Adversarial Diffusion Distillation

Axel Sauer Dominik Lorenz Andreas Blattmann Robin Rombach

Stability AI

Code: https://github.com/Stability-AI/generative-models *Model weights*: https://huggingface.co/stabilityai/

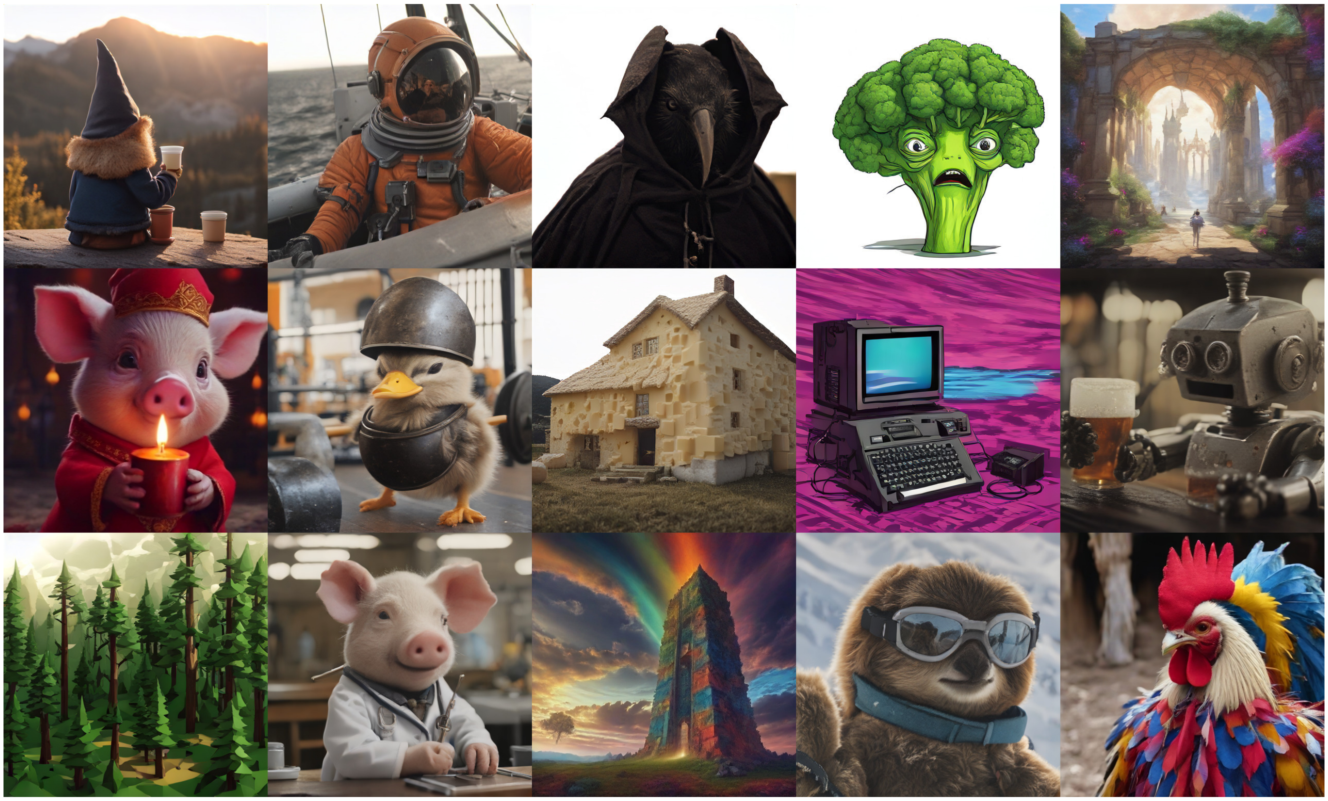


Figure 1. **Generating high-fidelity** 512^2 **images in a single step.** All samples are generated with a single U-Net evaluation trained with adversarial diffusion distillation (ADD).

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.

1. Introduction

Diffusion models (DMs) [20, 63, 65] have taken a central role in the field of generative modeling and have recently enabled remarkable advances in high-quality image- [3, 53, 54] and video- [4, 12, 21] 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) [14, 26, 27], on the other hand, are characterized by their single-step formulation and inherent speed. But despite attempts to scale to large datasets[25, 58], 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.

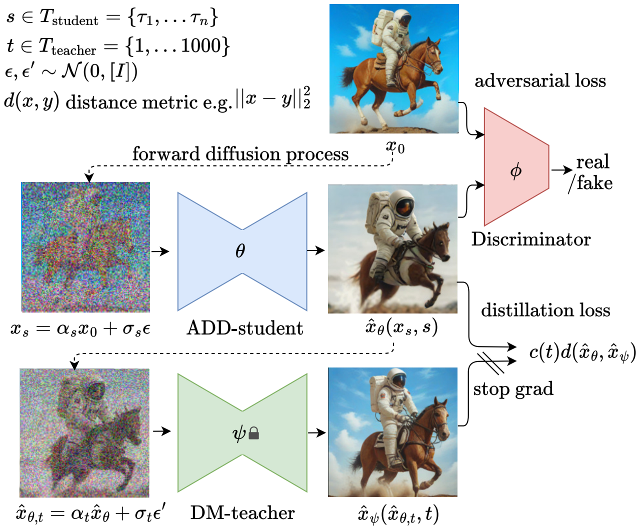
Our approach is conceptually simple: We propose *Adversarial Diffusion Distillation* (ADD), a general approach that reduces the number of inference steps of a pre-trained diffusion model to 1–4 sampling steps while maintaining high sampling fidelity and potentially further improving the overall performance 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) [51]. 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 [43]. 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 [19], 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 [59].

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 [38] and single-step GANs [59], 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 [3, 53, 54] and videos [4, 21], their iterative nature hinders real-time application. Latent diffusion models [54] attempt to solve this problem by representing images in a more computationally feasible latent space [11], but they still rely on the iterative application of large models with billions of parameters. In addition to utilizing faster samplers for diffusion models [8, 37, 64, 74], there is a growing body of research on model distillation such as progressive distillation [56] and guidance distillation [43]. 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 [66] 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 [38] focus on distilling latent diffusion models and achieve impressive performance at 4 sampling steps. Recently, LCM-LoRA [40] introduced a low-rank adaptation [22] training for efficiently learning LCM modules, which can be plugged into different checkpoints for SD and SDXL [50, 54]. InstaFlow [36] propose to use Rectified Flows [35] to facilitate a better distillation process.



