Adversarial Diffusion Distillation

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach

Stability AI

Abstract. We introduce Adversarial Diffusion Distillation (ADD), a novel training approach that efficiently samples large-scale foundational image diffusion models in just 1–4 steps while maintaining high image quality. We use score distillation to leverage large-scale off-the-shelf image diffusion models as a teacher signal in combination with an adversarial loss to ensure high image fidelity even in the low-step regime of one or two sampling steps. Our analyses show that our model clearly outperforms existing few-step methods (GANs, Latent Consistency Models) in a single step and reaches the performance of state-of-the-art diffusion models (SDXL) in only four steps. ADD is the first method to unlock single-step, real-time image synthesis with foundation models.

Keywords: Diffusion Models · GANs · Distillation

Fig. 1: Generating high-fidelity 512^2 images in a single step. All samples are generated with a single U-Net evaluation trained with adversarial diffusion distillation.

1 Introduction

Diffusion models (DMs) [18, 59, 62] have taken a central role in the field of generative modeling and have recently enabled remarkable advances in high-quality image- [1, 49, 50] and video- [2, 8, 17] synthesis. One of the key strengths of DMs is their scalability and iterative nature, which allows them to handle complex tasks such as image synthesis from free-form text prompts. However, the iterative inference process in DMs requires a significant number of sampling steps, which currently hinders their real-time application. Generative Adversarial Networks (GANs) [11, 23, 24], on the other hand, are characterized by their single-step formulation and inherent speed. But despite attempts to scale to large datasets [22, 55], GANs often fall short of DMs in terms of sample quality. The aim of this work is to combine the superior sample quality of DMs with the inherent speed of GANs.

Our approach is conceptually simple: We propose *Adversarial Diffusion Distillation* (ADD), a general approach that reduces the number of inference steps of a pre-trained diffusion model to 1–4 sampling steps while maintaining high sampling fidelity and potentially further improving the overall performance of the model. To this end, we introduce a combination of two training objectives: (i) an *adversarial loss* and (ii) a distillation loss that corresponds to *score distillation sampling* (SDS) [47]. The adversarial loss forces the model to directly generate samples that lie on the manifold of real images at each forward pass, avoiding blurriness and other artifacts typically observed in other distillation methods [39]. The distillation loss uses another pretrained (and fixed) DM as a teacher to effectively utilize the extensive knowledge of the pretrained DM and preserve the strong compositionality observed in large DMs. During inference, our approach does not use classifier-free guidance [16], further reducing memory requirements. We retain the model's ability to improve results through iterative refinement, which is an advantage over previous one-step GAN-based approaches [54].

Our contributions can be summarized as follows:

- We introduce ADD, a method for turning pretrained diffusion models into high-fidelity, real-time image generators using only 1–4 sampling steps.
- Our method uses a novel combination of adversarial training and score distillation, for which we carefully ablate several design choices.
- ADD significantly outperforms strong baselines such as LCM, LCM-XL [34] and single-step GANs [54], and is able to handle complex image compositions while maintaining high image realism at only a single inference step.
- Using four sampling steps, ADD-XL outperforms its teacher model SDXL-Base at a resolution of 512^2 px.

2 Background

While diffusion models achieve remarkable performance in synthesizing and editing high-resolution images [1, 49, 50] and videos [2, 17], their iterative nature hinders real-time application. Latent diffusion models [50] attempt to solve this problem by

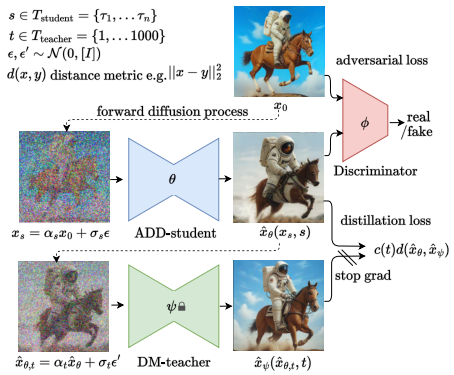

Fig. 2: Adversarial Diffusion Distillation. The ADD-student is trained as a denoiser that receives diffused input images x_s and outputs samples $\hat{x}_\theta(x_s, s)$ and optimizes two objectives: a) adversarial loss: the model aims to fool a discriminator which is trained to distinguish the generated samples $\hat{x}_\theta$ from real images x_0. b) distillation loss: the model is trained to match the denoised targets $\hat{x}_\psi$ of a frozen DM teacher.

