EPiC Series in Technology

Volume 3, 2026, Pages 1–10

Proceedings of AI for Accelerated Research Symposium

HiPA: Enabling One-Step Text-to-Image Diffusion via High-Frequency Promotion

Yifan Zhang[1,2], Bryan Hooi[2], Shuicheng Yan[2]

[1] MiroMind AI,
[2] National University of Singapore
yifan.zhang@miromind.ai

Abstract

Diffusion models have revolutionized text-to-image generation, but their real-world applications are hampered by the extensive inference time needed for hundreds of diffusion steps. Although progressive distillation and consistency distillation have been proposed to speed up diffusion sampling to 2-8 steps, they still fall short in one-step generation due to poor abilities to generate high-frequency content. To overcome this, we introduce High-frequency-Promoting Adaptation (HiPA), a parameter-efficient approach to enable one-step text-to-image diffusion. Grounded in the insight that high-frequency information is essential but highly lacking in one-step diffusion, HiPA aims to train one-step, low-rank adaptors to specifically enhance the under-represented high-frequency abilities of advanced diffusion models. The learned adaptors empower these diffusion models to generate high-quality images in just a single step. Compared with progressive distillation, HiPA achieves much better performance in one-step text-to-image generation (37.3 → 23.8 in FID-5k on MS-COCO 2017) and 28.6x training speed-up (108.8 → 3.8 A100 GPU days), requiring only 0.04% training parameters (7,740 million → 3.3 million). We also demonstrate HiPA's effectiveness in text-guided image editing, inpainting and super-resolution tasks, where our adapted models consistently deliver high-quality outputs in just one diffusion step.

1 Introduction

Text-to-image generation [1, 2, 3], aiming at synthesizing images from textual descriptions, has undergone a significant transformation with the advent of diffusion models [4, 5, 6, 7, 8, 9, 10]. These models, known for their multi-step denoising process, have set new benchmarks in the quality of generated images, marked by increased fidelity and detail [11]. However, the necessity for multiple diffusion steps, each meticulously refining the image, results in significantly long generation time. This diminishes the practicality of text-to-image diffusion models for real-time applications [12, 13, 14, 15, 16].

To mitigate this issue, Progressive distillation (PD) [12, 17, 18, 19] distills a T-step teacher into a $T/2$-step student iteratively, achieving 2–8 step generation but at high computational and parameter cost due to repeated training. Consistency distillation [13, 20] accelerates unguided diffusion, but its efficacy for text-to-image remains unclear. Extensions like latent consistency distillation [21] and LCM-LoRA [22] enable 2–4 step text-to-image generation, yet struggle with

J. Masnec, H.R. Karimian, P. Kordjamshidi and Y. Li (eds.), AIAS 2025 (EPiC Series in Technology, vol. 3), pp. 1–10

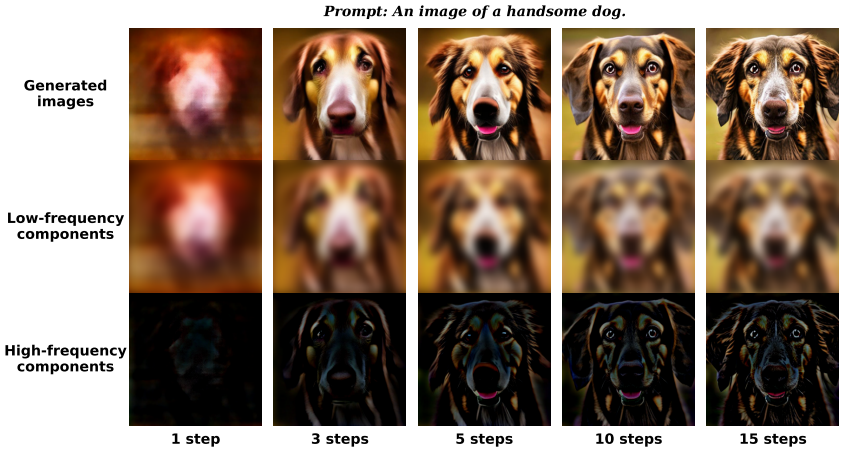

Figure 1: Illustration of text-to-image generation with different diffusion steps based on Stable Diffusion [1]. The capability for low-frequency content generation is attained with fewer diffusion steps, whereas the capability for high-frequency detail generation demands an increased number of steps. Notably, one-step diffusion images lack the complex high-frequency components, making them noticeably less realistic.

one-step high-frequency detail. Adversarial diffusion [23] improves quality but trains the full student model, reducing efficiency.

