Your Student is Better Than Expected: Adaptive Teacher-Student Collaboration for Text-Conditional Diffusion Models

Nikita Starodubcev Dmitry Baranchuk Artem Fedorov Artem Babenko
Yandex Research

https://github.com/yandex-research/adaptive-diffusion

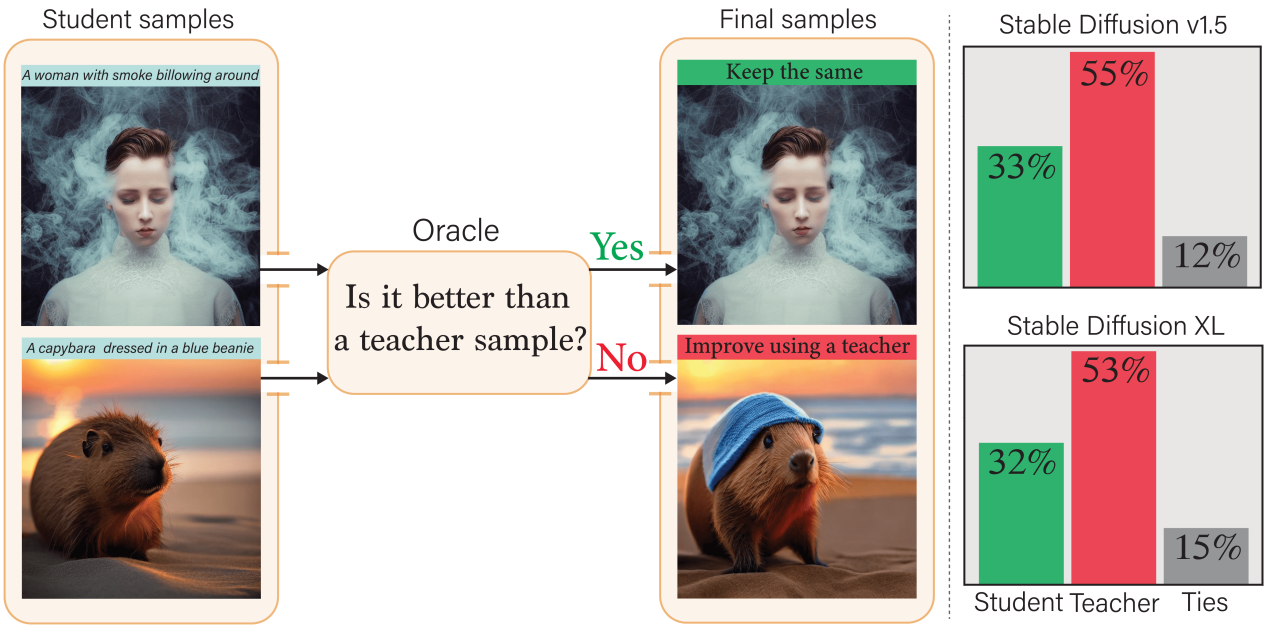

Figure 1. *Left:* Overview of the proposed approach. *Right:* Side-by-side comparison of SDv1.5 and SDXL with their few-step distilled versions. The distilled models surpass the original ones in a noticeable number of samples for the same text prompts and initial noise.

Abstract

Knowledge distillation methods have recently shown to be a promising direction to speedup the synthesis of large-scale diffusion models by requiring only a few inference steps. While several powerful distillation methods were recently proposed, the overall quality of student samples is typically lower compared to the teacher ones, which hinders their practical usage. In this work, we investigate the relative quality of samples produced by the teacher text-to-image diffusion model and its distilled student version. As our main empirical finding, we discover that a noticeable portion of student samples exhibit superior fidelity compared to the teacher ones, despite the "approximate" nature of the student. Based on this finding, we propose an adaptive collaboration between student and teacher diffusion models for effective text-to-image synthesis. Specifically, the distilled model produces an initial image sample, and then an oracle decides whether it needs further improvements with the teacher model. Extensive experiments demonstrate that the designed pipeline surpasses state-of-the-art text-to-image alternatives for various inference budgets in terms of human preference. Furthermore, the proposed approach can be naturally used in popular applications such as text-guided image editing and controllable generation.

1. Introduction

Large-scale diffusion probabilistic models (DPMs) have recently shown remarkable success in text-conditional image generation [32, 34, 38, 41] that aims to produce high quality images closely aligned with the user-specified text prompts. However, DPMs pose sequential synthesis leading to high inference costs opposed to feed-forward alternatives, e.g.,

GANs, that provide decent text-to-image generation results for a single forward pass [17, 43].

There are two major research directions mitigating the sequential inference problem of state-of-the-art diffusion models. One of them considers the inference process as a solution of a probability flow ODE and designs efficient and accurate solvers [18, 26, 27, 49, 58] reducing the number of inference steps down to $\sim$10 without drastic loss in image quality. Another direction represents a family of knowledge distillation approaches [12, 24, 25, 28, 30, 42, 44, 50] that learn the student model to simulate the teacher distribution requiring only $1-4$ inference steps. Recently, distilled text-to-image models have made a significant step forward [25, 28, 30, 44]. However, they still struggle to achieve the teacher performance either in terms of image fidelity and textual alignment [25, 28, 30] or distribution diversity [44]. Nevertheless, we hypothesize that text-to-image students may already have qualitative merits over their teachers. If so, perhaps it would be more beneficial to consider a teacher-student collaboration rather than focusing on replacing the teacher model entirely.

