Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation

Axel Sauer   **Frederic Boesel**   **Tim Dockhorn**

Andreas Blattmann   **Patrick Esser**   **Robin Rombach**

Stability AI

Figure 1: **Generating high-resolution multi-aspect images with *SD3-Turbo*.** All samples are generated with a maximum of four transformer evaluations trained with latent adversarial diffusion distillation (LADD).

Abstract

Diffusion models are the main driver of progress in image and video synthesis, but suffer from slow inference speed. Distillation methods, like the recently introduced adversarial diffusion distillation (ADD) aim to shift the model from many-shot to single-step inference, albeit at the cost of expensive and difficult optimization due to its reliance on a fixed pretrained DINOv2 discriminator. We introduce Latent Adversarial Diffusion Distillation (LADD), a novel distillation approach overcoming the limitations of ADD. In contrast to pixel-based ADD, LADD utilizes generative features from pretrained latent diffusion models. This approach simplifies training and enhances performance, enabling high-resolution multi-aspect ratio image synthesis. We apply LADD to Stable Diffusion 3 (8B) to obtain *SD3-Turbo*, a fast model that matches the performance of state-of-the-art text-to-image generators using only four unguided sampling steps. Moreover, we systematically investigate its scaling behavior and demonstrate LADD's effectiveness in various applications such as image editing and inpainting.

1 Introduction

While diffusion models [54, 18, 57, 30, 31] have revolutionized both synthesis and editing of images [10, 42, 41, 1, 43, 13, 9, 38] and videos [5, 4, 53, 2, 12, 19], their iterative nature remains a crucial shortcoming: At inference, a trained diffusion model usually requires dozens of network evaluations to approximate the probability path from noise to data. This makes sampling slow, in particular for large models, and limits real-time applications.

Naturally, a large body of work focuses on speeding up the sampling of diffusion models — both via improved samplers [55, 11, 66, 51] and distilled models that are trained to match the sample quality of their teacher models in fewer steps [44, 35, 34, 58]. Very recent distillation works aim at reducing the number of model evaluations to a single step, enabling real-time synthesis [63, 34, 62, 49, 28]. The best results in the one- and few-step regime are currently achieved with methods that leverage adversarial training [50, 62, 49, 28], forcing the output distribution towards the real image manifold. Adversarial Diffusion Distillation (ADD) [49] provides the current state-of-the-art method for single-step synthesis: By leveraging a pretrained DINOv2 [36] feature extractor as the backbone of the discriminator, ADD manages to distill SDXL [38] into a single-step, real-time text-to-image model.

However, while achieving impressive inference speed, ADD comes with a series of shortcomings: First, the usage of the fixed and pretrained DINOv2 network restricts the discriminator's training resolution to 518×518 pixels. Furthermore, there is no straightforward way to control the feedback level of the discriminator, e.g., for weighting global shape vs. local features differently. Finally, for distilling latent diffusion models, ADD needs to decode to RGB space, as the discriminator has not been trained in latent space, which significantly hinders high-resolution training $> 512^2$ pixels.

More generally, and in contrast to large language models [25, 20] and diffusion models [37, 13], current adversarial models do not strictly adhere to scaling laws, and stable training methods usually require extensive hyperparameter tuning. In fact, previous attempts at scaling GANs resulted in diminishing returns when scaling the generator [48, 24]. Even more surprisingly, smaller discriminator feature networks often offer better performance than their larger counterparts [49, 48]. These non-intuitive properties are a significant shortcoming for GAN practitioners: Models that follow scaling laws offer predictable improvements in performance, allowing for more strategic and cost-effective scaling, and ultimately better model development.

In this work, we present *Latent Adversarial Diffusion Distillation* (LADD), an approach that offers stable, scalable adversarial distillation of pretrained diffusion transformer models [37, 13] up to the megapixel regime: Instead of utilizing discriminative features of, e.g., self-supervised feature networks such as DINOv2, we leverage generative features of a pretrained diffusion model. While directly enabling multi-aspect training, this approach also offers a natural way to control the discriminator features: By targeted sampling of the noise levels during training, we can bias the discriminator features towards more global (high noise level) or local (low noise level) behavior. Furthermore, *distillation in latent space* allows for leveraging large student and teacher networks and avoids the expensive decoding step to pixel space, enabling high-resolution image synthesis. Consequently, LADD results in a significantly simpler training setup than ADD while outperforming all prior single-step approaches.