In this work, we focus on optimizing text-to-image diffusion models for one-step generation, explicitly aiming to streamline the inference efficiency compared to conventional multi-step generation. To figure this out, we delve into the multi-step generation process of text-to-image diffusion models, aiming to uncover what information one-step diffusion lacks, compared to its multi-step counterpart. As shown in Figure 1, we identify a critical aspect of the text-to-image generation process: the ability to generate low-frequency content is achieved with fewer diffusion steps, while the ability of high-frequency detail generation requires more steps. It is worth noting that one-step diffusion often struggles to produce rich high-frequency details, which, however, are essential for realistic image generation. Existing acceleration techniques, such as progressive distillation [17, 18], consistency distillation [13, 20, 24] and adversarial diffusion [23], overlook this crucial aspect, thus sacrificing high-frequency detail generation in one-step diffusion and leading to limited image quality.

In light of these findings, we propose a parameter-efficient High-frequency-Promoting Adaptation (HiPA) approach to accelerate existing advanced multi-step text-to-image generation models to one-step diffusion. Instead of conducting slow progressive distillation by training multiple student models with extensive parameters, HiPA trains low-rank adaptors to enable one-step diffusion, particularly promoting high-frequency detail generation abilities. Central to HiPA is a new diffusion adaptation loss, consisting of a spatial perceptual loss and a high-frequency promoted loss. The spatial perceptual loss is employed to ensure the adapted model has general generation abilities, affirming its structural coherence and semantic consistency in the generated images. Meanwhile, the high-frequency promoted loss, leveraging Fourier transform and edge detection, is specifically designed to enhance the subtle, yet crucial, high-frequency generation abilities. This dual-loss strategy effectively preserves detailed textures and edges that are often overlooked in one-step diffusion, facilitating rapid generation without significantly compromising image quality.

Our approach is validated through extensive experiments in one-step text-to-image gener-

Figure 2: Illustration of the impact of high-frequency components in enhancing image clarity for one-step text-to-image diffusion. Combining the high-frequency components from the 15-step images with the the low-frequency components from fewer-step images results in sharper images after Inverse Fourier Transform, while using one-step high-frequency components provides no clarity enhancement.

ation, demonstrating that HiPA outperforms compared methods in both visual fidelity and training efficiency, while requiring much fewer training parameters. As illustrated in Table 2b, compared to progressive distillation, HiPA significantly improves one-step text-to-image generation performance (37.3 to 23.8 in FID-5k on MS-COCO 2017), accelerates training by 28.6 times (108.8 to 3.8 A100 GPU days), and drastically reduces training parameter needs (7,740 million to just 3.3 million). To showcase HiPA's versatility, we extend its application to text-guided image editing, inpainting, and super-resolution tasks, where we reduce the number of diffusion steps to a single step. Promising results demonstrate HiPA's remarkable potential for efficient and practical use in various real-world applications.

2 Preliminary Studies

2.1 One-step diffusion lacks high-frequency generation abilties

To advance one-step text-to-image diffusion, we begin with an analysis of Stable Diffusion (SD) [1], aiming to dissect the nuances of images produced at different diffusion steps. As shown in Figure 1 (first row), we identify an essential characteristic of text-to-image generation: the images generated by one-step SD are notably blurry, and their quality dramatically improves with an increase in diffusion steps. To delve into this phenomenon, we leverage Discrete Fourier Transform [25] to differentiate between high and low-frequency information within the image, and then employ Inverse Fourier Transform [25] to reconstruct images for visualization. Figure 1 (second row) shows that the model's capability to generate low-frequency content, such as the foundational elements and the underlying scene of the image, is achieved with fewer diffusion steps. This indicates an efficient grasp of the image's basic structure and thematic essence early in the diffusion process. Conversely, the ability to generate high-frequency details, which involves refining textures and adding intricate elements to enhance image realism and complexity, requires a greater number of diffusion steps for precision (see Figure 1, last row).