representing images in a more computationally feasible latent space [9], but they still rely on the iterative application of large models with billions of parameters. In addition to utilizing faster samplers for diffusion models [6,33,60,69], there is a growing body of research on model distillation such as progressive distillation [52] and guidance distillation [39]. These approaches reduce the number of iterative sampling steps to 4-8, but may significantly lower the original performance. Furthermore, they require an iterative training process. Consistency models [61] address the latter issue by enforcing a consistency regularization on the ODE trajectory and demonstrate strong performance for pixel-based models in the few-shot setting. LCMs [34] focus on distilling latent diffusion models and achieve impressive performance at 4 sampling steps. Recently, LCM-LoRA [36] introduced a low-rank adaptation [19] training for efficiently learning LCM modules, which can be plugged into different checkpoints for SD and SDXL [46,50]. InstaFlow [32] propose to use Rectified Flows [31] to facilitate a better distillation process.

All of these methods share common flaws: samples synthesized in four steps often look blurry and exhibit noticeable artifacts. At fewer sampling steps, this problem is further amplified. GANs [11] can also be trained as standalone single-step models for text-to-image synthesis [22, 54]. Their sampling speed is impressive, yet the performance lags behind diffusion-based models. In part, this can be attributed to the finely balanced GAN-specific architectures necessary for stable training of the adversarial objective. Scaling these models and integrating advances in neural network architectures without disturbing the balance is notoriously challenging. Additionally, current state-of-the-art text-to-image GANs do not have a method like classifier-free guidance available which is crucial for DMs at scale.

Score Distillation Sampling [47] also known as Score Jacobian Chaining [64] is a recently proposed method that has been developed to distill the knowledge of foundational T2I Models into 3D synthesis models. While the majority of SDS-based works [41, 47, 64, 65] use SDS in the context of per-scene optimization for 3D objects, the approach has also been applied to text-to-3D-video-synthesis [58] and in the context of image editing [13].

Recently, the authors of [10] have shown a strong relationship between score-based models and GANs and propose Score GANs, which are trained using score-based diffusion flows from a DM instead of a discriminator. Similarly, Diff-Instruct [38], a method which generalizes SDS, enables to distill a pretrained diffusion model into a generator without discriminator.

Conversely, there are also approaches which aim to improve the diffusion process using adversarial training. For faster sampling, Denoising Diffusion GANs [66] are introduced as a method to enable sampling with few steps. To improve quality, a discriminator loss is added to the score matching objective in Adversarial Score Matching [21] and the consistency objective of CTM [26].

Our method combines adversarial training and score distillation in a hybrid objective to address the issues in current top performing few-step generative models.

3 Method

Our goal is to generate high-fidelity samples in as few sampling steps as possible, while matching the quality of state-of-the-art models [5, 46, 49, 51]. The adversarial objective [11, 56] naturally lends itself to fast generation as it trains a model that outputs samples on the image manifold in a single forward step. However, attempts at scaling GANs to large datasets [54, 55] observed that is critical to not solely rely on the discriminator, but also employ a pretrained classifier or CLIP network for improving text alignment. As remarked in [54], overly utilizing discriminative networks introduces artifacts and image quality suffers. Instead, we utilize the gradient of a pretrained diffusion model via a score distillation objective to improve text alignment and sample quality. Furthermore, instead of training from scratch, we initialize our model with pretrained diffusion model weights; pretraining the generator network is known to significantly improve training with an adversarial loss [12]. Lastly, instead of utilizing a decoder-only architecture used for GAN training [23, 24], we adapt a standard diffusion model framework. This setup naturally enables iterative refinement.

3.1 Training Procedure

Our training procedure is outlined in Fig. 2 and involves three networks: The ADD-student is initialized from a pretrained UNet-DM with weights θ, a discriminator with trainable weights ϕ, and a DM teacher with frozen weights ψ. During training, the ADD-student generates samples $\hat{x}_\theta(x_s, s)$ from noisy data x_s. The noised data points are produced from a dataset of real images x_0 via a forward diffusion

process $x_s = \alpha_s x_0 + \sigma_s \epsilon$. In our experiments, we use the same coefficients α_s and σ_s as the student DM and sample s uniformly from a set $T_{\text{student}} = \{\tau_1, ..., \tau_n\}$ of N chosen student timesteps. In practice, we choose $N = 4$. Importantly, we set $\tau_n = 1000$ and enforce zero-terminal SNR [29] during training, as the model needs to start from pure noise during inference.

