FreeScale: Unleashing the Resolution of Diffusion Models via Tuning-Free Scale Fusion

Supplementary Material

Overview. In the supplementary material, we introduce implementation details in Section A, show more evaluations in Section B, exhibit more results in Section C, and finally, discuss limitations and future work in Section D.

A. Implementation Details

During sampling, we perform DDIM sampling [48] with 50 denoising steps, setting DDIM eta to 0. For image generation, the base inference resolution of SDXL is 1024×1024 pixels, and the scale of the classifier-free guidance is set to 7.5. For video generation, the base inference resolution of VideoCrafter2 is 320×512, the video length is 16 frames, and the scale of the classifier-free guidance is set to 12.0.

For tailored self-cascade upscaling, we set $K = 700$ in Equation 3 for all experiments. And in Equation 4, α is set as a scaler, 2, by default. To avoid excessive and messy textures in generating 8k images, α is reduced to 1. In Figure 4, α is 3 and 0.5 in the targeted and other areas, respectively. Users can further adjust these parameters according to the detailed requirements of different images. For restrained dilated convolution, the dilation factor d in Equation 5 is equal to the resolution level (1 represents original resolution, 2 represents the twice height and width). For scale fusion, the kernel size is $2 \times \sqrt{height \times width \div (1024 \times 1024)} - 1$ and the standard deviation is 1 in Equation 7.

Datasets. We evaluate image generation on the LAION-5B dataset [46] with 1024 randomly sampled captions. Specifically, to better align with human preference, we randomly selected prompts from the LAION-Aesthetics-V2-6.5plus dataset to evaluate image generation. The LAION-Aesthetics-V2-6.5plus is a subset of the LAION 5B dataset, characterized by its high visual quality, where images have scored 6.5 or higher according to aesthetic prediction models. Regarding the evaluation of video generation, we use randomly sampled 512 captions from the WebVid-10M dataset [1].

B. More Evaluation

B.1. Comparison with Super-Resolution

Different from traditional super-resolution (SR) tasks. Higher-resolution generation aims to tap the potential of the pre-trained model itself. Therefore, the performance of the higher-resolution generation method is based on the base model rather than another additional SR model. We compare our method with a super-resolution post-processing setting: SDXL+Real-ESRGAN [52]. As shown in Table 4, FreeScale achieves competitive performance in quantitative metrics. As reported in most previously published related works [14, 20], higher-resolution generation methods are hard to beat SR methods completely on quantitative metrics due to the difference in difficulty between the two tasks. However, Figure 8 shows that FreeScale is not inferior to SDXL+Real-ESRGAN in visual quality, and adds more details. In addition, SR methods will faithfully follow the low-resolution input while FreeScale can regenerate the original blurred areas based on the prior knowledge that the model has learned (the eyes and logos in Figure 8).

Table 4. **Image quantitative comparisons with super-resolution.** Compared to super-resolution post-processing setting SDXL+Real-ESRGAN, FreeScale also achieves competitive performance. As reported in most previously published related works, higher-resolution generation methods are hard to beat SR methods completely on quantitative metrics due to the difference in difficulty between the two tasks.

Method	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑
SDXL+Real-ESRGAN [52]	43.476	0.000	73.524	0.024	12.599
Ours	49.796	0.004	71.369	0.029	12.572

Table 5. **User study.** Users are required to pick the best one among our proposed FreeScale with the other baseline methods in terms of image-text alignment, image quality, and visual structure.

Method	Text Alignment	Image Quality	Visual Structure
SDXL-DI [40]	0.87%	0.00%	0.00%
ScaleCrafter [20]	7.83%	5.22%	7.83%
DemoFusion [14]	17.39%	14.35%	18.26%
FouriScale [25]	2.17%	2.61%	1.74%
Ours	**71.74%**	**77.83%**	**72.17%**

Table 6. **User study for Video Generation.** Users are required to pick the best one among our proposed FreeScale with the other baseline methods in terms of text alignment, cover quality, and video quality.