We apply LADD to the current state-of-the-art text-to-image model Stable Diffusion 3 [13] and obtain *SD3-Turbo*, a multi-aspect megapixel generator that matches its teacher's image quality in only four sampling steps. In summary, the core contributions of our work are

- *SD3-Turbo*, a fast foundation model supporting high-resolution multi-aspect image generation from text prompts, see Fig. 1 and Fig. 2,

- a significantly simplified distillation formulation that outperforms LADD's predecessor ADD [49] and a systematic study of LADD's scaling behavior,

- a demonstration of the versatility of our approach via two exemplary applications: image editing and image inpainting.

We will make code and model weights publicly available.

2

A high-quality photo of a spaceship that looks like the head of a horse.

A group of quirky robot animals, with parts made of different metals and machinery, playing in a futuristic park with holographic trees.

An anthropomorphic clock character in a bustling city square, interacting with time-themed creatures.

A macro shot of a flower with a bee wearing sunglasses on it that holds a sign saying: "turbo!"

Photo of a T-Rex wearing a cap sitting at a bonfire with his human friend

A close-up shot of a skateboard on a colorful graffiti-filled backdrop in an urban setting, capturing the essence of street culture.

A realistic, detailed photograph of a baguette with human teeth. The baguette is wearing hiking boots and an old-school skiing suit.

Moloch whose eyes are a thousand blind windows, whose skyscrapers stand in the long streets, whose smoke-stacks and antennae crown the cities!

A photograph of a pig with a unicorn's horn.

A close-up of a woman's face, lit by the soft glow of a neon sign in a dimly lit, retro diner, hinting at a narrative of longing and nostalgia.

A dramatic shot of a classic detective in a trench coat and fedora, standing in a rain-soaked alleyway under a dim streetlight.

An origami eagle flying through a living room.

candid photo of santa in my living room placing boxes of cheese under the christmas tree

Figure 2: **More high-resolution multi-aspect images generated with *SD3-Turbo*.** All samples are generated with a maximum of four transformer evaluations. damn

3

$$x_T = \alpha_T x_0 + \sigma_T \varepsilon$$
$$x_0 = \frac{x_T - \sigma_T \varepsilon}{\alpha_T}$$
$$x_0 = \frac{x_t - t\varepsilon}{1-t}$$

2 Background

2.1 Diffusion Models

$$D(x_t, t) = x_t - t\, F_\theta(x_t, t)$$

Flow based model

Diffusion models learn to iteratively denoise Gaussian noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into data. The learnable component in diffusion models is a *denoiser* D that predicts the expected image $\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t, t]$ given a noisy image $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \varepsilon$. While in this work we focus on the rectified flow formulation [31] where $\alpha_t = 1 - t$ and $\sigma_t = t$ for $t \in [0, 1]$, and the denoiser is parameterized as $D(\mathbf{x}_t, t) = \mathbf{x}_t - t \cdot F_\theta(\mathbf{x}_t, t)$, where F_θ is a large neural network, our method is generally applicable to any diffusion model formalism. The denoiser can be trained via *score matching* [21, 60],

$$\min_\theta \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0), \varepsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t \sim p(t)} \left[\lambda(t) \| D(\mathbf{x}_t, t) - \mathbf{x}_0 \|_2^2 \right], \tag{1}$$

where $p(\mathbf{x}_0)$ is the empirical data distribution, $p(t)$ is a (continuous) distribution over $t \in [0, 1]$ and λ is a weighting function. After training, we can generate realistic samples by numerically solving a (stochastic) differential equation (backwards from $t=1$ to $t=0$) [57, 26], iteratively evaluating the learned denoiser D.

2.2 Diffusion Distillation

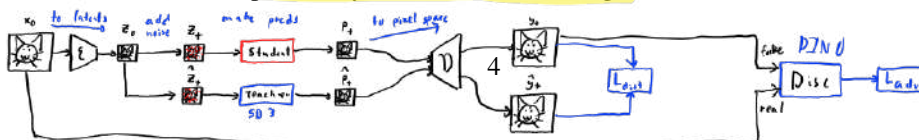*Space from t=1 to t=0 is very nonlinear despite being "linear"*

While the denoiser D learns to predict clean images with sharp high frequency details for sufficiently small t, it also learns to approximate the mean of the empirical data distribution for large t, resulting in a highly non-linear differential equation. Therefore, one needs to solve the differential equations with sufficiently small step sizes, resulting in many (expensive) evaluations of the network F_θ.