For the adversarial objective, the generated samples $\hat{x}_\theta$ and real images x_0 are passed to the discriminator which aims to distinguish between them. The design of the discriminator and the adversarial loss are described in detail in Sec. 3.2. To distill knowledge from the DM teacher, we diffuse student samples $\hat{x}_\theta$ with the teacher's forward process to $\hat{x}_{\theta,t}$, and use the teacher's denoising prediction $\hat{x}_\psi(\hat{x}_{\theta,t}, t)$ as a reconstruction target for the distillation loss $\mathcal{L}_{\text{distill}}$, see Section 3.3. Thus, the overall objective is

$$\mathcal{L} = \mathcal{L}^{\text{G}}_{\text{adv}}(\hat{x}_\theta(x_s, s), \phi) + \lambda \mathcal{L}_{\text{distill}}(\hat{x}_\theta(x_s, s), \psi) \tag{1}$$

While we formulate our method in pixel space, it is straightforward to adapt it to LDMs operating in latent space. When using LDMs with a shared latent space for teacher and student, the distillation loss can be computed in pixel or latent space. We compute the distillation loss in pixel space as this yields more stable gradients when distilling latent diffusion model [68].

3.2 Adversarial Loss

For the discriminator, we follow the proposed design and training procedure in [54] which we briefly summarize; for details, we refer the reader to the original work. We use a frozen pretrained feature network F and a set of trainable lightweight discriminator heads $\mathcal{D}_{\phi,k}$. For the feature network F, Sauer et al. [54] find vision transformers (ViTs) [7] to work well, and we ablate different choice for the ViTs objective and model size in Section 4. The trainable discriminator heads are applied on features F_k at different layers of the feature network.

To improve performance, the discriminator can be conditioned on additional information via projection [42]. Commonly, a text embedding c_{text} is used in the text-to-image setting. But, in contrast to standard GAN training, our training configuration also allows to condition on a given image. For $\tau < 1000$, the ADD-student receives some signal from the input image x_0. Therefore, for a given generated sample $\hat{x}_\theta(x_s, s)$, we can condition the discriminator on information from x_0. This encourages the ADD-student to utilize the input effectively. In practice, we use an additional feature network to extract an image embedding c_{img}.

Following [53, 54], we use the hinge loss [28] as the adversarial objective function. Thus the ADD-student's adversarial objective $\mathcal{L}_{\text{adv}}(\hat{x}_\theta(x_s, s), \phi)$ amounts to

$$\begin{aligned} &\mathcal{L}^{\text{G}}_{\text{adv}}(\hat{x}_\theta(x_s, s), \phi) \\ &\quad = -\mathbb{E}_{s,\epsilon,x_0}\Big[\sum_k \mathcal{D}_{\phi,k}(F_k(\hat{x}_\theta(x_s, s)))\Big], \end{aligned} \tag{2}$$

whereas the discriminator is trained to minimize

$$\begin{aligned}&\mathcal{L}_{\text{adv}}^{\text{D}}(\hat{x}_\theta(x_s, s), \phi)\\ &= \mathbb{E}_{x_0}\Big[\sum_k \max(0, 1 - \mathcal{D}_{\phi,k}(F_k(x_0))) + \gamma R1(\phi)\Big]\\ &\quad + \mathbb{E}_{\hat{x}_\theta}\Big[\sum_k \max(0, 1 + \mathcal{D}_{\phi,k}(F_k(\hat{x}_\theta)))\Big],\end{aligned} \tag{3}$$

where $R1$ denotes the R1 gradient penalty [40]. Rather than computing the gradient penalty with respect to the pixel values, we compute it on the input of each discriminator head $\mathcal{D}_{\phi,k}$. We find that the $R1$ penalty is particularly beneficial when training at output resolutions larger than 128^2 px.

3.3 Score Distillation Loss

The distillation loss in Eq. (1) is formulated as

$$\begin{aligned}&\mathcal{L}_{\text{distill}}(\hat{x}_\theta(x_s, s), \psi)\\ &\qquad = \mathbb{E}_{t,\epsilon'}\left[c(t) d(\hat{x}_\theta, \hat{x}_\psi(\text{sg}(\hat{x}_{\theta,t}); t))\right],\end{aligned} \tag{4}$$

where sg denotes the stop-gradient operation. Intuitively, the loss uses a distance metric d to measure the mismatch between generated samples x_θ by the ADD-student and the DM-teacher's outputs $\hat{x}_\psi(\hat{x}_{\theta,t}, t) = (\hat{x}_{\theta,t} - \sigma_t \hat{\epsilon}_\psi(\hat{x}_{\theta,t}, t))/\alpha_t$ averaged over timesteps t and noise ϵ'. Notably, the teacher is not directly applied on generations $\hat{x}_\theta$ of the ADD-student but instead on diffused outputs $\hat{x}_{\theta,t} = \alpha_t \hat{x}_\theta + \sigma_t \epsilon'$, as non-diffused inputs would be out-of-distribution for the teacher model [64].