Method	Text Alignment	Cover Quality	Video Quality
VC2-DI	5.38%	4.62%	3.85%
ScaleCrafter	4.62%	5.38%	0.77%
DemoFusion	30.00%	26.92%	30.77%
Ours	**60.00%**	**63.08%**	**64.62%**

B.2. User Study

In addition, we conducted a user study to evaluate our results on human subjective perception. Users are asked to watch the generated images of all the methods, where each

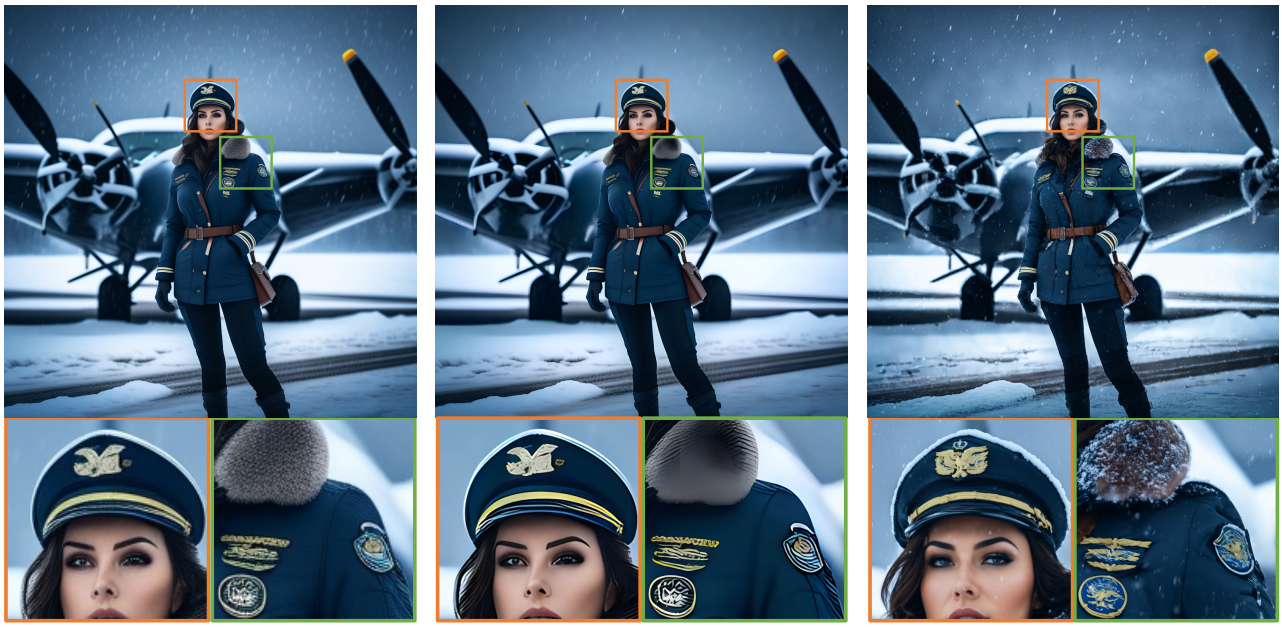

Figure 8. **Image qualitative comparisons with super-resolution.** FreeScale is not inferior to SDXL+Real-ESRGAN in visual quality, and adds more details. In addition, SR methods will faithfully follow the low-resolution input while FreeScale can regenerate the original blurred areas based on the prior knowledge that the model has learned. Best viewed **ZOOMED-IN**.

Table 7. **Video quantitative comparisons with other ablations.** Our final setting achieves the best or second-best scores for all metrics. The best results are marked in **bold**, and the second best results are marked by underline.

Method	FVD ↓	Dynamic Degree ↑	Aesthetic Quality ↑	Time (min) ↓
Dilated Up-Blocks	523.323	0.363	0.611	3.788
RGB Upsampling	**422.245**	0.381	0.604	3.799
Ours	484.711	**0.383**	**0.621**	**3.787**

example is displayed in a random order to avoid bias, and then pick the best one in three evaluation aspects. A total of 23 users were asked to pick the best one according to the image-text alignment, image quality, and visual structure, respectively. As shown in Table 5, our approach gains the most votes for all aspects, outperforming baseline methods by a large margin.

We also add a human study for video generation. Users were asked to pick the best one according to the text alignment, cover quality, and video quality, respectively. As shown in Table 6, our method still gains the most votes for all aspects, outperforming baseline approaches significantly.