In this paper, we take a closer look at images produced by distilled text-conditional diffusion models and observe that the student can generate some samples even better than the teacher. Surprisingly, the number of such samples is significant and sometimes reaches up to half of the empirical distribution. Based on this observation, we design an adaptive collaborative pipeline that leverages the superiority of student samples and outperforms both individual models alone for various inference budgets. Specifically, the student model first generates an initial image sample given a text prompt, and then an "oracle" decides if this sample should be updated using the teacher model at extra compute. The similar idea has recently demonstrated its effectiveness for large language models (LLMs) [5] and we show that it can be naturally applied to text-conditional diffusion models as well. Our approach is schematically presented in Figure 1. To summarize, our paper presents the following contributions:

- We reveal that the distilled student DPMs can outperform the corresponding teacher DPMs for a noticeable number of generated samples. We demonstrate that most of the superior samples correspond to the cases when the student model significantly diverges from the teacher.
- Based on the finding above, we develop an adaptive teacher-student collaborative approach for effective text-to-image synthesis. The method not only reduces the average inference costs but also improves the generative quality by exploiting the superior student samples.
- We provide an extensive human preference study illustrating the advantages of our approach for text-to-image generation. Moreover, we demonstrate that our pipeline can readily improve the performance of popular text-guided image editing and controllable generation tasks.

2. Related work

Diffusion Probabilistic Models (DPMs) [14, 48, 49] represent a class of generative models consisting of *forward* and *reverse* processes. The *forward* process $\{\boldsymbol{x}_t\}_{[0,T]}$ transforms real data $\boldsymbol{x}_0 \sim p_{\text{data}}(\boldsymbol{x}_0)$ into the noisy samples $\boldsymbol{x}_t$ using the transition kernels $\mathcal{N}\left(\boldsymbol{x}_t \mid \sqrt{1-\sigma_t}\boldsymbol{x}_0, \sigma_t \mathbf{I}\right)$ specifying σ_t according to the selected *noise schedule*.

The *reverse* diffusion process generates new data points by gradually denoising samples from a simple (usually standard normal) distribution. This process can be formulated as a *probabilistic-flow ODE* (PF-ODE) [46, 49], where the only unknown component is a *score function*, which is approximated with a neural network. The ODE perspective of the reverse process fosters designing a wide range of the specialized solvers [15, 18, 26, 27, 46, 57, 58] for efficient and accurate sampling. However, for text-to-image generation, one still needs $\sim$25 and more steps for the top performance.

Text-conditional diffusion models can be largely grouped into *cascaded* and *latent* diffusion models. The cascaded models [32, 41] generate a sample in several stages using separate diffusion models for different image resolutions. The latent diffusion models [34, 37, 38] first generate a low-resolution latent variable in the VAE [21] space and then apply its feedforward decoder to map the latent sample to the high-resolution pixel space. Thus, the latent diffusion models have significantly more efficient inference thanks to a single forward pass for the upscaling step.

Along with cascaded models, there are other works combining several diffusion models into a single pipeline. Some methods propose to use distinct diffusion models at different time steps [1, 8, 23, 55]. Others [25, 34] consider an additional model to refine the samples produced with a base model. In contrast, our method relies on the connection between student and teacher models and adaptively improves only selected student samples to reduce the inference costs.

Text-to-image diffusion models have also succeeded in text-guided image editing and personalization [3, 10, 19, 29, 31, 40]. Moreover, some methods allow controllable generation via conditioning on additional inputs, e.g., canny-edges, semantic masks, sketches [52, 56]. Our experiments show that the proposed pipeline is well-suited to these techniques.

Distillation of diffusion models is another pivotal direction for efficient diffusion inference [2, 25, 30, 42, 44, 47, 50]. The primary goal is to adapt the diffusion model parameters to represent the teacher image distribution for $1-4$ steps. Recently, consistency distillation (CD) [50] have demonstrated promising results on both classical benchmarks [20, 47] and text-to-image generation [28] but fall short of the teacher performance at the moment. Concurrently, adversarial diffusion distillation [44] could outperform the SDXL-Base [34] teacher for 4 steps in terms of image quality and prompt alignment. However, it significantly reduces the diversity of generated samples, likely due

Figure 2. **Student outperforms its teacher (SD1.5)**. *Left:* Text-conditional image synthesis. *Right:* Text-guided image editing (SDEdit [29]). The images within each pair are generated for the same initial noise sample.

to the adversarial training [11] and mode-seeking distillation technique [35]. Therefore, it is still an open question if a few-step distilled model can perfectly approximate the diffusion model on highly challenging and diverse distributions that are currently standard for text-conditional generation [45].