High-frequency abilities matter. To explore the role of high-frequency information in few-step generation, we conduct an experiment mixing high and low-frequency components of

Methods	FID-30k ↓	IS ↑	CLIP ↑
Stable Diffusion	355.2	2.0	0.11
• L2 spatial loss	131.2	8.3	0.16
• L2 spatial loss + low-frequency promotion	163.5	4.7	0.15
• L2 spatial loss + high-frequency promotion	**115.2**	**10.6**	**0.18**

Table 1: One-step generation results of Stable Diffusion adaptation on MS-COCO 2014.

images generated at different steps. Using Discrete Fourier Transform, we extract high- and low-frequency components from 1–15 step images, cross-combine them (e.g., high-frequency from 1 or 15 steps with low-frequency from 1, 2, 3, 5, or 10 steps), and reconstruct the results via Inverse Fourier Transform. As shown in Figure 2, high-frequency components from the 15-step image significantly enhance clarity and quality when combined with low-frequency components from fewer steps, while high-frequency components from the 1-step image yield blurry results regardless of the low-frequency source. This highlights the critical importance of high-frequency generation in text-to-image diffusion and its potential to improve one-step synthesis quality.

2.2 Promoting high-frequency generation boosts one-step diffusion

Based on the above observations, we hypothesize that enhancing high-frequency generation can improve one-step text-to-image diffusion. To validate this, we adapt Stable Diffusion (SD) by aligning one-step outputs with multi-step (e.g., 10 or 15 steps) outputs using different losses: L2 spatial loss, L2 with low-frequency promotion, and L2 with high-frequency promotion. High- and low-frequency promotion are achieved by aligning the respective Fourier-reconstructed components between one- and multi-step images. As shown in Table 1, L2 loss alone improves over the SD baseline, while adding low-frequency promotion degrades performance, suggesting it may hinder one-step diffusion. In contrast, augmenting L2 with high-frequency promotion yields the best results, improving realism (lower FID), diversity (higher IS), and textual fidelity (higher CLIP), confirming that promoting high-frequency generation is key to enhancing one-step diffusion.

3 Our Approach

Overall scheme. In light of the aforementioned insights, we propose a new parameter-efficient strategy, High-frequency-Promoting Adaptation (HiPA), to enable one-step text-to-image diffusion. Our approach diverges from previous methods like Progressive Distillation [17] and Adversarial Distillation [23], which focus on tuning the entire pre-trained diffusion models. Instead, HiPA aims to train a low-rank adaptor to enhance the one-step generation abilities of diffusion models. As shown in Figure 3, HiPA achieves this by aligning the images generated in a single step from the adapted model with those produced by the original, frozen model across multiple steps (*e.g.,* 15 steps). To promote high-frequency abilities, HiPA applies a composite adaptation loss. This loss integrates spatial perceptual loss with high-frequency promoted loss, collectively refining one-step generation for improved fidelity and high-frequency details:

$$L_{\text{adaptation}} = L_{\text{spatial}} + L_{\text{high-freq}}. \tag{1}$$

Spatial perceptual loss L_{spatial}. The spatial adaptation loss is defined as follows:

$$L_{\text{spatial}} = L\left(I^{\text{1-step}}_{\text{generated}}, I^{\text{multi}}_{\text{generated}}\right), \tag{2}$$