B.3. Ablation Study for Video Generation

We also conduct an ablation study for higher-solution video generation. As discussed in the method part, we adopt latent space upsampling in video generation. Table 7 shows that our final setting achieves the best or second-best scores for all metrics.

Figure 9. **Flexible aspect ratio generation.** FreeScale can directly achieve a flexible aspect ratio (the resolution must be a multiple of 512) without any adaptation.

C. More Results

C.1. Flexible Aspect Ratio Generation

As shown in Figure 9, FreeScale can directly achieve a flexible aspect ratio (the resolution must be a multiple of 512) without any adaptation. We also add quantitative experiments for 2048 × 4096 resolution. As shown in Table 8, FreeScale still achieves the best or second-best scores for all metrics.

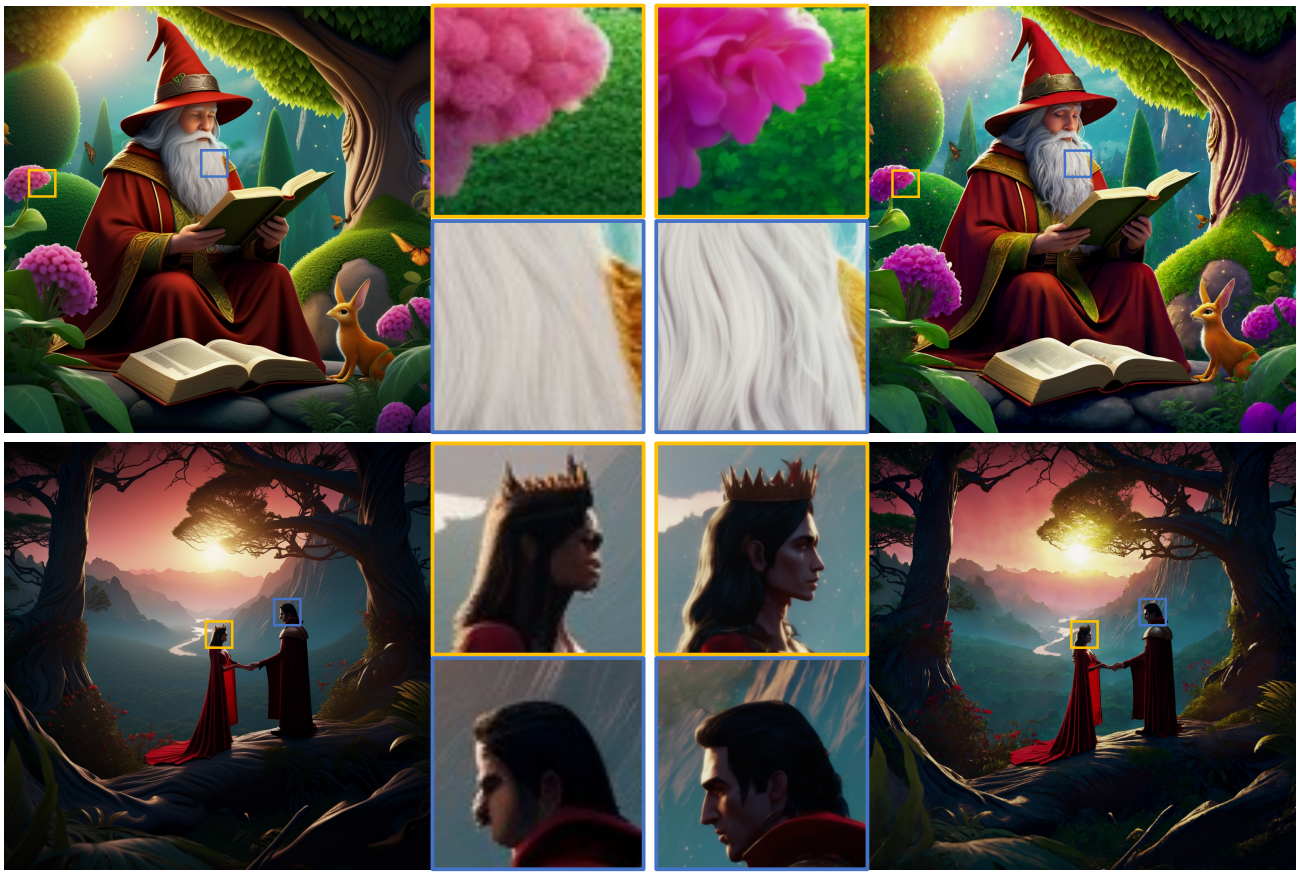

Figure 10. **Zoomed in details for the 8k image.** FreeScale may regenerate the original blurred areas at low resolution based on the prior knowledge that the model has learned. As shown in the bottom row, two originally chaotic and blurry faces are clearly outlined at 8k resolution. Best viewed **ZOOMED-IN**.

Figure 11. **Fast generation with SDXL-Turbo.** FreeScale can help SDXL-Turbo generate results at 2048^2 resolution with even 2 timesteps.

C.2. Fast Generation with SDXL-Turbo