3. Toward a unified teacher-student framework

Opposed to the purpose of replacing the expensive text-to-image diffusion models by more effective few-step alternatives, the present work suggests considering the distilled text-to-image models as a firm companion in a teacher-student collaboration.

In this section, we first explore the advantages of the distilled text-to-image models and then unleash their potential in a highly effective generative pipeline comprising the student and teacher models.

3.1. Delving deeper into the student performance

We start with a side-by-side comparison of the student and teacher text-to-image diffusion models. Here, we focus on Stable Diffusion v1.5[1] (SD1.5) as our main teacher model and distill it using consistency distillation [50]. The student details and sampling setup are presented in A. The similar analysis for a few other distilled models is provided in B.2.

In Figure 1 (Right), we provide the human votes for 600 random text prompts from COCO2014 [22] for SD1.5 and SDXL. The images within each pair are generated for the same initial noise sample. We observe that the students generally falls short of the teacher performance. However, interestingly, despite the initial intention to mimic the teacher model, $\sim$30% student samples were preferred over the teacher ones. A few visual examples in Figure 2 validate these results.

[1]https://huggingface.co/runwayml/stable-diffusion-v1-5

Therefore, we can formulate our first observation:

> The student can surpass its teacher in a substantial portion of image samples.

Below, we develop a more profound insight into this phenomenon.

Student-teacher similarity. First, we evaluate the student's ability to imitate the teacher. We compute pairwise distances between the student (S) and teacher (T) images generated for the same text prompts and initial noise. As a distance measure, we use DreamSim [9] tuned to be aligned with the human perception judgments. For evaluation, we consider 5000 prompts from the COCO2014 [22] validation split.

Primarily, we observe that many student samples are highly distinct from the teacher ones. A few image pairs are presented in Figure 3a. Figure 3c presents the human vote distribution for low ($0-20\%$), medium ($40-60\%$) and high ($80-100\%$) distance ranges. Interestingly, most of the student wins appear when its samples are highly different from the teacher ones. This brings us to our second observation:

> The student wins are more likely where its samples significantly differ from the teacher ones.

Also, we evaluate the relative gap in sample quality against the similarity between the teacher and student outputs. To measure the quality of individual samples, we use ImageReward [53], which shows a positive correlation with human preferences in terms of image fidelity and prompt alignment. The divergence in quality is calculated as the difference between the ImageReward scores for student and teacher samples. We observe that highly distinct samples

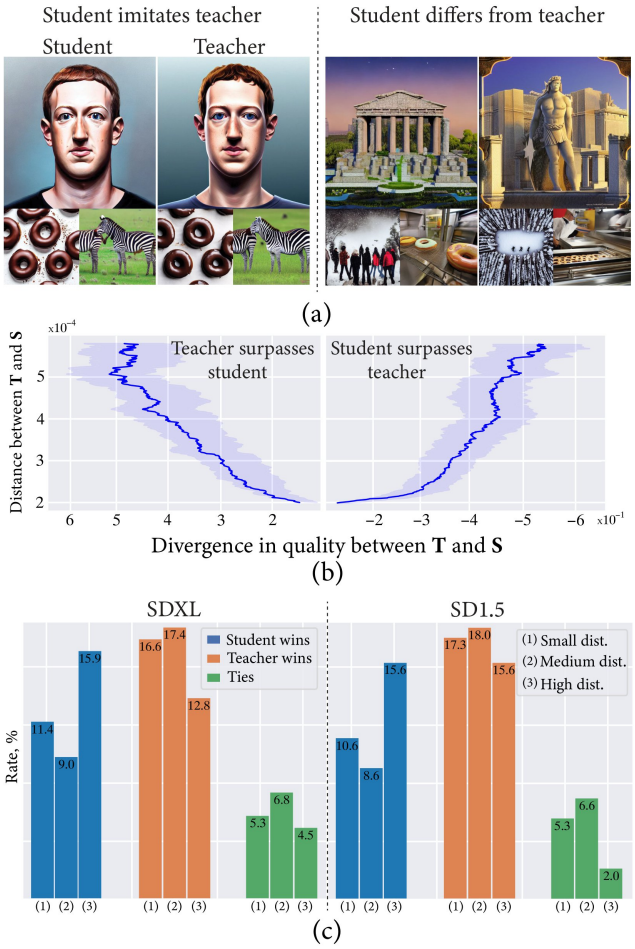

Figure 3. (a) Visual examples of similar (Left) and dissimilar (Right) teacher and student samples. (b) Similarity between the student and teacher samples w.r.t. the difference in sample quality. Highly distinct samples tend to be of different quality. (c) Human vote distribution for different distance ranges between student and teacher samples. Most of the student wins are achieved when the student diverges from the teacher.

likely have a significant difference in quality. Importantly, this holds for both student failures and successes, as shown in Figure 3b. Therefore, effectively detecting the positive student samples and improving the negative ones can potentially increase the generative performance.