In the following, we define the distance function $d(x, y) \coloneqq ||x - y||_2^2$. Regarding the weighting function $c(t)$, we consider two options: exponential weighting, where $c(t) = \alpha_t$ (higher noise levels contribute less), and score distillation sampling (SDS) weighting [47]. In the supplementary material, we demonstrate that with $d(x, y) = ||x - y||_2^2$ and a specific choice for $c(t)$, our distillation loss becomes equivalent to the SDS objective $\mathcal{L}_{\text{SDS}}$, as proposed in [47]. The advantage of our formulation is its ability to enable direct visualization of the reconstruction targets and that it naturally facilitates the execution of several consecutive denoising steps. Lastly, we also evaluate noise-free score distillation (NFSD) objective, a recently proposed variant of SDS [25].

4 Experiments

For our experiments, we train two models of different capacities, ADD-M (860M parameters) and ADD-XL (3.1B parameters). For ablating ADD-M, we use a Stable Diffusion (SD) 2.1 backbone [50], and for fair comparisons with other baselines, we use SD1.5. ADD-XL utilizes a SDXL [46] backbone. All experiments are conducted at a standardized resolution of 512x512 pixels; outputs from models generating higher resolutions are down-sampled to this size.

Fig. 3: Qualitative comparison to state-of-the-art fast samplers. Single step samples from our ADD-XL (top) and LCM-XL [36], our custom StyleGAN-T [54] baseline, InstaFlow [32] and MUSE. For MUSE, we use the *OpenMUSE* implementation and default inference settings with 16 sampling steps. For LCM-XL, we sample with 1, 2 and 4 steps. Our model outperforms all other few-step samplers in a single step.

Arch	Objective	FID ↓	CS ↑
ViT-S	DINOv1	21.5	0.312
ViT-S	DINOv2	**20.6**	**0.319**
ViT-L	DINOv2	24.0	0.302
ViT-L	CLIP	23.3	0.308

(a) Discriminator feature networks. Small, modern DINO networks perform best.

c_{text}	c_{img}	FID ↓	CS ↑
✗	✗	21.2	0.302
✓	✗	21.2	0.307
✗	✓	21.1	0.316
✓	✓	**20.6**	**0.319**

(b) Discriminator conditioning. Combining image and text conditioning is most effective.

Initialization	FID ↓	CS ↑
Random	293.6	0.065
Pretrained	**20.6**	**0.319**

(c) Student pretraining. A randomly initialized student network collapses.

Loss	FID ↓	CS ↑
$\mathcal{L}_{\text{adv}}$	20.8	0.315
$\mathcal{L}_{\text{dist}}$	315.6	0.076
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{dist,exp}}$	**20.6**	0.319
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{dist,sds}}$	22.3	0.325
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{dist,nfsd}}$	21.8	**0.327**

(d) Loss terms. Both losses are needed and exponential weighting of $\mathcal{L}_{\text{distill}}$ is beneficial.

Student	Teacher	FID ↓	CS ↑
SD2.1	SD2.1	**20.6**	0.319
SD2.1	SDXL	21.3	0.321
SDXL	SD2.1	29.3	0.314
SDXL	SDXL	28.41	**0.325**

(e) Teacher type. The student adopts the teacher's traits (SDXL has higher FID & CS).

Steps	FID ↓	CS ↑
1	**20.6**	0.319
2	20.8	**0.321**
4	20.3	0.317

(f) Teacher steps. A single teacher step is sufficient.

Table 1: ADD ablation study. We report COCO zero-shot FID_{5k} (FID) and CLIP score (CS). The results are calculated for a single student step. The default training settings are: DINOv2 ViT-S as the feature network, text and image conditioning for the discriminator, pretrained student weights, a small teacher and student model, and a single teacher step. The training length is 4000 iterations with a batch size of 128. Default settings are marked in gray.