Can learn to make the space more "flat" allowing larger steps.

For many applications, such as text-to-image generation, we are, however, only interested in the final (clean image) distribution at $t=0$ which can be obtained from a multitude of different differential equations. In particular, many distillation techniques attempt to learn "simpler" differential equations that result in the same distribution at $t=0$ however with "straighter", more linear, trajectories (which allows for larger step sizes and therefore less evaluations of the network F_θ). Progressive Distillation [44], for example, tries to distill two Euler steps into a single Euler step. This technique iteratively halves the number of steps required, however, it suffers from error accumulation as generally five or more rounds of distillation are needed to obtain a fast model. Reflow [31] is another distillation technique where new models are trained iteratively on synthetic data from older models, and therefore also suffers from error accumulation. In contrast, Consistency Distillation [58] distills models in a single stage without iterative application, however, the training process is quite unstable and requires advanced techniques such as distillation schedules [58], and extensive hyperparameter tuning. Improved techniques for both Consistency Distillation [56, 33, 15, 68] and Progressive Distillation [35, 28, 3] have since been introduced. The current top-performing distillation methods for text-to-image applications utilize adversarial training. In particular, Adversarial Diffusion Distillation (ADD) [49], uses a pretrained feature extractors as its discriminator, achieving performance on par with strong diffusion models such as SDXL [38] in only four steps.

3 Method

By leveraging a lower-dimensional latent space, latent diffusion models (LDMs) [42] significantly reduce memory requirements for training, facilitating the efficient scaling of to large model size and high resolutions. This advantage is exemplified by the recently introduced MMDiT family [13] of LDMs where the largest model (8B parameters) achieves state-of-the art text-to-image synthesis performance. Our goal is to distill such large LDMs efficiently for high-resolution, multi-aspect image synthesis. *Latent adversarial diffusion distillation* (LADD), simplifies the distillation process by eliminating the necessity of decoding back to the image space, thereby significantly reducing memory demands in comparison to its predecessor, ADD.

Distillation in latent space. An overview of LADD and comparison to ADD is shown in Fig. 3. In ADD, the ADD-student receives noised input images x_t at the timestep t and generates samples $\hat{x}_\theta(x_t, t)$ aiming to optimize for two objectives: an adversarial loss L_{adv}, which involves deceiving a discriminator, and a distillation loss $L_{distill}$, which involves matching the denoised output to that of a frozen DM teacher. LADD introduces two main modifications: the unification of discriminator and teacher model, and the adoption of synthetic data for training.