$s \in T_{\text{student}} = \{\tau_1, \dots \tau_n\}$
$t \in T_{\text{teacher}} = \{1, \dots 1000\}$
$\epsilon, \epsilon' \sim \mathcal{N}(0, [I])$
$d(x, y)$ distance metric e.g. $||x - y||_2^2$
adversarial loss
forward diffusion process
x_0
ϕ
real /fake
Discriminator
θ
$x_s = \alpha_s x_0 + \sigma_s \epsilon$
ADD-student
$\hat{x}_\theta(x_s, s)$
distillation loss
$c(t) d(\hat{x}_\theta, \hat{x}_\psi)$
stop grad
ψ
$\hat{x}_{\theta,t} = \alpha_t \hat{x}_\theta + \sigma_t \epsilon'$
DM-teacher
$\hat{x}_\psi(\hat{x}_{\theta,t}, t)$

Figure 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.

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 [14] can also be trained as standalone single-step models for text-to-image synthesis [25, 59]. 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 [51] also known as Score Jacobian Chaining [68] 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 [45, 51, 68, 69] use SDS in the context of

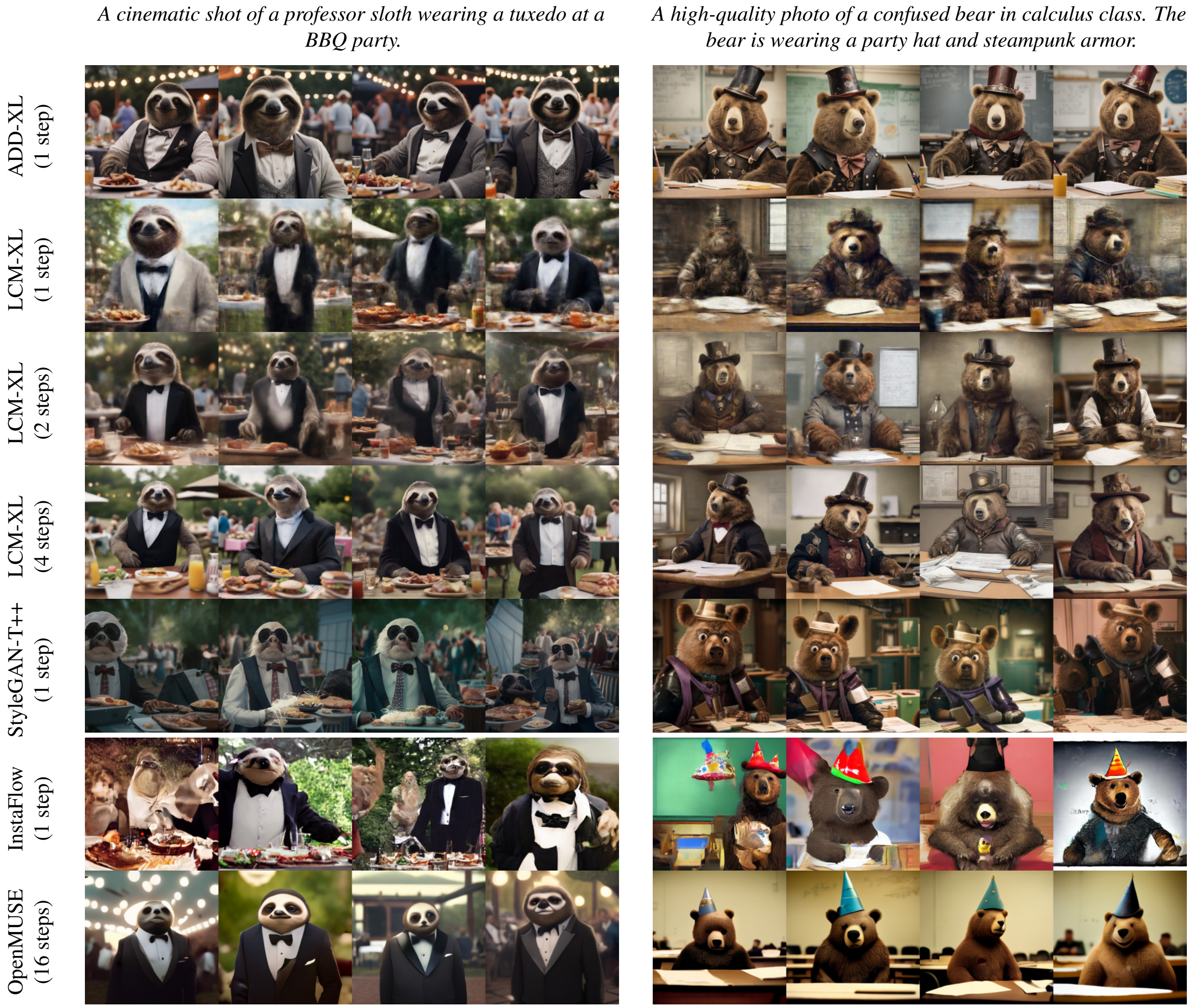

A cinematic shot of a professor sloth wearing a tuxedo at a BBQ party.
A high-quality photo of a confused bear in calculus class. The bear is wearing a party hat and steampunk armor.
ADD-XL (1 step)
LCM-XL (1 step)
LCM-XL (2 steps)
LCM-XL (4 steps)
StyleGAN-T++ (1 step)
InstaFlow (1 step)
OpenMUSE (16 steps)

Figure 3. **Qualitative comparison to state-of-the-art fast samplers.** Single step samples from our ADD-XL (top) and LCM-XL [40], our custom StyleGAN-T [59] baseline, InstaFlow [36] 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.

per-scene optimization for 3D objects, the approach has also been applied to text-to-3D-video-synthesis [62] and in the context of image editing [16].

Recently, the authors of [13] 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 [42], 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 [70] 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 [24] and the consistency objective of CTM [29].