Image complexity. Then, we describe the connection of the similarity between student and teacher samples with the teacher image complexity. To estimate the latter, we use the ICNet model [7] learned on a large-scale human annotated dataset. The results are presented in Figure 4. We notice that larger distances between student and teacher outputs are more typical for complex teacher samples. In other words, the student mimics its teacher for plain images, e.g., close-up faces, while acting more as an independent model for more intricate ones. Figure 4b confirms that significant changes in image quality are observed for more complex images.

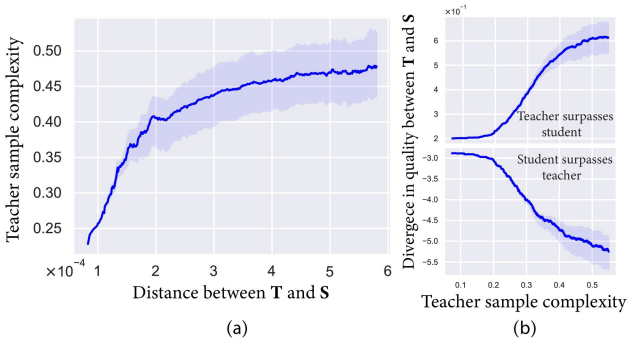

Figure 4. **Effect of image complexity**. (a) More similar student and teacher samples corresponds to simpler images and vice versa. (b) The student and teacher largely diverge in image quality on the complex teacher samples.

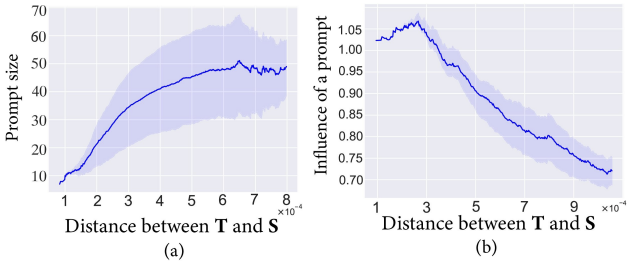

Figure 5. **Effect of text prompts**. (a) Shorter prompts usually lead to more similar student and teacher samples. (b) The student and teacher tend to generate more similar images when the student relies heavily on the text prompt.

Text prompts. Then, we analyse the connection of the student-teacher similarity with the prompt length. Figure 5 demonstrates that shorter prompts typically lead to more similar teacher and student samples. Here, the prompt length equals to the number of CLIP tokens. Intuitively, longer prompts are more likely to describe intricate scenes and object compositions than shorter ones. Note that long prompts can also carry low textual informativeness and describe concepts of low complexity. We hypothesize that this causes high variance in Figure 5a.

Also, we report the prompt influence on the student generation w.r.t. the student-teacher similarity in Figure 5b. We estimate the prompt influence by aggregating student cross-attention maps. More details are in B.1. The student tends to imitate the teacher if it relies heavily on the text prompt.

Trajectory curvature. Previously, it was shown to be beneficial to straighten the PF-ODE trajectory before distillation [24, 25]. We investigate the effect of the trajectory curvature on the similarity between the teacher and student samples and its correlation with the teacher sample complexity. We estimate the trajectory curvatures following [4] and observe that straighter trajectories lead to more similar student and teacher samples (Figure 6a). In addition, we

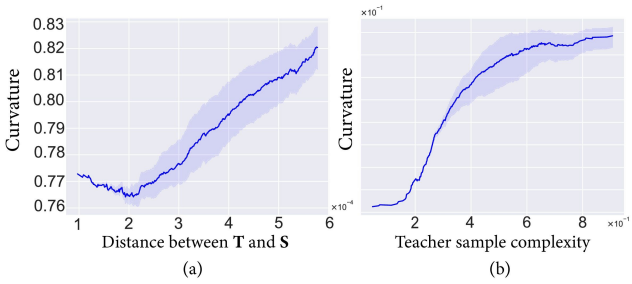

Figure 6. **Effect of teacher trajectory curvature**. (a) The student samples resemble the teacher ones for less curved trajectories. (b) Straighter trajectories usually correspond to plainer teacher images.

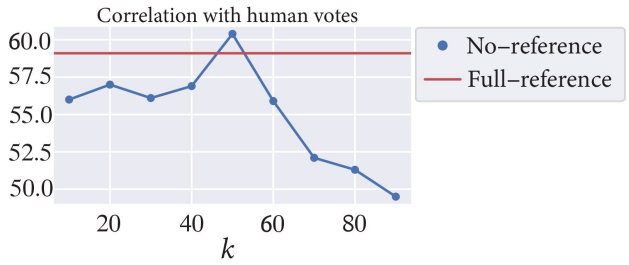

Figure 7. **Full-reference vs no-reference decision-making**. Usually, one can find a k-th percentile of the ImageReward scores providing the correlation with human votes similar to the full-reference comparisons but without observing the teacher samples.

show that the trajectory curvature correlates with the teacher sample complexity (Figure 6b).

To sum up, we conclude that the student largely diverges from the teacher on the samples that are challenging in different respects. Interestingly, the superior student samples often occur in these cases.