FreeScale can easily be compatible with other models with similar structures. SDXL-Turbo [45] is a distilled version of SDXL [40] and can produce similar quality results with $2 \sim 4$ timesteps. However, SDXL-Turbo can only generate results at 512^2 resolution due to the knowledge loss during distillation. As shown in Figure 11, FreeScale can help SDXL-Turbo generate results at 2048^2 resolution.

Table 8. **Image quantitative comparisons with baselines in 2048×4096 resolution.** FreeScale still achieves the best or second-best scores for all metrics.

Method	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑
SDXL-DI	97.493	0.026	**38.273**	**0.009**	7.258
ScaleCrafter	97.235	0.032	107.582	0.050	8.001
DemoFusion	<u>72.196</u>	<u>0.019</u>	91.264	0.044	<u>10.622</u>
FouriScale	95.891	0.032	118.306	0.061	8.422
Ours	**54.704**	**0.004**	<u>65.584</u>	<u>0.025</u>	**11.323**

C.3. Gallery of 8k Images

Figure 12 illustrates the effectiveness of FreeScale on generating ultra-high-resolution images (*i.e.*, 8k-resolution images). As shown in Figure 10, FreeScale effectively enhances local details without compromising the original visual structure or introducing object repetitions. Different from simple super-resolution, FreeScale may regenerate the original blurred areas at low resolution based on the prior knowledge that the model has learned. In Figure 10, two