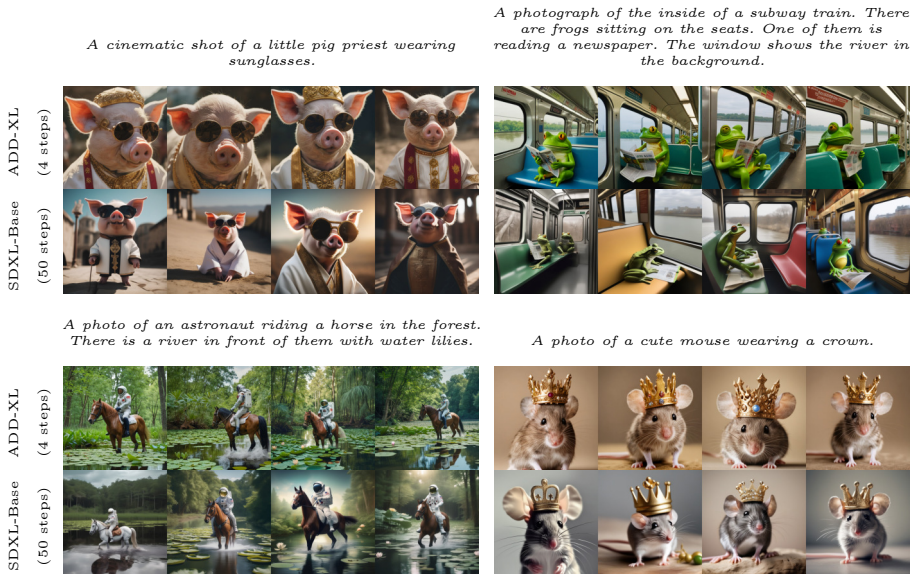

Fig. 4: Qualitative comparison to the teacher model. ADD-XL can outperform its teacher model SDXL in the multi-step setting. The adversarial loss boosts realism, particularly enhancing textures (fur, fabric, skin) while reducing oversmoothing, commonly observed in diffusion model samples. ADD-XL's overall sample diversity tends to be lower.

We employ a distillation weighting factor of $\lambda = 2.5$ across all experiments. Additionally, the R1 penalty strength γ is set to 10^{-5}. For discriminator conditioning, we use a pretrained CLIP-ViT-g-14 text encoder [48] to compute text embeddings c_{text} and the CLS embedding of a DINOv2 ViT-L encoder [43] for image embeddings c_{img}. For the baselines, we use the best publicly available models: Latent diffusion models [46,50] (SD1.5[1], SDXL[2]) cascaded pixel diffusion models [51] (IF-XL[3]), distilled diffusion models [35,37] (LCM-1.5, LCM-1.5-XL[4]), and OpenMUSE [5] [44], a reimplementation of MUSE [4], a transformer model specifically developed for fast inference. Note that we compare to the SDXL-Base-1.0 model without its additional refiner model; this is to ensure a fair comparison. As there are no public state-of-the-art GAN models, we retrain StyleGAN-T [54] with our improved discriminator. This baseline (StyleGAN-T++) significantly outperforms the previous best GANs in FID and CS, see supplementary. We quantify sample quality via FID [15] and text alignment via CLIP score [14]. For CLIP score, we use ViT-g-14 model trained on LAION-2B [57]. Both metrics are evaluated on 5k samples from COCO2017 [30].

4.1 Ablation Study

Our training setup opens up a number of design spaces regarding the adversarial loss, distillation loss, initialization, and loss interplay. We conduct an ablation study on several choices in Table 1; key insights are highlighted below each table. We will discuss each experiment in the following.

Discriminator feature networks. (Table 1a). Recent insights by Stein et al. [63] suggest that ViTs trained with the CLIP [48] or DINO [3,43] objectives are particularly well-suited for evaluating the performance of generative models. Similarly, these models also seem effective as discriminator feature networks, with DINOv2 emerging as the best choice. Interestingly, ViT-S outperforms ViT-L, which is a counterintuitive result. Similar results have been obtained in [54], suggesting that discriminators don't strictly adhere to scaling laws.

Discriminator conditioning. (Table 1b). Similar to prior studies, we observe that text conditioning of the discriminator enhances results. Notably, image conditioning outperforms text conditioning, and the combination of both c_{text} and c_{img} yields the best results.

Student pretraining. (Table 1c). Our experiments demonstrate the importance of pretraining the ADD-student. Being able to use pretrained generators is a significant advantage over pure GAN approaches. A problem of GANs is the lack of scalability; both Sauer et al. [54] and Kang et al. [22] observe a saturation of performance after a certain network capacity is reached. This observation contrasts the generally smooth scaling laws of DMs [45]. However, ADD can

[1] https://github.com/CompVis/stable-diffusion
[2] https://github.com/Stability-AI/generative-models
[3] https://github.com/deep-floyd/IF
[4] https://huggingface.co/latent-consistency/lcm-lora-sdxl
[5] https://huggingface.co/openMUSE