3.2. Method

In this section, we propose an adaptive collaborative approach consisting of three steps: 1) Generate a sample with the student model; 2) Decide if the sample needs further improvement; 3) If so, refine or regenerate the sample with the teacher model.

Student generation step produces an initial sample $\mathcal{X}^{\mathbf{S}}$ for a given context and noise. This work considers consistency distillation [50] as a primary distillation framework and uses multistep consistency sampling [50] for generation.

Adaptive step leverages our finding that many student samples may exhibit superior quality. Specifically, we seek an "oracle" that correctly detects superior student samples. For this role, we consider an individual sample quality estimator $\mathbf{E}$. In particular, we use the current state-of-the-art automated estimator, ImageReward (IR) [53] that is learned to imitate human preferences for text-to-image generation.

Then, comparing the scores of the teacher and student samples, one can conclude which one is better. However, in practice, we avoid expensive teacher inference to preserve the efficiency of our approach. Therefore, a decision must be made having access only to the student sample $\mathcal{X}^{\mathbf{S}}$. To address this problem, we introduce a cut-off threshold τ which is a k-th percentile of the IR score tuned on a hold-out subset of student samples. The details on the threshold tuning are described in C. During inference, the IR score is calculated only for $\mathcal{X}^{\mathbf{S}}$. If it exceeds the threshold τ, we accept the sample and avoid further teacher involvement. Interestingly, we observe that it is often possible to reproduce the accuracy of the full-reference estimation by varying τ (see Figure 7). Also, note that IR calculation costs are negligible compared to a single diffusion step, see D.7.

Improvement step engages the teacher to improve the quality of the rejected student samples. We consider two teacher involvement strategies: *regeneration* and *refinement*. The former simply applies the teacher model to produce a new sample from scratch for the same text prompt and noise. The refinement is inspired by the recent work [34]. Specifically, $\mathcal{X}^{\mathbf{S}}$ is first corrupted with a Gaussian noise controlled by the rollback value $\sigma \in [0, 1]$. Higher σ leads to more pronounced changes. We vary σ between 0.3 and 0.75 in our experiments. Then, the teacher starts sampling from the corrupted sample following the original noise schedule and using an arbitrary solver, e.g., DPM-Solver [26]. Note that refinement requires significantly fewer steps to produce the final sample than generation from scratch. Intuitively, the refinement strategy aims to fix the defects of the student sample. At the same time, the regeneration strategy may be useful if $\mathcal{X}^{\mathbf{S}}$ is poorly aligned with the text prompt in general. Our experiments below confirm this intuition.

4. Experiments

We evaluate our approach for text-to-image synthesis, text-guided image editing and controllable generation. The results confirm that the proposed adaptive approach can outperform the baselines for various inference budgets.

4.1. Text-guided image synthesis

In most experiments, we use Stable Diffusion v1.5 (SD1.5) as a teacher model and set the classifier-free guidance scale to 8. To obtain a student model, we implement consistency distillation (CD) for latent diffusion models and distill SD1.5 on the 80M subset of LAION2B [45]. The resulting model demonstrates decent performance for 5 steps of multistep consistency sampling with the guidance scale 8.

Metrics. We first consider FID [13], CLIP score [36] and ImageReward [53] as automatic metrics. ImageReward is selected due to a higher correlation with human preferences compared to FID and CLIP scores. OpenCLIP ViT-bigG [6] is used for CLIP score calculation. For evaluation, we use 5000 text prompts from the COCO2014 validation set [22].

Figure 8. Qualitative comparison of our adaptive refinement approach to the SD1.5 teacher model.