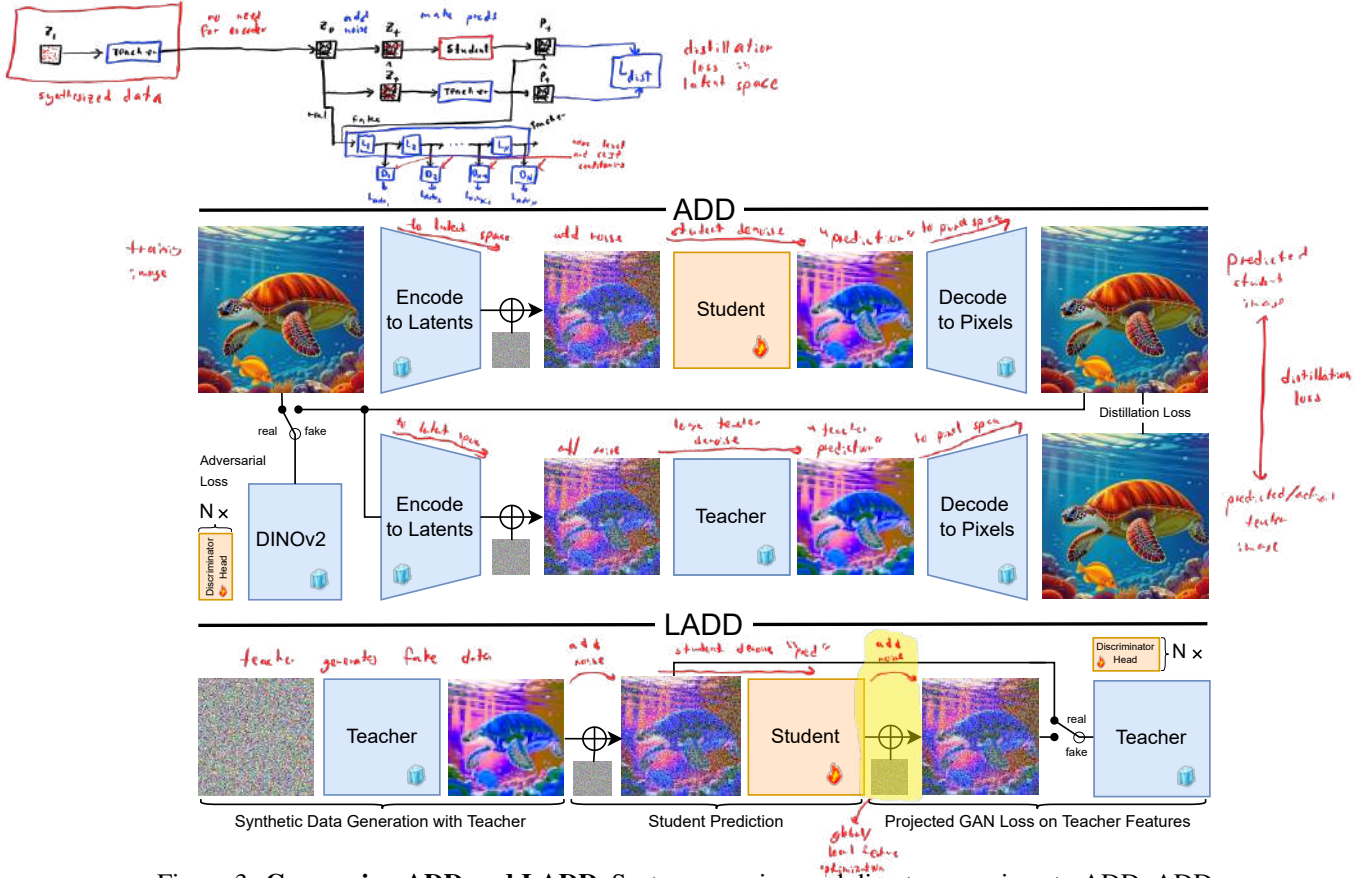*LADD has the same discriminator/teacher and uses synthetic data*

Figure 3: **Comparing ADD and LADD.** System overview and direct comparison to ADD. ADD (top two rows) computes a distillation loss in pixel space and an adversarial loss on top of DINOv2 features, thereby requiring expensive decoding from latent space to pixel space. In LADD (bottom row), we use the teacher model for synthetic data generation and its features for the adversarial loss, which allows us to train purely in the latent space.

Unifying teacher and discriminator. Instead of decoding and applying a discriminator in image space, we operate exclusively on latents. First, we renoise the generated latents at timestep $\hat{t}$ drawn from a logit-normal distribution, following [13]. We then apply the teacher model to the noised latents, extracting the full token sequence after each attention block. On each token sequence, we apply independent discriminator heads. Additionally, each discriminator is conditioned on the noise level and pooled CLIP embeddings.

ADD leverages the Projected GAN paradigm [46], i.e., applying independent discriminators on features obtained from pretrained features network. We can distinguish these feature networks depending on the pretraining task which is either *discriminative* (classification, self-supervised objective) or *generative* (diffusion objective). Utilizing generative features presents several key benefits over discriminative ones:

- **Efficiency and Simplification.** Generative features eliminate the need for decoding to image space, thereby saving memory and simplifying the overall system compared to ADD. Another possible option is training a discriminative feature network in latent space, yet, discriminative pretraining is non-trivial and top-performing approaches require significant engineering [8, 36].

- **Noise-level specific feedback.** Generative features vary with noise level, providing structured feedback at high noise levels and texture-related feedback at low noise levels [1, 32]. By adjusting the parameters of the noise sampling distribution, we gain direct control over discriminator behavior, aligning with the standard practice of loss weighting in diffusion model training [26, 13]

- **Multi-Aspect Ratio (MAR).** Since the teacher model is trained on MAR data, it inherently generates relevant features for the discriminators in in this setting.

- **Alignment with Human Perception.** Discriminative models exhibit a notable *texture bias* [14], prioritizing texture over global shape, unlike humans who tend to rely on global shape. Jaini et al. [22] demonstrates that generative models possess a shape bias closely resembling that of humans and achieve near human-level accuracy on out-of-distribution

tasks. This suggests that leveraging pretrained generative features for adversarial training could enhance alignment with human perception.