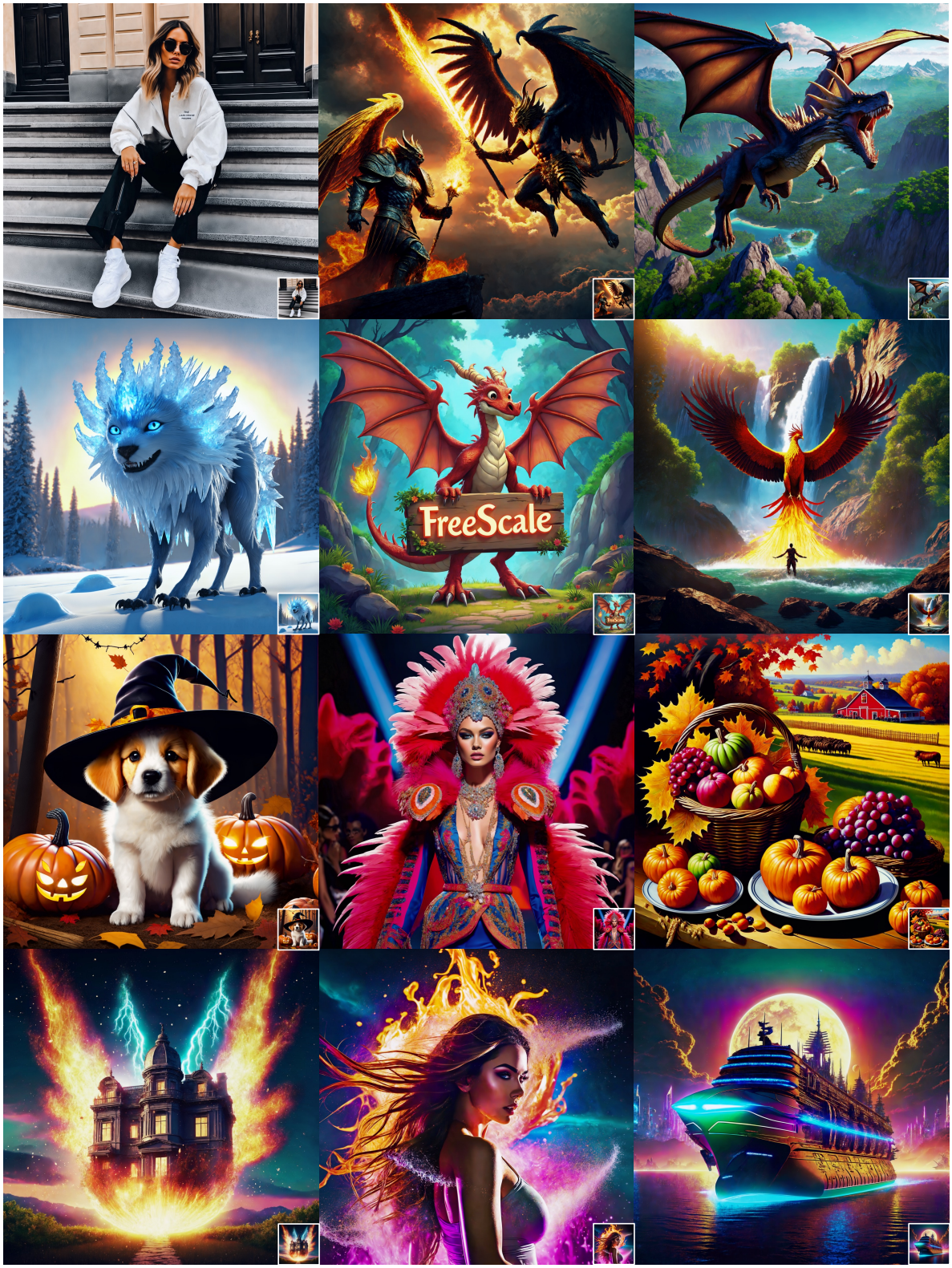

Figure 12. **Gallery of generated 8k images.** We place the original-resolution result in the lower right corner for reference. FreeScale effectively enhances local details without compromising the visual structure or introducing object repetitions. Best viewed **ZOOMED-IN**.

Figure 13. **Structure gap.** UNet-based LDMs and DiT-based LDMs will face different challenges in the higher-resolution generation task. UNet-based LDMs face repetition problems while DiT-based LDMs face blur problems.

originally chaotic and blurry faces are clearly outlined at 8k resolution.

Visual Enhancement. FreeScale also supports using existing images to replace the intermediate $1\times$ result. Compared to SDXL [40], FLUX [33] is better in visual text generation. In the center of Figure 12, we first use FLUX to generate the intermediate $1\times$ result, a dragon with "FreeScale". Then we utilize the remaining pipeline of FreeScale to generate the final 8k-resolution result. In this sense, FreeScale is also a tool to upscale resolution and enhance detail.

D. Limitations and Future Work

Inference Cost. We employ the scale fusion only in the self-attention layers thus bringing negligible time cost. And the omitted time steps almost offset the additional cost of tailored self-cascade upscaling. As a result, the inference cost of FreeScale is close to the direct inference by the base model. However, the inference cost is still huge for ultra-high-resolution generation. In future work, when users require image generation at resolutions exceeding 8k, memory constraints may be mitigated through multi-GPU inference strategies, while computational efficiency can be enhanced by employing inference acceleration techniques.

Knowledge Limitation. Even ignoring the limitations of the computation, there is a limit to the upscaling capability of FreeScale. When the desired resolution is beyond the prior knowledge that the model has learned, no more details can be reasonably added. In other words, the endless higher-resolution result will have either the same level of detail or unnatural messy detail. In addition, as a tuning-free framework, FreeScale's performance relies heavily on base models. During the tailored self-cascade process, the intermediate $1\times$ result is equivalent to direct inference with base models. Some artifacts caused by inherently flawed (*e.g.*, extra legs), will be inherited in further upscaling.