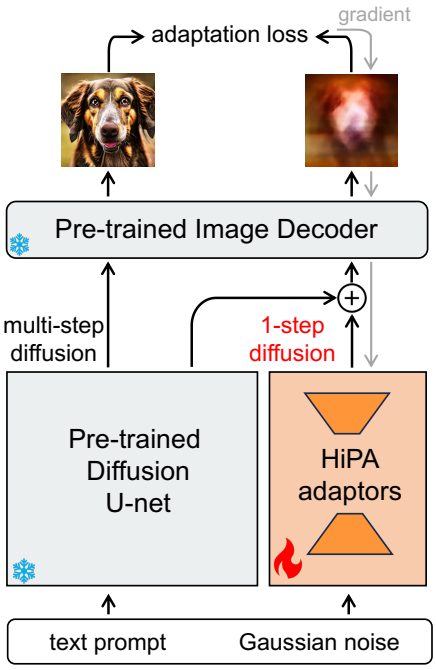

Figure 3: An illustration of our parameter-efficient HiPA approach.

where $I^{\text{1-step}}_{\text{generated}}$ and $I^{\text{multi-step}}_{\text{generated}}$ are the images generated by the HiPA-adapted model in a single step and those by the original diffusion model with multiple steps. Here, $L(\cdot,\cdot)$ can be any distance loss. In this work, we adopt the Deep Image Structure and Texture Similarity (DISTS) metric [26], since it goes beyond pixel-level differences to capture perceptual dissimilarities between images, considering both structural and textural characteristics.

High-frequency promoted loss $L_{\text{high-freq}}$. To effectively promote high-frequency abilities, we apply two complementary strategies to extract high-frequency information: Fourier transform and edge detection.

For the Fourier strategy, we first apply Discrete Fourier Transform (DFT) [25] to the generated image $I_{\text{generated}}$, transforming it from the spatial to the frequency domain. The high-frequency components are then extracted through high-pass filtering, followed by Inverse Fourier Transform (IFT)[25] to reconstruct the high-frequency image I_{freq}. This process can be described by: $I_{\text{freq}} = \text{IFT}\Big(\text{DFT}(I_{\text{generated}}) \odot M_{\text{high}}(u,v)\Big)$, where $M_{\text{high}}(u,v)$ is a high-pass filter in frequency domain.

Meanwhile, we apply the Sobel operator [27] to extract edge information of images. This operator computes the image gradient ∇I through convolution of I with pre-defined horizontal and vertical kernels G_x and G_y, thereby highlighting significant intensity transitions. The detected edge image I_{edge} is computed by: $I_{\text{edge}} = \sqrt{(I_{\text{generated}} * G_x)^2 + (I_{\text{generated}} * G_y)^2}$, where $*$ represents convolution, and the Sobel kernels are defined as $G_x = [-1,0,1;-2,0,2;-1,0,1]$ and $G_y = [-1,-2,-1;0,0,0;1,2,1]$.

Based on the extracted high-frequency information, we design the high-frequency promoted loss by aligning the high-frequency details of the one-step images with those of the multi-step

Figure 4: One-step text-guided image generation on MS-COCO 2014 (512×512) by Stable Diffusion (SD), Latent Consistency Distillation, and HiPA. These visual results show that our approach can generate high-quality images in a single diffusion step.

images:

$$L_{\text{high-freq}} = L\left(I_{\text{freq}}^{\text{1-step}}, I_{\text{freq}}^{\text{multi}}\right) + L\left(I_{\text{edge}}^{\text{1-step}}, I_{\text{edge}}^{\text{multi}}\right). \tag{3}$$

We also adopt the DISTS metric as $L(\cdot,\cdot)$ for the loss. As a result, the learned HiPA adaptor enables the adapted one-step model to emulate the superior quality of its multi-step counterpart, particularly in high-frequency generation abilities.

4 Experiments

In this section, we evaluate the effectiveness and versatility of our method in one-step text-to-image diffusion. We mainly use the MS-COCO 2017 training set [28] for diffusion model adaptation, and use the COCO 2014/2017 validation set for evaluation. This dataset offers diverse textual-visual content, making it an ideal benchmark for method evaluation. Moreover, we use three main evaluation metrics: Fréchet Inception Distance (FID), Inception Score (IS), and CLIP score (ViT-g/14).