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-

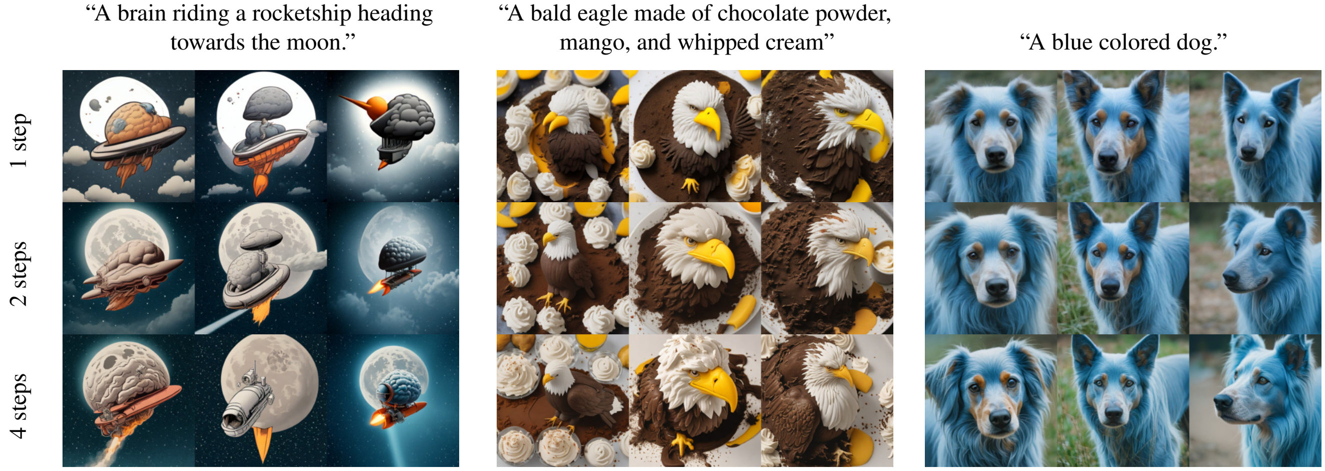

“A brain riding a rocketship heading towards the moon.”
“A bald eagle made of chocolate powder, mango, and whipped cream”
“A blue colored dog.”
1 step
2 steps
4 steps

Figure 4. **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.

of-the-art models [7, 50, 53, 55]. The adversarial objective [14, 60] 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 [58, 59] 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 [59], 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 [15]. Lastly, instead of utilizing a decoder-only architecture used for GAN training [26, 27], 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 [33] 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

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

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 [72].

3.2. Adversarial Loss

For the discriminator, we follow the proposed design and training procedure in [59] 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. [59] find vision transformers (ViTs) [9] 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 [46]. Com-

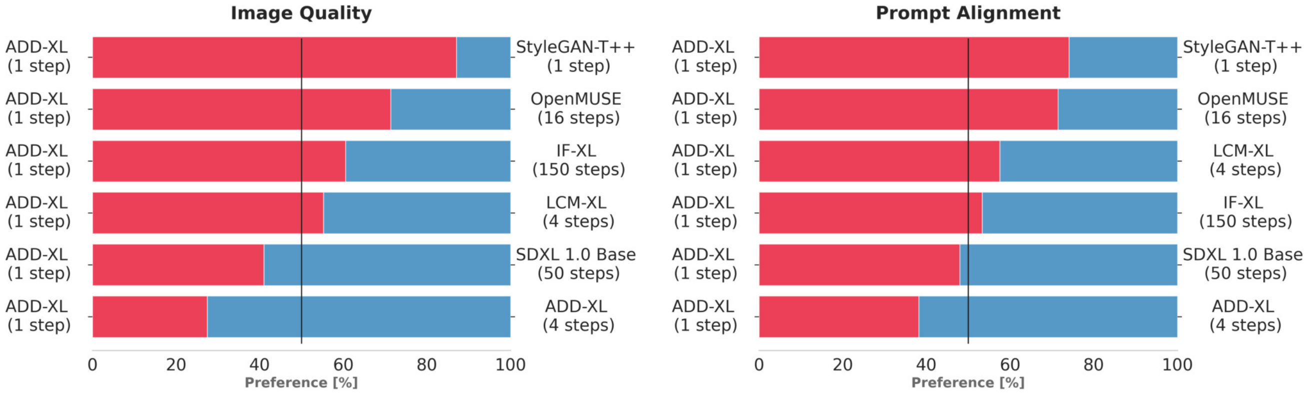


Figure 5. **User preference study (*single step*).** We compare the performance of ADD-XL (1-step) against established baselines. ADD-XL model outperforms all models, except SDXL in human preference for both image quality and prompt alignment. Using more sampling steps further improves our model (bottom row).

monly, 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 [57, 59], we use the hinge loss [32] 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

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

where $R1$ denotes the R1 gradient penalty [44]. 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

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

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 [68].

In the following, we define the distance function $d(x, y) := ||x - y||_2^2$. 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 [51]. 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 [51]. 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 [28].

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 [54], and for fair comparisons with other baselines, we use SD1.5. ADD-XL utilizes a SDXL [50] backbone. All experiments are conducted at a standardized resolution of 512x512 pixels; outputs from models generating higher resolutions are down-sampled to this size.

We employ a distillation weighting factor of $\lambda = 2.5$ across all experiments. Additionally, the R1 penalty strength

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
$\mathcal{L}_{\text{distill}}$	315.6	0.076
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{distill,exp}}$	**20.6**	0.319
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{distill,sds}}$	22.3	0.325
$\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{distill,nfsd}}$	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.

γ is set to 10^{-5}. For discriminator conditioning, we use a pretrained CLIP-ViT-g-14 text encoder [52] to compute text embeddings c_{text} and the CLS embedding of a DINOv2 ViT-L encoder [47] for image embeddings c_{img}. For the baselines, we use the best publicly available models: Latent diffusion models [50, 54] (SD1.5[1], SDXL[2]) cascaded pixel diffusion models [55] (IF-XL[3]), distilled diffusion models [39, 41] (LCM-1.5, LCM-1.5-XL[4]), and OpenMUSE [5] [48], a reimplementation of MUSE [6], 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 [59] 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 [18] and text alignment via CLIP score [17]. For CLIP score, we use ViT-g-14 model trained on LAION-2B [61]. Both metrics are evaluated on 5k samples from COCO2017 [34].

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. [67] suggest that ViTs trained with the CLIP [52] or DINO [5, 47] 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.

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. [59] and Kang et al. [25] observe a saturation of performance after a certain network capacity is reached. This observation contrasts the generally smooth scaling laws of DMs [49]. However, ADD can effectively leverage larger pretrained DMs (see Table 1c) and benefit from stable DM pretraining.

Loss terms. (Table 1d). We find that both losses are essential. The distillation loss on its own is not effective, but when combined with the adversarial loss, there is a noticeable improvement in results. Different 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 [23] could be considered.