For the discriminator architecture, we mostly follow [48, 49]. However, instead of utilizing 1D convolution in the discriminator, we reshape the token sequence back to its original spatial layout, and transition to 2D convolutions. Switching from 1D to 2D convolutions circumvents a potential issue in the MAR setting, where a 1D discriminator would process token sequences of varying strides for different aspect ratios, potentially compromising its efficacy.

CFG does not work in one-step methods

Leveraging synthetic data. Classifier-free guidance (CFG) [17] is essential for generating high-quality samples. However, in one-shot scenarios, CFG simply oversaturates samples rather than improving text-alignment [48]. This observation suggests that CFG works best in settings with multiple steps, allowing for corrections of oversaturation issues ins most cases. Additional techniques like dynamic thresholding further ameliorate this issue [43].

synthetic data more aligned with prompt than human data

Text-alignment varies significantly across natural datasets. For instance, while COCO [29] images reach an average CLIP [1] score [39] of 0.29, top-performing diffusion models can achieve notably higher CLIP scores, e.g. SD3 attains a CLIP score of 0.35 on COCO prompts. CLIP score is an imperfect metric, yet, the large score differential between natural and synthetic data suggests that generated images are better aligned for a given prompt on average. To mitigate this issue and avoid additional complexity that is introduced by an auxiliary distillation loss as in ADD, we opt for synthetic data generation via the teacher model at a constant CFG value. This strategy ensures high and relatively uniform image-text aligned data and can be considered as an alternative approach for distilling the teacher's knowledge.

As LADD eliminates the need for decoding, we can directly generate latents with the teacher model and omit the additional encoding step for real data. For conditioning of the teacher, we sample prompts from the original training dataset of SD3.

4 Experiments

In this section, we evaluate our approach in the single-step setting, i.e., starting from pure noise inputs. For evaluation, we compute the CLIP score on all prompts from DrawBench [43] and PartiPrompts [64]. We train for 10k iterations and the default model for the student, teacher, and data generator is an MMDiT with a depth of 24 ($\sim$2B parameters) if not explicitly stated otherwise. Accordingly, the qualitative outputs in this section are generally of lower quality than the ones of our final (larger) model.

4.1 Teacher noise distribution

denoising from low noise results in a lack of global coherence, but good local features.

Fig. 4 illustrates the effect of different parametrization for the logit-normal distributions $\pi(t; m, s)$ of the teacher. When biasing the distribution towards low noise values, we observe missing global coherence while textures and local patches look realistic. Lacking global coherence is a common problem in adversarial training and additional losses such as classifier or CLIP guidance are often introduced to improve image quality [47, 48]. While increasing the bias towards higher noise levels improves coherence, excessively high noise levels can detrimentally affect texture and fine details. We find $\pi(t; m = 1, s = 1)$ to be solid choice which we will use for the remainder of this work.

4.2 Synthetic data

We aim to answer two questions: Does synthetic data lead to improvements in image-text alignment over real data? And, is an additional distillation loss $L_{distill}$ necessary? Fig. 5 displays the findings. Training with synthetic data significantly outperforms training with real data. While a distillation loss benefits training with real data, it offers no advantage for synthetic data. Thus, training on synthetic data can be effectively conducted using only an adversarial loss.

---

[1] We compute CLIP score using the ViT-g-14 model available at https://github.com/mlfoundations/open_clip
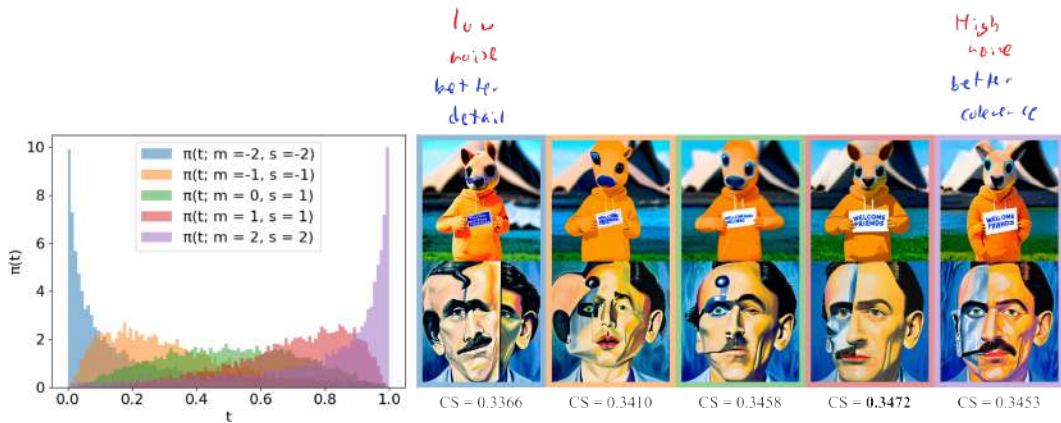
Figure 4: **Controlling the teacher noise distribution.** We vary the parameters of a logit-normal distribution for biasing the sampling of the teacher noise level. Shifting to higher noise improves overall coherence. When biasing towards very high noise levels ($m = 2, s = 2$), we observe a loss of fine details.