Method	#Steps	Time (s)	FID ↓	CLIP ↑
DPM Solver [33]	25	0.88	20.1	0.318
	8	0.34	31.7	0.320
Progressive Distillation [39]	1	0.09	37.2	0.275
	2	0.13	26.0	0.297
	4	0.21	26.4	0.300
CFG-Aware Distillation [27]	8	0.34	24.2	0.300
InstaFlow-0.9B [32]	1	0.09	23.4	0.304
InstaFlow-1.7B [32]	1	0.12	22.4	0.309
UFOGen [67]	1	0.09	22.5	0.311
ADD-M	1	0.09	**19.7**	**0.326**

Table 2: Distillation Comparison We compare ADD to other distillation methods via COCO zero-shot FID_{5k} (FID) and CLIP score (CS). All models are based on SD1.5.

effectively leverage larger pretrained DMs (see Table 1c) and benefit from stable DM pretraining. Generally, adversarial training demands carefully designed, specialized architectures to work well (e.g. equalized learning rates). A pretrained student appears to circumvent this need and allows leveraging the vast design space and the stable pretraining of diffusion models.

Loss terms. (Table 1d). We find that both losses are beneficial. The adversarial loss functions decently well on its own which we attribute to our discriminator design. The distillation loss on its own is not effective and cannot improve over the initial performance of the non-distilled student in the single step setting. However, when combined with the adversarial loss, there is a noticeable improvement in results, particularly CLIP score, ie. the text alignment, is boosted noticeably. Different distillation loss weighting schedules lead to different behaviours, the exponential schedule tends to yield more diverse samples, as indicated by lower FID, SDS and NFSD schedules improve quality and text alignment. While we use the exponential schedule as the default setting in all other ablations, we opt for the NFSD weighting for training our final model. Choosing an optimal weighting function presents an opportunity for improvement. Alternatively, scheduling the distillation weights over training, as explored in the 3D generative modeling literature [20] could be considered.

Given the strong performance of pure adversarial training, we hypothesize that by continued research into adversarial training and further improved discriminator architectures, it is possible to achieve similar performance to our current setup with an adversarial loss only.

Teacher type. (Table 1e). Interestingly, a bigger student and teacher does not necessarily result in better FID and CS. Rather, the student adopts the teachers characteristics. SDXL obtains generally higher FID, possibly because of its less diverse output, yet it exhibits higher image quality and text alignment [46].

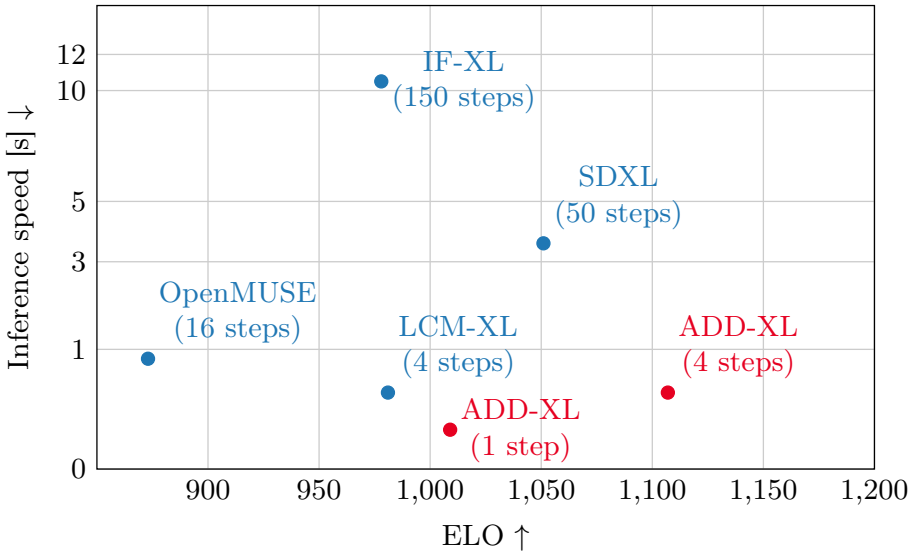

Fig. 5: Performance vs. speed. We visualize the results reported in Fig. 7 in combination with the inference speeds of the respective models. The speeds are calculated for generating a single sample at resolution 512x512 on an A100 in mixed precision.

Teacher steps. (Table 1f). While our distillation loss formulation allows taking several consecutive steps with the teacher by construction, we find that several steps do not conclusively result in better performance.

4.2 Quantitative Comparison to State-of-the-Art

For our main comparison with other approaches, we refrain from using automated metrics, as user preference studies are more reliable [46]. In the study, we aim to assess both prompt adherence and the overall image. As a performance measure, we compute win percentages for pairwise comparisons and ELO scores when comparing several approaches. For the reported ELO scores we calculate the mean scores between both prompt following and image quality. Details on the ELO score computation and the study parameters are listed in the supplementary material.