[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

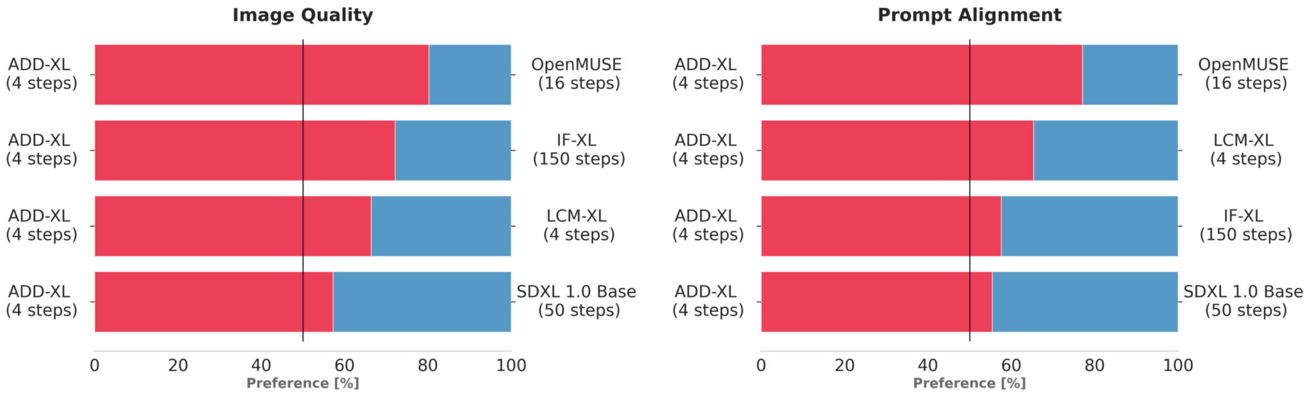


Figure 6. **User preference study (*multiple steps*).** We compare the performance of ADD-XL (4-step) against established baselines. Our ADD-XL model outperforms all models, including its teacher SDXL 1.0 (base, no refiner) [50], in human preference for both image quality and prompt alignment.

Method	#Steps	Time (s)	FID ↓	CLIP ↑
DPM Solver [37]	25 8	0.88 0.34	20.1 31.7	0.318 0.320
Progressive Distillation [43]	1 2 4	0.09 0.13 0.21	37.2 26.0 26.4	0.275 0.297 0.300
CFG-Aware Distillation [31]	8	0.34	24.2	0.300
InstaFlow-0.9B [36] InstaFlow-1.7B [36]	1 1	0.09 0.12	23.4 22.4	0.304 0.309
UFOGen [71]	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.

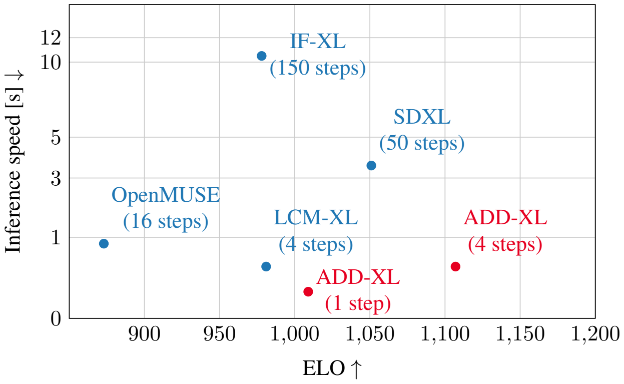


Figure 7. **Performance vs. speed.** We visualize the results reported in Fig. 6 in combination with the inference speeds of the respective models. The speeds are calculated for generating a single sample at resolution 512x512 on an A100 in mixed precision.

Teacher 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 [50].

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.

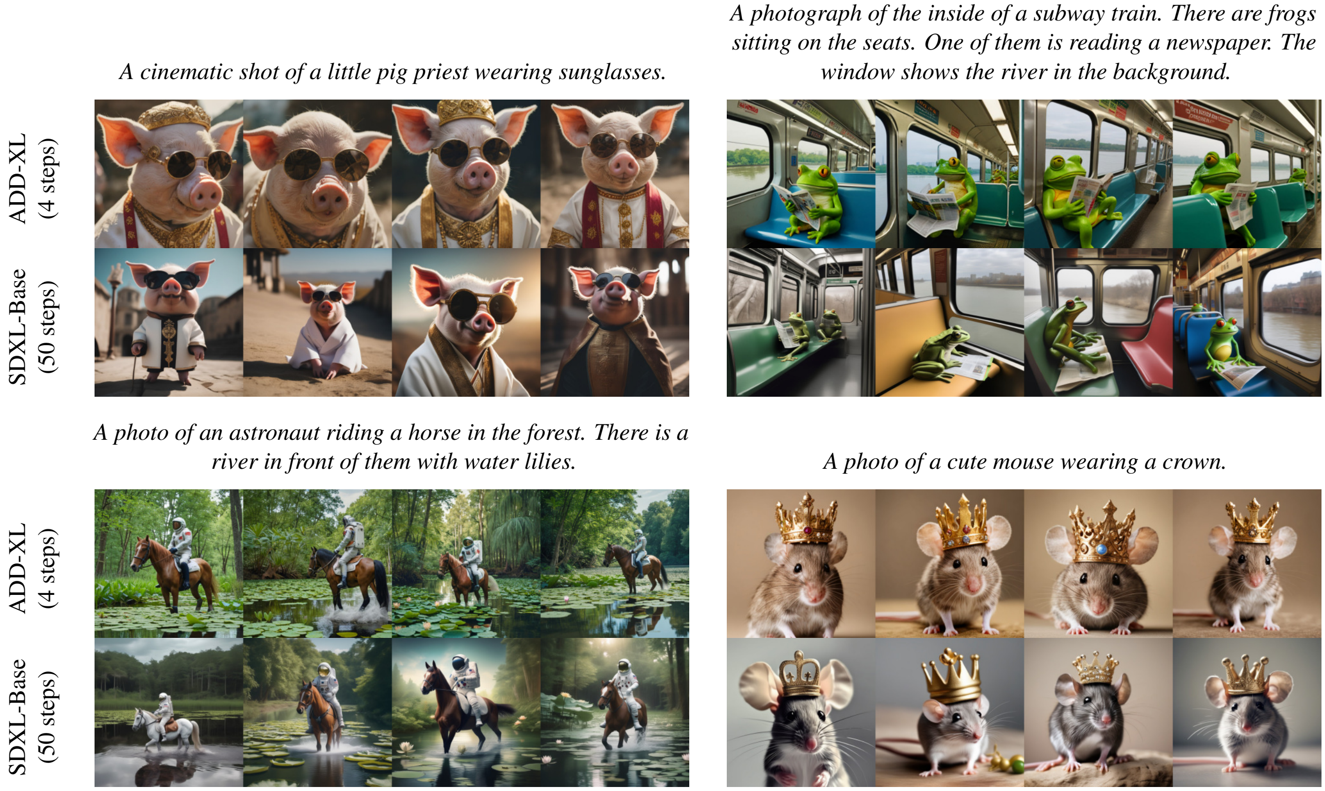
4.2. Quantitative Comparison to State-of-the-Art

For our main comparison with other approaches, we refrain from using automated metrics, as user preference studies are more reliable [50]. 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. 5 and Fig. 6 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. 7 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.