Figure 5: **Synthetic data improves image-text alignment.** We compare outputs for a fixed seed and the prompts "panda scientist mixing chemicals" and "a red car on a scenic road above a cliff." When training on real data, an additional distillation $L_{distill}$ improves details and thereby increases image-text alignment. Training on synthetic data substantially outperforms training on real data rendering the distillation loss obsolete.

4.3 Latent distillation approaches

Consistency Distillation [58] is another recent and popular approach for distillation. Latent consistency models (LCM) [33, 34] leverage consistency distillation for LDMs where training is conducted exclusively in latent space, similarly to LADD. For a fair comparison, we train the same student model with LCM and LADD. We observe much higher volatility for LCM than for LADD training, i.e., outcomes vastly differ for small changes in hyperparameters, different random seeds, and training iterations. For LCM, we run a hyperparameter grid search over the *skipping-step* [33], noise schedule, and full-finetuning (with and without EMA target [56]) vs LoRA-training [34] and select the best checkpoint out of all runs and over the course of training. For LADD, we train only once and select the last checkpoint. As Fig. 6 shows, LADD outperforms LCM by a large margin.

As discussed in Section 2, Consistency Distillation may require heavy hyperparameter tuning. To the best of our knowledge, we are the first work that attempting LCM training on Diffusion Transformers [37, 13], and it may be possible that we have not explore the hyperparameter space well enough. We want to highlight that LCM can potentially achieve more impressive results, as shown by SDXL-LCM [34, 33] to which we compare in Section 5.1. We hypothesize that larger models may facilitate LCM training, as evidenced by the substantial improvement when transitioning from SD1.5-LCM to SDXL-LCM [33]. Nonetheless, our experimental findings indicate that LADD can distill both small and large models effectively and without extensive hyperparameter tuning.
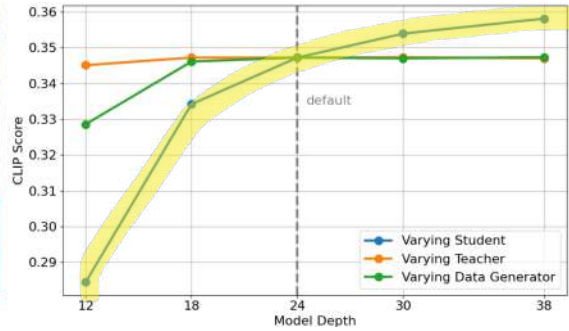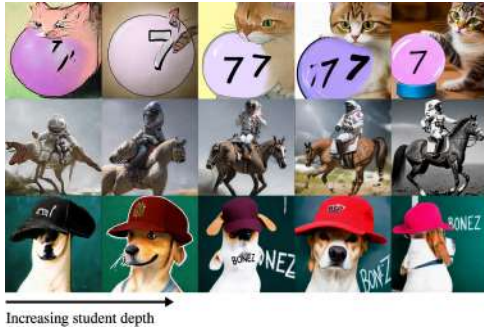
4.4 Scaling Behavior

We consider three dimension for scaling model size: student, teacher, and data generator. For the following experiments, we keep two dimensions constant at the default setting (depth=24), allowing variation in just one. We utilize the models of the scaling study evaluated in [13].

