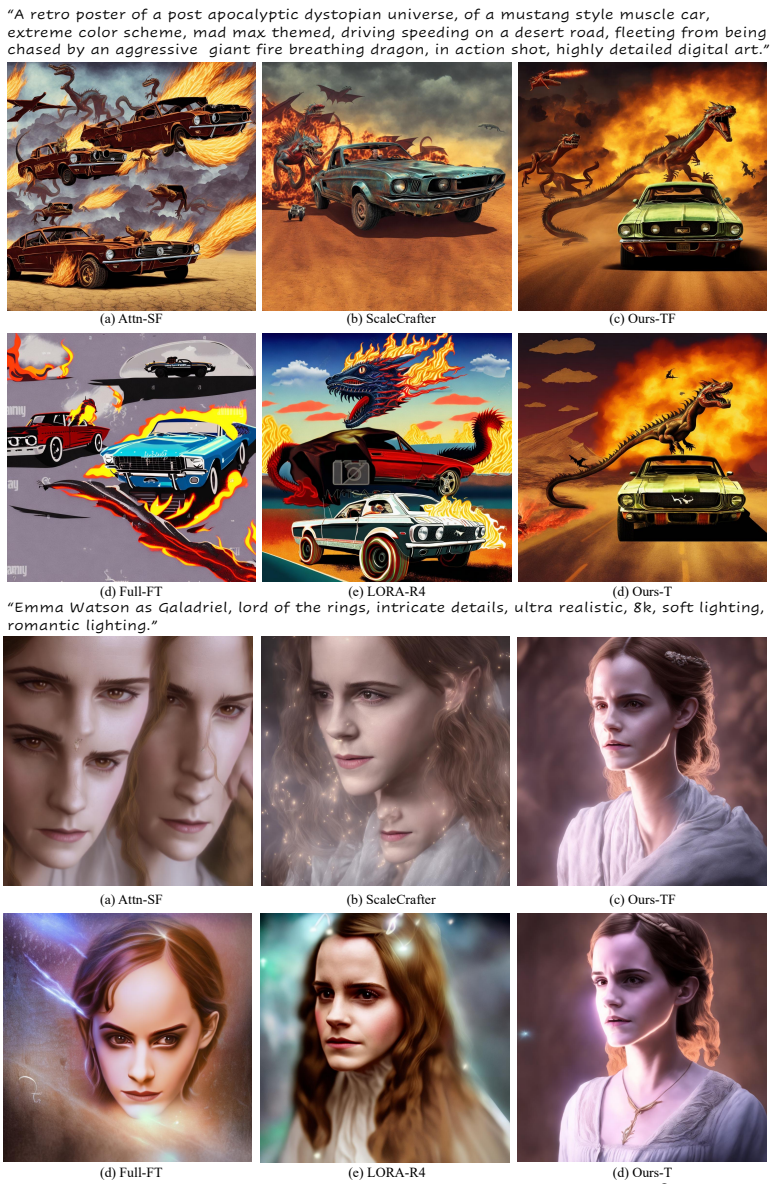

Fig. 7: Visual comparisons between Turing-free methods with 1024^2: (a) Attn-SF [4], (b) ScaleCrafter [2], (c) Ours-TF; and Tuning methods: (d) Full Finetuning (Full-FT) tuning with $18k$ steps, (e) LORA-R4 tuning with $18k$ steps, (d) Ours-T tuning with $4k$ steps, of the higher-resolution adaptation to $4\times$ higher resolutions over SD 2.1 base model. Please zoom in for more details.