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 com-



A cinematic shot of a little pig priest wearing sunglasses.

A photograph of the inside of a subway train. There are frogs sitting on the seats. One of them is reading a newspaper. The window shows the river in the background.

A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.

A photo of a cute mouse wearing a crown.

Figure 8. **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.

parisons in the supplementary material. Fig. 3 compares ADD-XL (1 step) against the best current baselines in the few-steps regime. Fig. 4 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. 8 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. We combine an adversarial and a score distillation objective to distill the public Stable Diffusion [54] and SDXL [50] 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.

Acknowledgements

We would like to thank Jonas Müller for feedback on the draft, the proof, and typesetting; Patrick Esser for feedback on the proof and building an early model demo; Frederic Boesel for generating data and helpful discussions; Minguk Kang and Taesung Park for providing GigaGAN samples; Richard Vencu, Harry Saini, and Sami Kama for maintaining the compute infrastructure; Yara Wald for creative sampling support; and Vanessa Sauer for her general support.

References

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Cather-

ine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. 13

[2] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. 13

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *ArXiv*, abs/2211.01324, 2022. 1, 2

[4] A. Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *Proc. ICML*, 2023. 6

[7] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 4

[8] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems*, 35:30150–30166, 2022. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[10] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. 13

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. 2

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 1

[13] Jean-Yves Franceschi, Mike Gartrell, Ludovic Dos Santos, Thibaut Issenhuth, Emmanuel de Bézenac, Mickaël Chen, and Alain Rakotomamonjy. Unifying gans and score-based diffusion as generative particle models. *arXiv preprint arXiv:2305.16150*, 2023. 3

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014. 1, 2, 4

[15] Timofey Grigoryev, Andrey Voynov, and Artem Babenko. When, why, and which pretrained gans are useful? *ICLR*, 2022. 4

[16] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023. 3

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. EMNLP*, 2021. 6

[18] 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. *NeurIPS*, 2017. 6, 12

[19] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 2

[20] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 1

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022. 1, 2

[22] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 2

[23] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 6

[24] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020. 3

[25] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 1, 2, 6, 14

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. 1, 4, 14

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 8107–8116, 2019. 1, 4

[28] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 5

[29] 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. 3

[30] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 12

[31] 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. *arXiv preprint arXiv:2306.00980*, 2023. 7

[32] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 5

[33] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed, 2023. 4

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[35] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[36] 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, 3, 7, 15

[37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 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. 2, 7

[38] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *ArXiv*, abs/2310.04378, 2023. 2, 13

[39] 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. 6

[40] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolin'ario Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *ArXiv*, abs/2311.05556, 2023. 2, 3, 13, 15

[41] 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. 6

[42] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *arXiv preprint arXiv:2305.18455*, 2023. 3

[43] 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, 7

[44] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 5

[45] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12663–12673, 2022. 2

[46] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 4

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[48] Suraj Patil, William Berman, and Patrick von Platen. Amused: An open muse model. https://github.com/huggingface/diffusers, 2023. 6, 15

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 6

[50] 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. 2, 4, 5, 6, 7, 8, 12, 13

[51] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5, 12

[52] 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. 6, 12

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1, 2, 4

[54] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1, 2, 5, 6, 8

[55] 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. 4, 6

[56] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *CoRR*, abs/2202.00512, 2022. 2

[57] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 5

[58] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 1, 4

[59] 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. *Proc. ICML*, 2023. 2, 3, 4, 5, 6, 14

[60] Juergen Schmidhuber. Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991), 2020. 4

[61] 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. In *NeurIPS*, 2022. 6

[62] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 3

[63] Jascha Narain Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015. 1

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[65] Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. 1

[66] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023. 2

[67] George Stein, Jesse C Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, J Eric T Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *arXiv preprint arXiv:2306.04675*, 2023. 6, 14

[68] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2, 5

[69] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *ArXiv*, abs/2305.16213, 2023. 2

[70] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 3

[71] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. *arXiv preprint arXiv:2311.09257*, 2023. 7

[72] Chun-Han Yao, Amit Raj, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Artic3d: Learning robust articulated 3d shapes from noisy web image collections. *arXiv preprint arXiv:2306.04619*, 2023. 4

[73] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 12

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

Appendix

A. SDS As a Special Case of the Distillation Loss

If we set the weighting function to $c(t) = \frac{\alpha_t}{2\sigma_t} w(t)$ where $w(t)$ is the scaling factor from the weighted diffusion loss as in [51] and choose $d(x, y) = ||x - y||_2^2$, the distillation loss in Eq. (4) is equivalent to the score distillation objective:

$$
\begin{aligned}
&\frac{d}{d\theta}\mathcal{L}_{\text{distill}}^{\text{MSE}} \\
&\quad = \mathbb{E}_{t,\epsilon'}\left[c(t)\frac{d}{d\theta}||\hat{x}_\theta - \hat{x}_\psi(\text{sg}(\hat{x}_{\theta,t}); t)||_2^2\right] \\
&\quad = \mathbb{E}_{t,\epsilon'}\left[2c(t)[\hat{x}_\theta - \hat{x}_\psi(\hat{x}_{\theta,t}; t)]\frac{d\hat{x}_\theta}{d\theta}\right] \\
&\quad = \mathbb{E}_{t,\epsilon'}\left[\frac{\alpha_t}{\sigma_t}w(t)[\frac{1}{\alpha_t}(\hat{x}_{\theta,t} - \hat{x}_{\theta,t}) + \hat{x}_\theta - \hat{x}_\psi(\hat{x}_{\theta,t}; t)]\frac{d\hat{x}_\theta}{d\theta}\right] \\
&\quad = \mathbb{E}_{t,\epsilon'}\left[\frac{1}{\sigma_t}w(t)[(\alpha_t\hat{x}_\theta - \hat{x}_{\theta,t}) - (\alpha_t\hat{x}_\psi(\hat{x}_{\theta,t}; t) - \hat{x}_{\theta,t})]\frac{d\hat{x}_\theta}{d\theta}\right] \\
&\quad = \mathbb{E}_{t,\epsilon'}\left[\frac{w(t)}{\sigma_t}[-\sigma_t\epsilon' + \sigma_t\hat{\epsilon}_\theta(\hat{x}_{\theta,t}; t)]\frac{d\hat{x}_\theta}{d\theta}\right] \\
&\quad = \frac{d}{d\theta}\mathcal{L}_{\text{SDS}}
\end{aligned}
\tag{5}
$$

B. Details on Human Preference Assessment

For the evaluation results presented in Figures 5 to 7, we employ human evaluation and do not rely on commonly used metrics for quality assessment of generative models such as FID [18] and CLIP-score [52], since these have been shown to capture more fine grained aspects like aesthetics and scene composition only insufficiently [30, 50]. However these categories in particular have become more and more important when comparing current state-of-the-art text-to-image models. We evaluate all models based on 100 selected prompts from the PartiPrompts benchmark [73] with the most relevant categories (excluding prompts from the category *basic*). More details on how the study was conducted Appendix B.1 and the rankings computed Appendix B.2 are listed below.