Fig. 6 and Fig. 7 present the study results. The most important results are: First, ADD-XL outperforms LCM-XL (4 steps) with a single step. Second, ADD-XL can beat SDXL (50 steps) with four steps in the majority of comparisons. This makes ADD-XL the state-of-the-art in both the single and the multiple steps setting. Fig. 5 visualizes ELO scores relative to inference speed. Lastly, Table 2 compares different few-step sampling and distillation methods using the same base model. ADD outperforms all other approaches including the standard DPM solver with eight steps.

Fig. 6: User preference study (*single step*). We compare the performance of ADD-XL (1-step) against established baselines. ADD-XL model outperforms all models, except SDXL in human preference for both image quality and prompt alignment. Using more sampling steps further improves our model (bottom row).

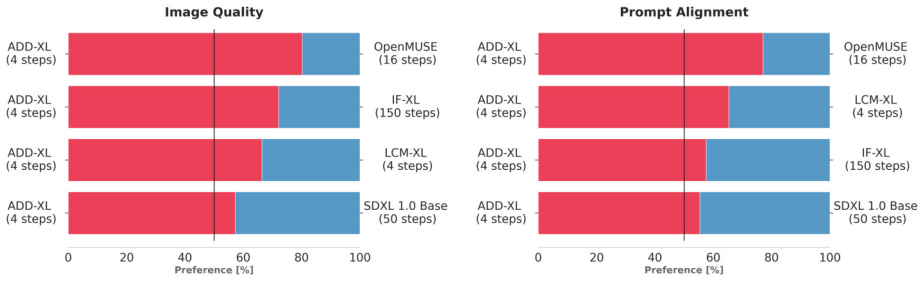

Fig. 7: User preference study (*multiple steps*). We compare the performance of ADD-XL (4-step) against established baselines. Our ADD-XL model outperforms all models, including its teacher SDXL 1.0 (base, no refiner) [46], in human preference for both image quality and prompt alignment.

4.3 Qualitative Results

To complement our quantitative studies above, we present qualitative results in this section. To paint a more complete picture, we provide additional samples and qualitative comparisons in the supplementary material. Fig. 3 compares ADD-XL (1 step) against the best current baselines in the few-steps regime. Fig. 8 illustrates the iterative sampling process of ADD-XL. These results showcase our model's ability to improve upon an initial sample. Such iterative improvement represents another significant benefit over pure GAN approaches like StyleGAN-T++. Lastly, Fig. 4 compares ADD-XL directly with its teacher model SDXL-Base. As indicated by the user studies in Section 4.2, ADD-XL outperforms its teacher in both quality and prompt alignment. The enhanced realism comes at the cost of slightly decreased sample diversity.

5 Discussion

This work introduces *Adversarial Diffusion Distillation*, a general method for distilling a pretrained diffusion model into a fast, few-step image generation model.

Fig. 8: Qualitative effect of sampling steps. We show qualitative examples when sampling ADD-XL with 1, 2, and 4 steps. Single-step samples are often already of high quality, but increasing the number of steps can further improve the consistency (e.g. second prompt, first column) and attention to detail (e.g. second prompt, second column). The seeds are constant within columns and we see that the general layout is preserved across sampling steps, allowing for fast exploration of outputs while retaining the possibility to refine.

We combine an adversarial and a score distillation objective to distill the public Stable Diffusion [50] and SDXL [46] models, leveraging both real data through the discriminator and structural understanding through the diffusion teacher. Our approach performs particularly well in the ultra-fast sampling regime of one or two steps, and our analyses demonstrate that it outperforms all concurrent methods in this regime. Furthermore, we retain the ability to refine samples using multiple steps. In fact, using four sampling steps, our model outperforms widely used multi-step generators such as SDXL, IF, and OpenMUSE.

Our model enables the generation of high quality images in a single-step, opening up new possibilities for real-time generation with foundation models.