Fig. 7 presents the results. Student model size significantly impacts performance, surpassing both data quality and teacher model size in influence. Consequently, larger student models do not only demonstrate superior performance as diffusion models [13], but that performance advantage is

LADD outperforms LCM

LCM (4 steps)
CS = 0.2712

LADD (1 step)
CS = **0.3472**

Figure 6: **Comparing latent distillation approaches.** We distill an MMDiT (depth=24) with both LCM and LADD. For LADD, we use the same model as a teacher and data generator. We find that LADD consistently outperforms LCM in a single step.

student benefits from scale

Increasing student depth

Figure 7: **Scaling behaviour.** We ablate the size of student, teacher, and data generator model. Our default setting is a depth of 24 for all models and we vary one dimension at a time. A tangible difference is particularly noticeable when varying student depth. We show samples for a fixed seed and the following prompts: "a cat patting a crystal ball with the number 7 written on it in black marker", "an astronaut riding a horse in a photorealistic style", and "a dog wearing a baseball cap backwards and writing BONEZ on a chalkboard" (*left, top to bottom*).

effectively transferred to their distilled versions. While teacher models and data quality contribute to improvements, their benefits plateau, indicating diminishing returns beyond certain thresholds. This pattern suggests a strategy for optimizing resource allocation, especially under memory constraints, by prioritizing larger student models while allowing for smaller teacher models without substantially compromising performance.

4.5 Direct preference optimization.

For better human preference alignment, we finetune our models via *Diffusion DPO* ([61]), an adaption of the Direct Preference Optimization (DPO) [40] technique to diffusion models. In particular, we introduce learnable Low-Rank Adaptation (LoRA) matrices (of rank 256) for all linear layers into the teacher model and finetune it for 3k iterations with the DPO objective. For the subsequent LADD training, we use the DPO-finetuned model for student, teacher, and data generation. Interestingly, we find that we can further improve our LADD-student model by reapplying the original DPO-LoRA matrices. The resulting model achieves a win rate of 56% in a human preference study against the initial, non-DPO LADD-student evaluated at a single step. The human preference study follows the procedures outlined in Section A. DPO is even more impactful in the multi-step setting, as shown in the qualitative examples in Fig. 8.

5 Comparison to State-of-the-Art

Our evaluations begin with the text-to-image synthesis setting. We then progress to image-to-image tasks, demonstrating the universal applicability of our distillation approach. We adopt a training strategy that incorporates both full and partial noise inputs to enable multi-step inference. For multi-step inference we employ a flow consistency sampler.

Figure 8: **Applying DPO to LADD students.** Samples are generated by our best 8B model at 4 steps. After LADD training, we apply pretrained DPO-LoRA matrices to our student, which adds more details, fixes duplicates objects (e.g. car wheels), improves hands, and increases overall visual appeal (*bottom*).
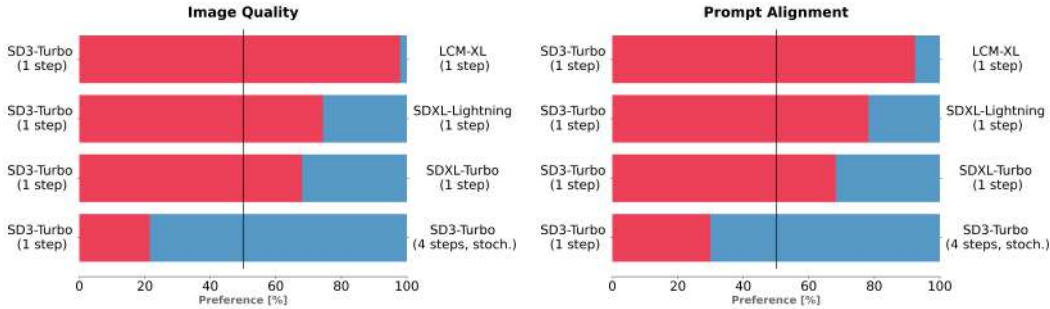

Figure 9: **User preference study (*single step*).** We compare the performance of our model against established baselines. Our model clearly outperforms all other baselines in human preference for both image quality and prompt alignment. Using more sampling steps further improves our model's results (bottom row).

We train across four discrete timesteps $t \in [1, 0.75, 0.5, 0.25]$. For two- and four-step inference, we found the consistency sampler proposed in [58] to work well. For two step inference, we evaluate the model at $t \in [1, 0.5]$. At higher resolutions ($> 512^2$ pixels), an initial warm-up phase is crucial for training stability; thus, we start with lower noise levels (initial probability distribution $p = [0, 0, 0.5, 0.5]$). After 500 iterations, the focus shifts towards full noise ($p = [0.7, 0.1, 0.1, 0.1]$) to refine single-shot performance. Lastly, MAR training follows the binning strategy outlined in [38, 13].