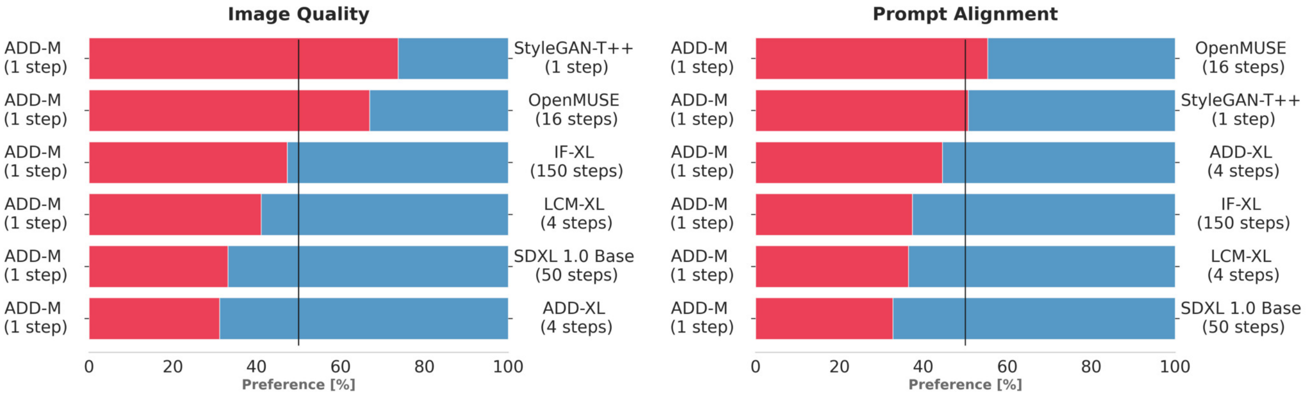


Figure 9. **User preference study (*single step*).** We compare the performance of ADD-M (1-step) against established baselines.

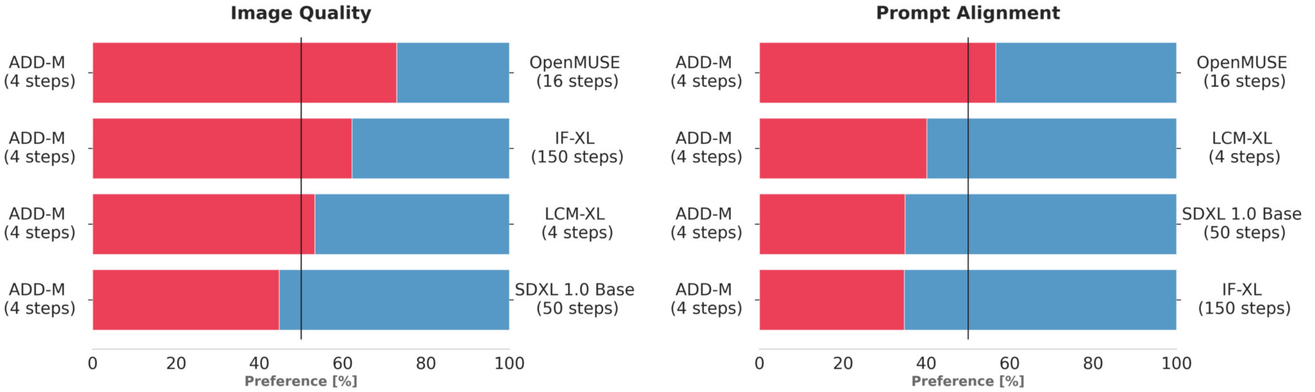


Figure 10. **User preference study (*multiple steps*).** We compare the performance of ADD-XL (4-step) against established baselines.

B.1. Experimental Setup

Given all models for one particular study (e.g. ADD-XL, OpenMUSE[6], IF-XL[7], SDXL [50] and LCM-XL[8] [38, 40] in Figure 7) we compare each prompt for each pair of models (1v1). For every comparison, we collect an average of four votes per task from different annotators, for both visual quality and prompt following. Human evaluators, recruited from the platform *Prolific*[9] with English as their first language, are shown two images from different models based on the same text prompt. To prevent biases, evaluators are restricted from participating in more than one of our studies. For the prompt following task, we display the text prompt above the two images and ask, “Which image looks more representative of the text shown above and faithfully follows it?” For the visual quality assessment, we do not show the prompt and instead ask, “Which image is of higher quality and aesthetically more pleasing?”. Performing a complete assessment between all pair-wise comparisons gives us robust and reliable signals on model performance trends and the effect of varying thresholds. The order of prompts and the order between models are fully randomized. Frequent attention checks are in place to ensure data quality.

B.2. ELO Score Calculation

To calculate rankings when comparing more than two models based on 1v1 comparisons we use ELO Scores (higher-is-better) [10] which were originally proposed as a scoring method for chess players but have more recently also been applied to compare instruction-tuned generative LLMs [1, 2]. For a set of competing players with initial ratings R_{init} participating in a series of zero-sum games the ELO rating system updates the ratings of the two players involved in a particular game based on the expected and and actual outcome of that game. Before the game with two players with ratings R_1 and R_2, the expected outcome for the two players are calculated as

$$E_1 = \frac{1}{1 + 10^{\frac{R_2 - R_1}{400}}}\,, \tag{6}$$

$$E_2 = \frac{1}{1 + 10^{\frac{R_1 - R_2}{400}}}\,. \tag{7}$$

After observing the result of the game, the ratings R_i are updated via the rule

$$R_i^{'} = R_i + K \cdot (S_i - E_i)\,, \quad i \in \{1, 2\} \tag{8}$$

where S_i indicates the outcome of the match for player i. In our case we have $S_i = 1$ if player i wins and $S_i = 0$ if player i looses. The constant K can be see as weight putting emphasis on more recent games. We choose $K = 1$ and bootstrap the final ELO ranking for a given series of comparisons based on 1000 individual ELO ranking calculations with randomly shuffled order. Before comparing the models we choose the start rating for every model as $R_{\text{init}} = 1000$.