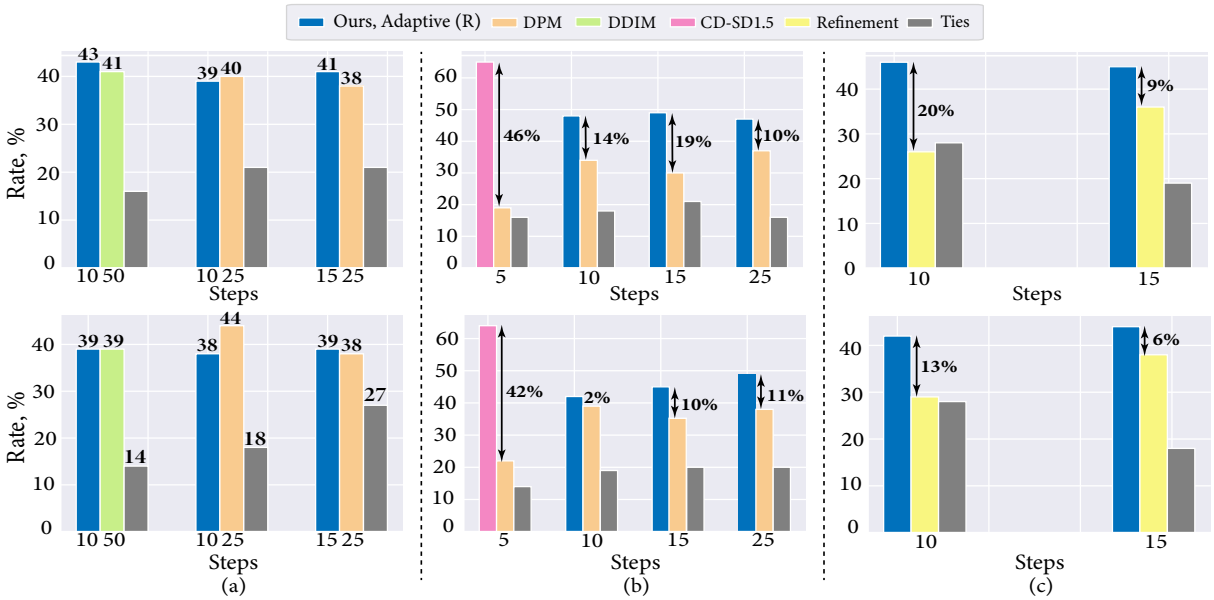

Figure 9. **User preference study (SD1.5)**. (a) Comparison of our approach to the top-performing teacher configurations. (b) Comparison to the teacher model with DPM-Solver for the same average number of steps. (c) Comparison to the refinement strategy without the adaptive step for the same average number of steps. *Top row:* LAION-Aesthetic text prompts. *Bottom row:* COCO2014 text prompts. For our adaptive approach, we use the refinement strategy (R).

Also, we evaluate user preferences using side-by-side comparisons conducted by professional assessors. We select 600 random text prompts from the COCO2014 validation set and 600 from LAION-Aesthetics. More details on the human evaluation pipeline are provided in D.1.

Configuration. For our adaptive approach, we consider both refinement (R) and regeneration (G) strategies using a second order multistep DPM solver [27] and vary the number of sampling steps depending on the average inference budget. As a sample estimator $\mathbf{E}$, we consider ImageReward, except for the CLIP score evaluation. For each inference budget, we tune the hyperparameters σ and τ on the hold-out prompt set. The exact values are provided in D.2.

Baselines. We consider the teacher performance as our main baseline and use DDIM [46] for 50 steps and a second order multistep DPM solver [27] for lower steps. In addition, we compare to the refining strategy on top of all student samples, without the adaptive step. This baseline is inspired by

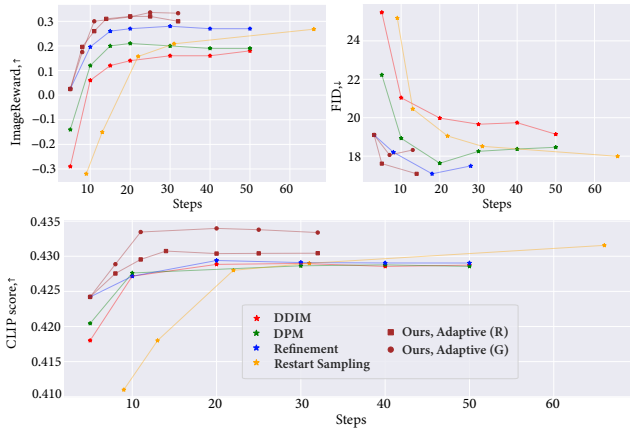

Figure 10. **Automated evaluation (SD1.5)**. Comparison of the FID, CLIP and ImageReward scores for different number of sampling steps on 5K text prompts from COCO2014. The proposed collaborative approach outperforms all the baselines. The adaptive pipeline with the regeneration strategy (G) demonstrates higher textual alignment (CLIP score), while the refinement strategy (R) improves the image fidelity (FID).

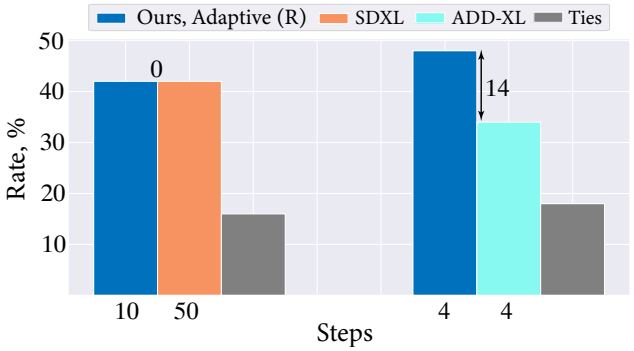

Figure 11. **User preference study (SDXL).** *Left:* Comparison of the adaptive approach with CD-SDXL to the top-performing teacher setup. *Right:* Comparison of the adaptive approach with ADD-XL to ADD-XL for the same average number of steps.

the recent results [34] demonstrating the advantages of the refinement stage itself. Also, we provide the comparison with Restart Sampling [54].

Results. The quantitative and qualitative results are presented in Figures 9, 10 and Figure 12, respectively. According to the automatic metrics, our approach outperforms all the baselines. Specifically, in terms of CLIP scores, the adaptive regeneration strategy demonstrates superior performance compared to the refining-based counterpart. On the other hand, the adaptive refining strategy is preferable in terms of FID scores. We assume that the refinement strategy essentially improves the image fidelity and does not significantly alter the textual alignment due to the relatively small rollback values. In terms of ImageReward, both adaptive strategies perform equally.