4.1 One-step text-guided image generation

In this work, we focus on adapting Stable Diffusion (SD) v2.1 [1] for fast generating high-fidelity images from text descriptions. Our method trains the adaptors for 5 epochs on the COCO

Methods	Step	FID-30k ↓	IS ↑	CLIP ↑	Inference time
Stable Diffusion [1]	25	9.40	31.83	0.29	8.6 hours
Stable Diffusion [1]	1	355.21	1.97	0.11	2.3 hours
Latent Consistency [21]	1	195.35	4.46	0.20	3.1 hours
LCM-LoRA [22]	1	94.72	9.45	0.22	3.2 hours
HiPA (ours)	1	13.91	28.09	0.31	2.5 hours

(a) One-step generation performance on COCO 2014. Guidance scale is 2. Inference time (30k images) is on A100 GPU. All methods are based on Stable Diffusion.

Methods	Step	FID-5k ↓	Training time	# Params. trained
Stable Diffusion [1]	8	31.7	6,250 Days	860M
SnapFusion [18]	8	24.2	-	3×848M
Progressive Distillation [17]	1	37.2	108.8 Days	9×860M
2-Rectified Flow [29]	1	47.0	75.2 Days	860M
InstaFlow-0.9B [29]	1	23.4	199.2 Days	860M
Latent Consistency [21]	1	204.0	1.3 Days	860M
SD-Turbo [23]	1	29.9	-	860M
LCM-LoRA [22]	1	153.0	-	67.5M
TCD [24]	1	80.6	5 Days	64.6M
HiPA (ours)	1	23.8	3.8 Days	3.3M

(b) Generation performance on COCO 2017 and training costs (A100 GPU days).

Table 2: Comparison of one-step text-to-image generation methods.

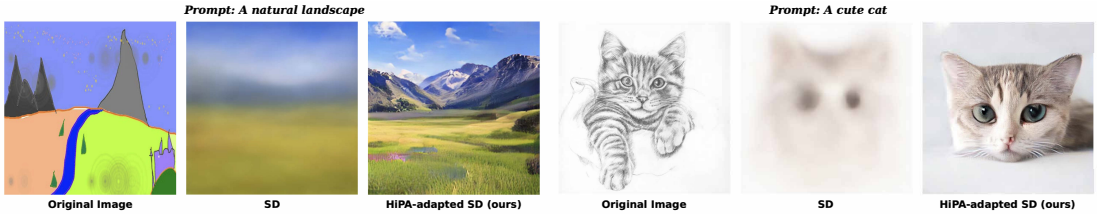

Figure 5: One-step text-guided image editing by the original Stable Diffusion (SD) and our HiPA-adapted SD model.

2017 training set, followed by qualitative and quantitative assessments. The one-step generated images, as shown in Figure 4, show a remarkable improvement in quality over one-step SD, Latent Distillation [21], LCM-LoRA [22], TCD [24] and SD-Turbo [23]. This is backed by the quantitative results in Table 2, where HiPA outperforms existing methods across FID, IS, and CLIP. These improvements are not just numerical; they translate into perceptibly more realistic images (lower FID), greater diversity (higher IS), and better alignment with textual descriptions (higher CLIP). Importantly, HiPA enables one-step text-to-image diffusion, offering a considerable boost in inference efficiency over the original multi-step SD. This advancement is attained with only a slight reduction in performance, effectively balancing speed and quality, making it highly suitable for real-world applications.

A more critical superiority of HiPA lies in its training efficiency (cf. Table 2b). Unlike Progressive Distillation that necessitates training multiple student models (up to 9 for one-step generation), HiPA solely trains low-rank adaptors. This reduces the training time to a mere 3.8 A100 GPU days, compared to Progressive Distillation's extensive 108.8 days, and also significantly cuts down on the number of training parameters—from hundreds of millions in other models to just 3.3 million in HiPA. These advantages position HiPA as an effective and efficient solution, setting a new standard in one-step text-to-image diffusion.