[6] https://huggingface.co/openMUSE
[7] https://github.com/deep-floyd/IF
[8] https://huggingface.co/latent-consistency/lcm-lora-sdxl
[9] https://app.prolific.com

C. GAN Baselines Comparison

For training our state-of-the-art GAN baseline StyleGAN-T++, we follow the training procedure outlined in [59]. The main differences are extended training (∼2M iterations with a batch size of 2048, which is comparable to GigaGAN's schedule [25]), the improved discriminator architecture proposed in Section 3.2, and R1 penalty applied at each discriminator head.

Fig. 11 shows that StyleGAN-T++ outperforms the previous best GANs by achieving a comparable zero-shot FID to GigaGAN at a significantly higher CLIP score. Here, we do not compare to DMs, as comparisons between model classes via automatic metrics tend to be less informative [67]. As an example, GigaGAN achieves FID and CLIP scores comparable to SD1.5, but its sample quality is still inferior, as noted by the authors.

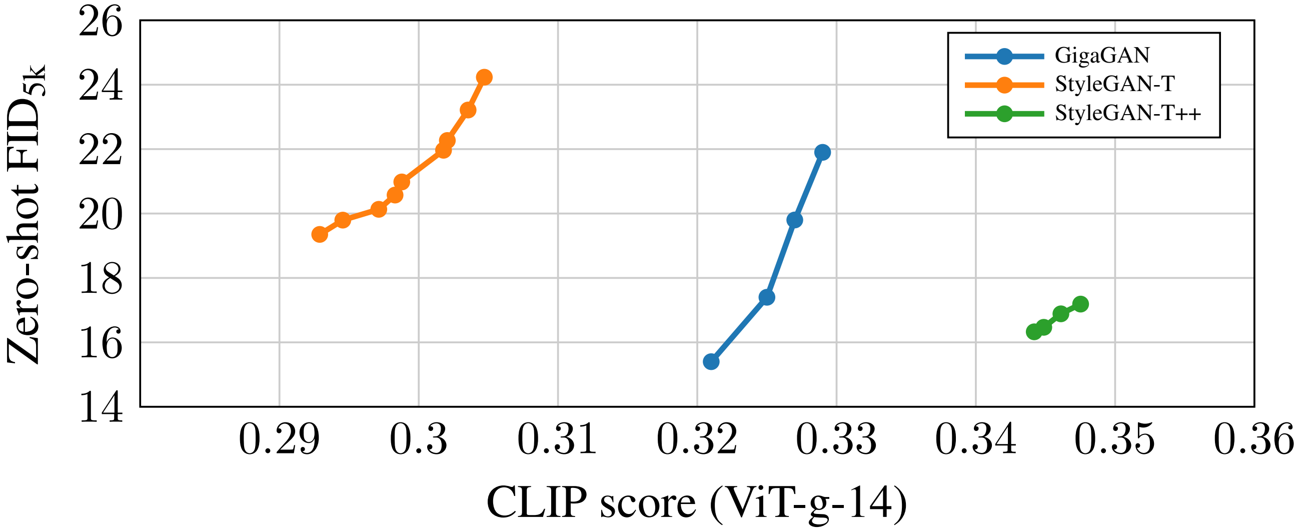


Figure 11. **Comparing text alignment tradeoffs at 256 × 256 pixels.** We compare FID–CLIP score curves of StyleGAN-T, StyleGAN-T++, and GigaGAN. For increasing CLIP score, all methods use via decreasing truncation [26] for values $\psi = \{1.0, 0.9, \ldots, 0.3\}$.

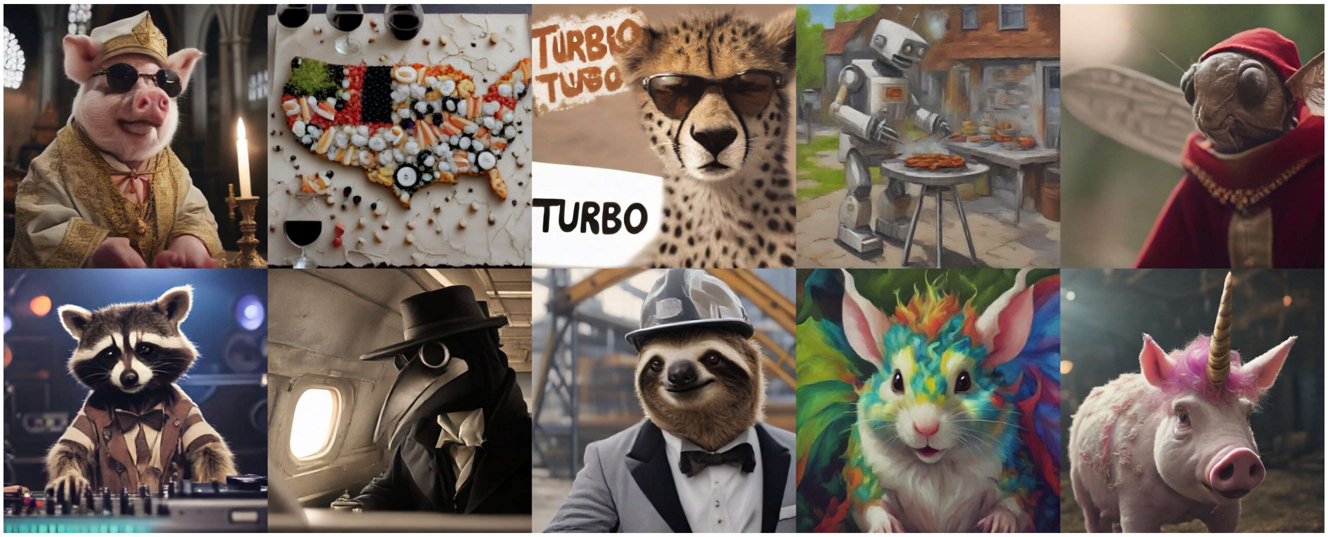
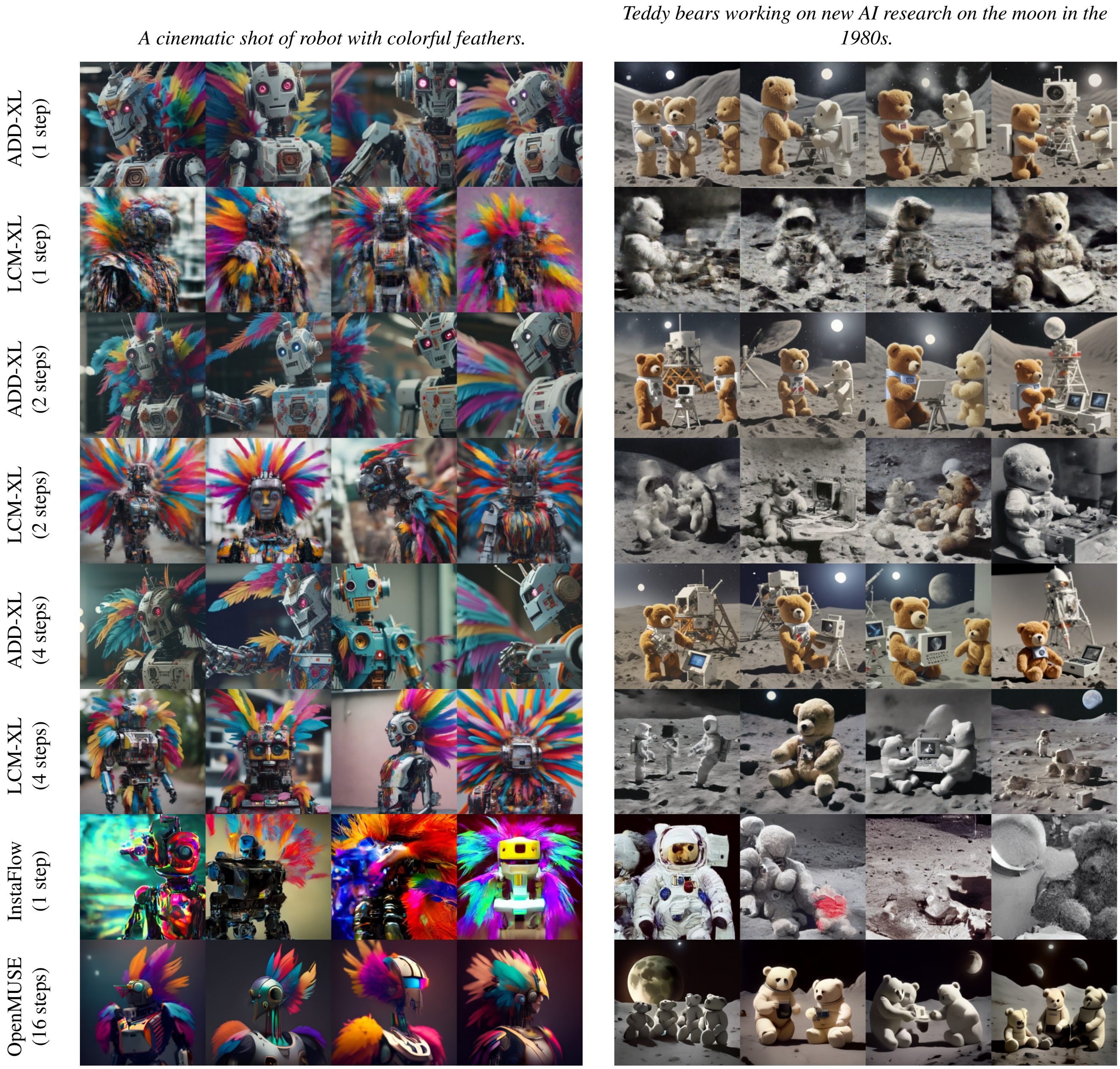