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.Y.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. ArXiv **abs/2211.01324** (2022), https://api.semanticscholar.org/CorpusID:253254800
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22563–22575 (2023), https://api.semanticscholar.org/CorpusID:258187553
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. Proc. ICML (2023)
5. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
6. Dockhorn, T., Vahdat, A., Kreis, K.: Genie: Higher-order denoising diffusion solvers. Advances in Neural Information Processing Systems **35**, 30150–30166 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models (2023)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12868–12878 (2020), https://api.semanticscholar.org/CorpusID:229297973
10. Franceschi, J.Y., Gartrell, M., Santos, L.D., Issenhuth, T., de Bézenac, E., Chen, M., Rakotomamonjy, A.: Unifying gans and score-based diffusion as generative particle models. arXiv preprint arXiv:2305.16150 (2023)
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**, 139 – 144 (2014), https://api.semanticscholar.org/CorpusID:1033682
12. Grigoryev, T., Voynov, A., Babenko, A.: When, why, and which pretrained gans are useful? ICLR (2022)
13. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)
14. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proc. EMNLP (2021)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. NeurIPS (2017)
16. Ho, J.: Classifier-free diffusion guidance. ArXiv **abs/2207.12598** (2022), https://api.semanticscholar.org/CorpusID:249145348
17. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A.A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition

video generation with diffusion models. ArXiv **abs/2210.02303** (2022), https://api.semanticscholar.org/CorpusID:252715883
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. ArXiv **abs/2006.11239** (2020), https://api.semanticscholar.org/CorpusID:219955663
19. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021), https://api.semanticscholar.org/CorpusID:235458009
20. Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J., Zhang, L.: Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422 (2023)
21. Jolicoeur-Martineau, A., Piché-Taillefer, R., Combes, R.T.d., Mitliagkas, I.: Adversarial score matching and improved sampling for image generation. arXiv preprint arXiv:2009.05475 (2020)
22. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4396–4405 (2018), https://api.semanticscholar.org/CorpusID:54482423
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8107–8116 (2019), https://api.semanticscholar.org/CorpusID:209202273
25. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023)
26. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. arXiv preprint arXiv:2310.02279 (2023)
27. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. arXiv preprint arXiv:2306.00980 (2023)
28. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017)
29. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed (2023)
30. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
31. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2022)
32. Liu, X., Zhang, X., Ma, J., Peng, J., Liu, Q.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. arXiv preprint arXiv:2309.06380 (2023)
33. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems **35**, 5775–5787 (2022)
34. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. ArXiv **abs/2310.04378** (2023), https://api.semanticscholar.org/CorpusID:263831037

35. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
36. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. ArXiv **abs/2311.05556** (2023), https://api.semanticscholar.org/CorpusID:265067414
37. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556 (2023)
38. Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., Zhang, Z.: Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. arXiv preprint arXiv:2305.18455 (2023)
39. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023)
40. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
41. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12663–12673 (2022), https://api.semanticscholar.org/CorpusID:253510536
42. Miyato, T., Koyama, M.: cgans with projection discriminator. arXiv preprint arXiv:1802.05637 (2018)
43. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
44. Patil, S., Berman, W., von Platen, P.: Amused: An open muse model. https://github.com/huggingface/diffusers (2023)
45. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
46. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
47. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
49. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. ArXiv **abs/2204.06125** (2022), https://api.semanticscholar.org/CorpusID:248097655
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10674–10685 (2021), https://api.semanticscholar.org/CorpusID:245335280
51. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-

to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
52. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. CoRR **abs/2202.00512** (2022), https://arxiv.org/abs/2202.00512
53. Sauer, A., Chitta, K., Müller, J., Geiger, A.: Projected gans converge faster. Advances in Neural Information Processing Systems **34**, 17480–17492 (2021)
54. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. Proc. ICML (2023)
55. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. ACM SIGGRAPH 2022 Conference Proceedings (2022), https://api.semanticscholar.org/CorpusID:246441861
56. Schmidhuber, J.: Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991) (2020)
57. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
58. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al.: Text-to-4d dynamic scene generation. arXiv preprint arXiv:2301.11280 (2023)
59. Sohl-Dickstein, J.N., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. ArXiv **abs/1503.03585** (2015), https://api.semanticscholar.org/CorpusID:14888175
60. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
61. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:257280191
62. Song, Y., Sohl-Dickstein, J.N., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. ArXiv **abs/2011.13456** (2020), https://api.semanticscholar.org/CorpusID:227209335
63. Stein, G., Cresswell, J.C., Hosseinzadeh, R., Sui, Y., Ross, B.L., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, J.E.T., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. arXiv preprint arXiv:2306.04675 (2023)
64. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
65. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. ArXiv **abs/2305.16213** (2023), https://api.semanticscholar.org/CorpusID:258887357
66. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021)
67. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023), https://api.semanticscholar.org/CorpusID:265221033

68. Yao, C.H., Raj, A., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Artic3d: Learning robust articulated 3d shapes from noisy web image collections. arXiv preprint arXiv:2306.04619 (2023)
69. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. arXiv preprint arXiv:2204.13902 (2022)