In the human evaluation, we consider two nominations:

Method	Steps	DINOv2 ↓	ImageReward ↑
CD-SD1.5	5	0.674 ± .004	0.192 ± .037
SD1.5, DDIM	50	0.665 ± .007	0.183 ± .024
SD1.5, DDIM	25	0.665 ± .002	0.183 ± .022
SD1.5, DPM	25	0.667 ± .005	0.179 ± .020
Refinement	30	0.710 ± .005	0.383 ± .033
Ours	30	0.669 ± .006	0.281 ± .008

Table 1. Comparison of SDEdit using different approaches in terms of reference preservation and editing quality for the strength 0.6.

i) *acceleration*, where our approach aims to reach the performance of SD1.5 using 50 DDIM steps or 25 DPM steps; ii) *quality improvement*, where the adaptive method is compared to the baselines for the same average number of steps. The results for the acceleration nomination are presented in Figure 9a. The proposed method achieves the teacher performance for $5\times$ and up to $2.5\times$ fewer steps compared to DDIM50 and DPM25, respectively. The results for the second nomination (Figure 9b,c) confirm that our approach consistently surpasses alternatives using the same number of steps on average. In particular, the adaptive method improves the generative performance by up to 19% and 20% compared to the teacher and refining strategy without the adaptive step, respectively.

SDXL results. In addition to SD1.5 experiments, we evaluate our pipeline using the recent CD-SDXL [28] and ADD-XL [44] which are both distilled from the SDXL-Base model [34]. Our approach with CD-SDXL stands against the top-performing SDXL setting: 50 steps of the DDIM sampler. For ADD-XL, we provide the comparison for 4 steps where ADD-XL has demonstrated exceptionally strong generative performance in terms of human preference [44]. In both settings, our approach uses the adaptive refinement strategy with the UniPC solver [58]. Note that both the SDXL-Base and SDXL-Refiner [34] models can be used for refinement. In our experiments, we observe that the refiner suits slightly better for fixing minor defects while the teacher allows more pronounced changes. Thus, we use the refiner for low σ values and the base model for the higher ones. More setup details are provided in D.3.

The results are presented in Figure 11. We observe that the adaptive approach using CD-SDXL achieves the quality of the SDXL model, being $5\times$ more efficient on average. Moreover, the proposed scheme improves the performance of ADD-XL by 14% in terms of human preference.

In D.5, we also investigate how our approach affects the distribution diversity. D.4 aims to reveal the potential gains of our approach if the oracle accuracy increases in the future.

4.2. Text-guided image editing

This section applies our approach for text-guided image editing using SDEdit [29]. We add noise to an image, alter the text prompt and denoise it using the student model first. If