4.2 One-step text-guided image editing

We proceed with experiments on one-step text-guided image editing based on SDEdit [30]. Following SDEdit, we first perturb the input image with Gaussian noise in the latent space, and then apply the original or our HiPA-adapted SD models for one-step diffusion. As showcased in Figure 5, our adapted model yields high-quality style-transfer images in just a single step, demonstrating a noticeable improvement over the one-step image editing capabilities of the standard SD.

Figure 6: One-step text-guided image inpainting by Stable Diffusion (SD) inpainting model and our HiPA-adapted SD inpainting model.

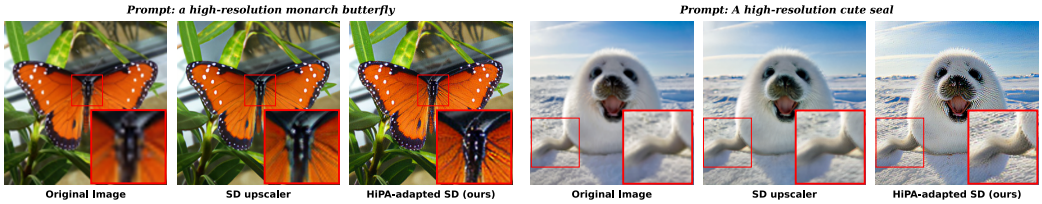

Figure 7: One-step text-guided image super-resolution by Stable Diffusion (SD) upscaler model and our HiPA-adapted upscaler model.

4.3 One-step text-guided image inpainting

We next apply HiPA to accelerate latent inpainting diffusion, based on the SD v2 inpainting model. This model is a fine-tuned variant of SD, enhanced to handle masks and masked images with additional input channels. Specifically, we apply HiPA to adapt the inpainting model on COCO 2017 for 1 epoch, where we retain the central image content while masking out 50% of the peripheral pixels. We train HiPA adaptors to effectively inpaint these masked regions in just one diffusion step, aligning the output with that of the original inpainting model's 15-step diffusion. The qualitative results in Figure 6 highlight HiPA's capability to facilitate fast, effective text-guided image inpainting for real applications.

4.4 One-step text-guided super-resolution

We further extend HiPA to accelerate latent super-resolution diffusion, based on the widely available SD 4x-upscaler model. This model, a specialized variant of SD, is tailored for text-guided super-resolution. We adapt this model on COCO 2017 for 2,000 iterations, beginning with images downsampled to 128×128. We train the HiPA adaptors to upscale these images to 512×512 in a single diffusion step by aligning their outputs with those from the original model's 15-step diffusion. The visualized results in Figure 7 highlight the capacity of HiPA to perform fast text-guided image super-resolution in real scenarios.

5 Conclusion

To advance one-step text-to-image diffusion, we have introduced High-frequency-Promoting Adaptation (HiPA). HiPA adeptly addresses the computational and qualitative dilemmas posed by existing text-to-image diffusion models. By integrating parameter-efficient adaptation with high-frequency promotion, HiPA not only accelerates the text-to-image diffusion process, but also amplifies the high-frequency details essential for generating photorealistic images. Our empirical evidence underscores HiPA's superiority over conventional methods, demonstrating it as a practical and parameter-efficient approach in real-time text-to-image diffusion generation.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[2] 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 *Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, 2023.

[8] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

[9] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024.

[10] Yifan Zhang and Bryan Hooi. Hipa: Enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. *arXiv preprint arXiv:2311.18158*, 2023.

[11] James Betker, Gabriel Goh, et al. Improving image generation with better captions. In *OpenAI Technical Report*, pages 8821–8831, 2023.

[12] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.

[13] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

[14] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. In *Advances in Neural Information Processing Systems*, 2023.

[15] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.

[16] Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In *Advances in Neural Information Processing Systems*, volume 34, pages 29848–29860, 2021.

[17] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[18] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In *Advances in Neural Information Processing Systems*, 2023.

[19] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel

Zheng, Walter Talbot, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

[20] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arxiv*, 2023.

[21] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[22] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[23] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv*, 2023.

[24] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation. *arXiv*, 2024.

[25] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[26] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

[27] Kenneth R Castleman. *Digital image processing*. Prentice Hall Press, 1996.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

[29] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023.

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.