Structure Gap. DiT-based LDMs (*e.g.*, FLUX [33] and CogVideoX [55]), have showcased impressive visual generation capabilities recently. However, UNet-based LDMs and DiT-based LDMs will face different challenges in the higher-resolution generation task. As shown in Figure 13, UNet-based LDMs face repetition problems while DiT-based LDMs face blur problems. Most previous higher-resolution generation methods either support the UNet-based LDMs (DemoFusion [14], and FouriScale [25]) or DiT-based LDMs (I-MAX [15]), in line with the common sense that different problems require different strategies to solve. To support the DiT based structure, FreeScale also needs to be customized specifically.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2, 3, 5

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 5

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[7] Boyuan Cao, Jiaxin Ye, Yujie Wei, and Hongming Shan. Ap-ldm: Attentive and progressive latent diffusion model for training-free high-resolution image generation. *arXiv preprint arXiv:2410.06055*, 2024. 3

[8] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, pages 170–188. Springer, 2022. 5

[9] 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. 2

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2:

Overcoming data limitations for high-quality video diffusion models, 2024. 1, 2, 5, 7

[11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1, 2

[12] Jiaxiang Cheng, Pan Xie, Xin Xia, Jiashi Li, Jie Wu, Yuxi Ren, Huixia Li, Xuefeng Xiao, Min Zheng, and Lean Fu. Resadapter: Domain consistent resolution adapter for diffusion models. 2024. 3

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[14] 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. 2, 3, 4, 5, 7, 1

[15] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*, 2024. 5

[16] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. *arXiv preprint arXiv:2402.10491*, 2024. 2, 3, 4

[17] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[18] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6603–6612, 2024. 3

[19] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022. 2

[20] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4, 5, 7, 1

[21] 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. 5

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[24] 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

[25] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. *arXiv preprint arXiv:2403.12963*, 2024. 2, 3, 5, 7, 1

[26] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023. 5

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

[28] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. 2024. 2

[29] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36:70847–70860, 2023. 3

[30] Gwanghyun Kim, Hayeon Kim, Hoigi Seo, Dong Un Kang, and Se Young Chun. Beyondscene: Higher-resolution human-centric scene generation with pretrained diffusion. In *European Conference on Computer Vision*, pages 126–142. Springer, 2024. 3

[31] Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. *arXiv preprint arXiv:2406.18459*, 2024. 3

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 7

[33] Black Forest Labs. Flux.1 : An advanced state-of-the-art generative deep learning model. Technical report, Black Forest Labs, 2024. 5

[34] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36:50648–50660, 2023. 3

[35] Mingbao Lin, Zhihang Lin, Wengyi Zhan, Liujuan Cao, and Rongrong Ji. Cutdiffusion: A simple, fast, cheap, and strong diffusion extrapolation method. *arXiv preprint arXiv:2404.15141*, 2024. 3

[36] Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. Accdiffusion: An accurate method for higher-resolution image generation. *arXiv preprint arXiv:2407.10738*, 2024. 3

[37] Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. 2024. 3

[38] Xinyu Liu, Yingqing He, Lanqing Guo, Xiang Li, Bu Jin, Peng Li, Yan Li, Chi-Min Chan, Qifeng Chen, Wei Xue, et al. Hiprompt: Tuning-free higher-resolution generation with hierarchical mllm prompts. *arXiv preprint arXiv:2409.02919*, 2024. 3

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 5, 3

[41] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 5

[42] 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. 3

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[44] 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. 5

[45] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 3

[46] 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

[47] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *CVPR*, 2024. 2, 5, 7

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[49] 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. *arXiv preprint arXiv:2309.03350*, 2023. 3

[50] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[51] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report, 2023. 1, 2

[52] 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*, pages 1905–1914, 2021. 1

[53] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 2023. 2

[54] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 2

[55] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 5

[56] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *CVPR*, 2024.

[57] 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. 1, 2

[58] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, pages 145–161. Springer, 2024. 3

[59] Qingping Zheng, Yuanfan Guo, Jiankang Deng, Jianhua Han, Ying Li, Songcen Xu, and Hang Xu. Any-size-diffusion: Toward efficient text-driven synthesis for any-size hd images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7571–7578, 2024. 3