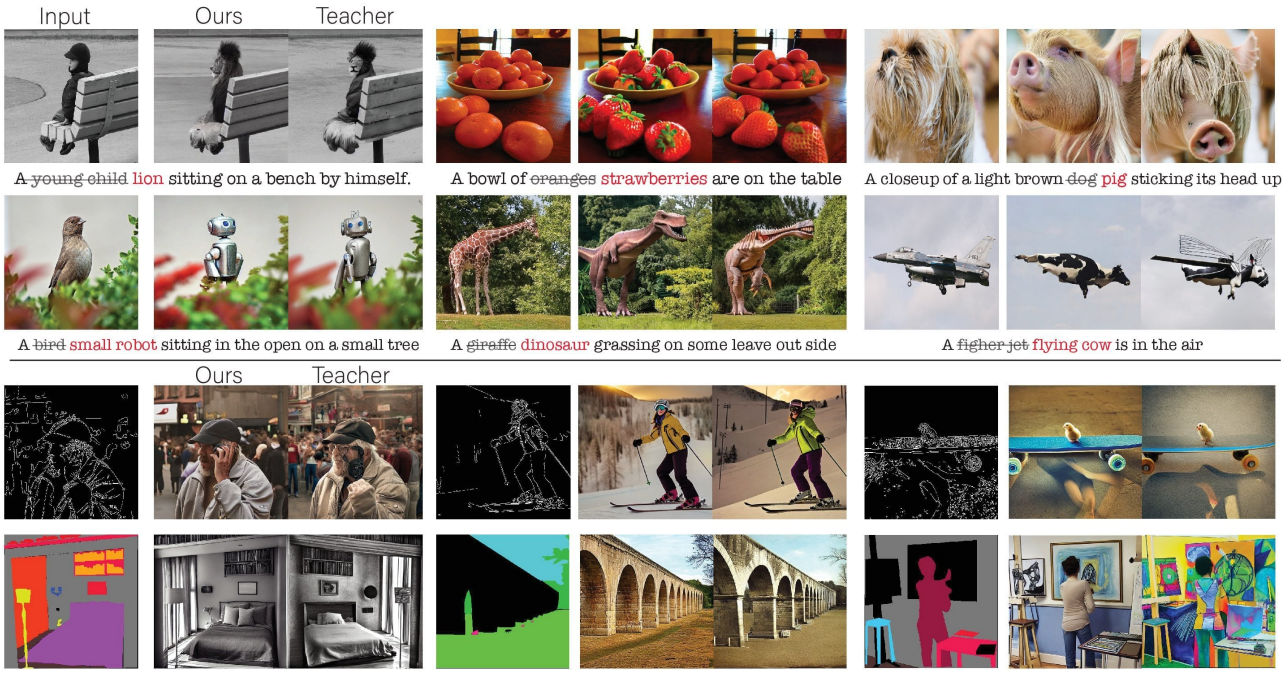

Figure 12. Visual examples produced with our approach and the top-performing teacher (SD1.5) configuration. *Top:* Text-guided image editing with SDEdit [29]. *Bottom:* Controllable image generation with Canny edges and semantic segmentation masks using ControlNet [56].

the edited image does not exceed the threshold τ, the teacher model is used for editing instead. In the editing setting, we observe that the refinement strategy significantly reduces similarity with the reference image due to the additional noising step. Thus, we apply the regeneration strategy only.

In these experiments, the SD1.5 and CD-SD1.5 models are considered. As performance measures, we use ImageReward for editing quality and DINOv2 [33] for reference preservation. For evaluation, 100 text prompts from COCO2014 are manually prepared for the editing task.

Results. Table 1 provides evaluation results for a SDEdit noising strength value 0.6. The proposed method demonstrates a higher ImageReward score compared to the baselines with similar reference preservation scores. In addition, we present the performance for different editing strength values in Figure 13. Our approach demonstrates a better trade-off between reference preservation and editing quality. We provide qualitative results in Figure 12.

4.3. Controllable image generation

Finally, we consider text-to-image generation using Canny edges and semantic segmentation masks as an additional context and use ControlNet [56] for this task. We use ControlNet pretrained on top of SD1.5 and directly plug it into the distilled model (CD-SD1.5). Interestingly, the model pretrained for the teacher model fits the student model surprisingly well without any further adaptation.

For the teacher model, the default ControlNet sampling configuration is used: 20 sampling steps of the UniPC [58] solver. In our adaptive approach, we use the refinement strategy with 10 steps of the same solver. For performance evaluation, we conduct the human preference study for each task on 600 examples and provide more details in D.6.

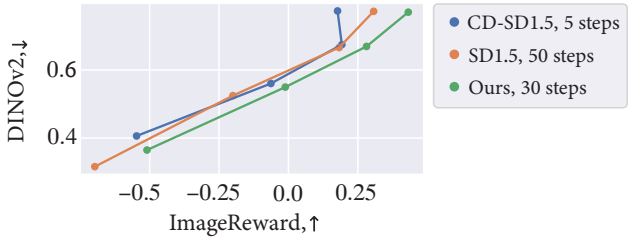

Figure 13. SDEdit performance for different strength values in terms of reference preservation (DINOv2) and editing quality (IR).

Results. According to the human evaluation, our approach outperforms the teacher (20 steps) by 19% (9 steps) and 4% (11 steps) for Canny edges and semantic segmentation masks, respectively. The visual examples are in Figure 12.

5. Conclusion

This work investigates the performance of the distilled text-to-image models and demonstrates that they may consistently outperform the teachers on many samples. We design an adaptive text-to-image generation pipeline that takes advantage of successful student samples and, in combination with the teacher model, outperforms other alternatives for low and high inference budgets.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[2] 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. 2

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[4] Defang Chen, Zhenyu Zhou, Jian-Ping Mei, Chunhua Shen, Chun Chen, and Can Wang. A geometric perspective on diffusion models. *arXiv preprint arXiv:2305.19947*, 2023. 4, 1, 2

[5] Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance, 2023. 2

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022. 5

[7] Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01):1–17, 2023. 4, 1, 2

[8] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023. 2

[9] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[12] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference Generative Modeling*, 2023. 2

[13] 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. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021. 2

[16] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. 1

[17] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 2

[20] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023. 2, 1

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014. 2

[22] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 5

[23] Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, and Yu Wang. Oms-dpm: Optimizing the model schedule for diffusion probabilistic models. *arXiv preprint arXiv:2306.08860*, 2023. 2

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2, 4, 1

[25] 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. 2, 4

[26] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffu-

sion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 2, 5

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2, 6

[28] 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. 2, 7, 1, 3, 4

[29] 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. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3, 7, 8

[30] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2, 1

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[33] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 8

[34] 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, 7, 3, 4

[35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[38] 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. 1, 2

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[40] 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. 2022. 2

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 2

[42] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[43] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. 2023. 2

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023. 2, 7, 3, 4

[45] 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. 3, 5, 1

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 6

[47] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023. 2

[48] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 1

[50] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 2, 3, 5, 1

[51] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. 1

[52] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[53] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 3, 5

[54] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. *arXiv preprint arXiv:2306.14878*, 2023. 7

[55] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 2

[56] 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. 2, 8

[57] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 2

[58] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023. 2, 7, 8, 3

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 4