Figure 12. **Additional single step 512^2 images generated with ADD-XL.** All samples are generated with a single U-Net evaluation trained with adversarial diffusion distillation (ADD).

D. Additional Samples

We show additional one-step samples as in Figure 1 in Figure 12. An additional qualitative comparison as in Figure 4 which demonstrates that our model can further refine quality by using more than one sampling step is provided in Figure 14, where we show that, while sampling quality with a single step is already high, more steps can give higher diversity and better spelling capabilities. Lastly, we provide an additional qualitative comparison of ADD-XL to other state-of-the-art one and few-step models in Figure 13.



A cinematic shot of robot with colorful feathers.

Teddy bears working on new AI research on the moon in the 1980s.

Figure 13. **Additional qualitative comparisons to state of the art fast samplers.** Few step samples from our ADD-XL and LCM-XL [40], InstaFlow [36], and OpenMuse [48].

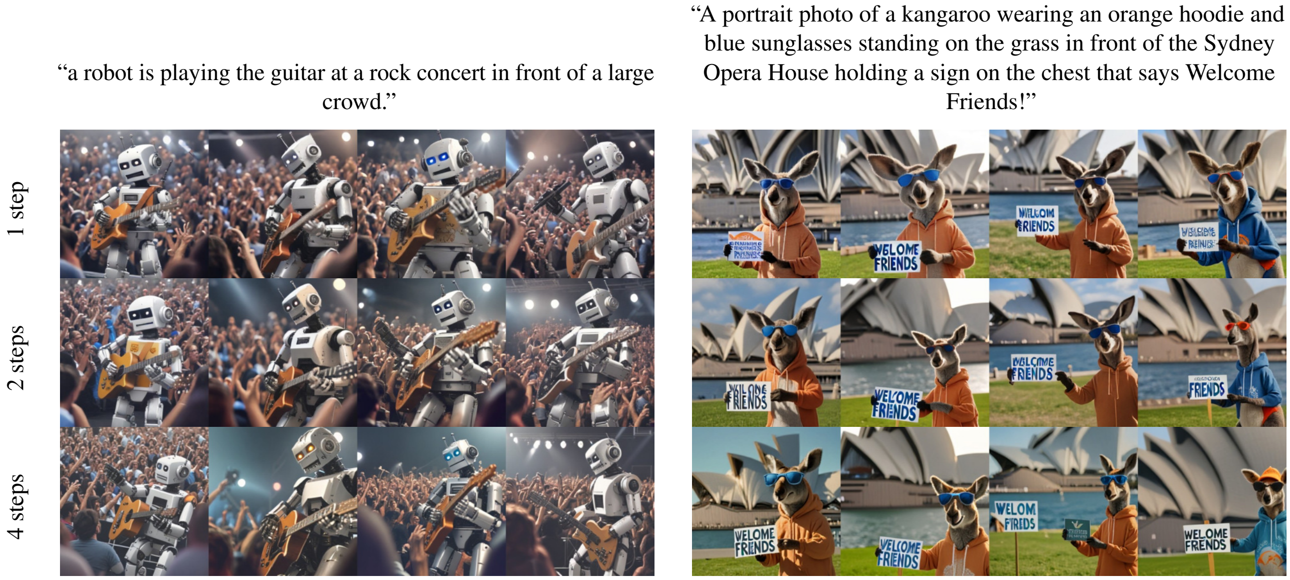

“a robot is playing the guitar at a rock concert in front of a large crowd.”
“A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!”
1 step
2 steps
4 steps

Figure 14. **Additional results on the qualitative effect of sampling steps.** Similar to Figure 4, 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 diversity (left) and spelling capabilities (right). 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.