5.1 Text-to-Image Synthesis

For our main comparison to other approaches, we conduct user preference studies, assessing image quality and prompt alignment, see Section A for details. Fig. 9 presents the results in the single step setting. SD3-Turbo clearly outperforms all baselines in both image quality and prompt alignment. Taking four steps instead of one significantly improves results further which we also illustrate in Fig. 11. We also evaluate SD3-Turbo at four steps against various state-of-the-art text-to-image models in Fig. 10. SD3-Turbo reaches the same image quality as its teacher model SD3, but in four instead of 50 steps. Although there is a slight reduction in prompt alignment relative to SD3, SD3-Turbo still beats strong baselines like Midjourney v6. We provide high-resolution, multi-aspect samples from SD3-Turbo in Fig. 1 and Fig. 2.
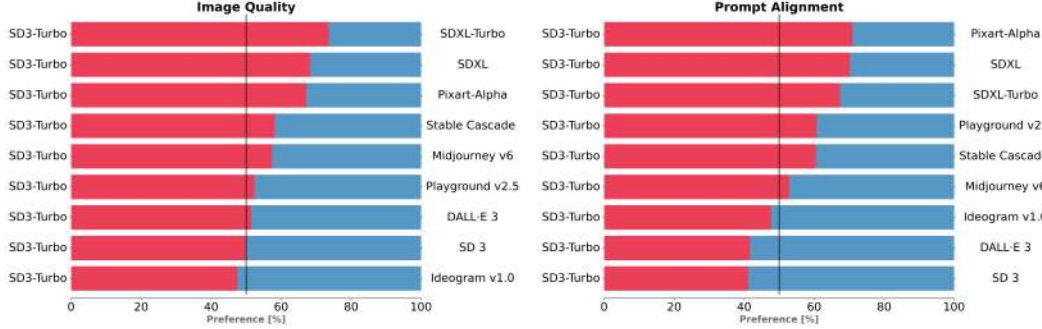
Figure 10: **User preference study (*multiple steps*).** We compare SD3-Turbo 1024^2-MAR to SOTA text-to-image generators. Our model, using four sampling steps, outperforms or is on par with all evaluated systems. We use default settings for all other multi-step samplers and four steps for SDXL-Turbo. For the SDXL-Turbo comparison, we downsample the SD3-Turbo outputs to 512^2 pixels.

Figure 11: **Qualitative effect of sampling steps.** We show qualitative examples when sampling SD3-Turbo with 1 and 4 steps; seeds are constant within columns. †: We only show the first sentence of the prompt to save space. The remainder is as follows: *The painting depicts a tranquil river between mountains. a ship gently bobbing in the water and entering the living room. The river's edge spills onto the wooden floor, merging the world of art with reality. The living room is adorned with tasteful furniture and a warm, inviting atmosphere., cinematic, photo, poster.*.

5.2 Image-to-Image Synthesis

It is straightforward to apply LADD to tasks other than text-to-image synthesis. To validate this claim, we apply LADD to instruction-guided image editing and image inpainting. We first continue training the pretrained text-to-image diffusion model with the diffusion objective and the dataset adjusted for the respective task. We refer to these models as *SD3-edit* (depth=24) and *SD3-inpainting* (depth=18) respectively. We then apply LADD as described in Sec. 3 to distill the image-to-image models, resulting in *SD3-edit Turbo* and *SD3-inpainting Turbo*.

Image Editing For the image editing task we consider instruction-based editing [7]. Following [7, 52], we condition on the input image via channel-wise concatenation and train on paired data with edit instructions. We use the synthetic InstrucPix2Pix dataset, for which we follow [6] and upsample the original 512^2 pixel samples using SDXL [38]. Similar to [52] we use additional data from bidirectional controlnet tasks (canny edges, keypoints, semantic segmentation, depth maps, HED lines) as well as object segmentation. During sampling, we guide the edit model with a nested classifier-free guidance formulation [17, 7], which allows us to utilize different strengths w for the image and text conditioning.

Fig. 12 shows the effectiveness of the distilled model especially for style editing tasks and object swaps by integrating the edited object well with the scene. We attribute this improved harmonization capability compared to other approaches to the adversarial loss. In Fig. 13 (Left), we plot the trade-off between CLIP image similarity and CLIP image editing direction similarity [39, 7]. We observe that our student model matches the performance of its teacher in a single step. The notable increase in
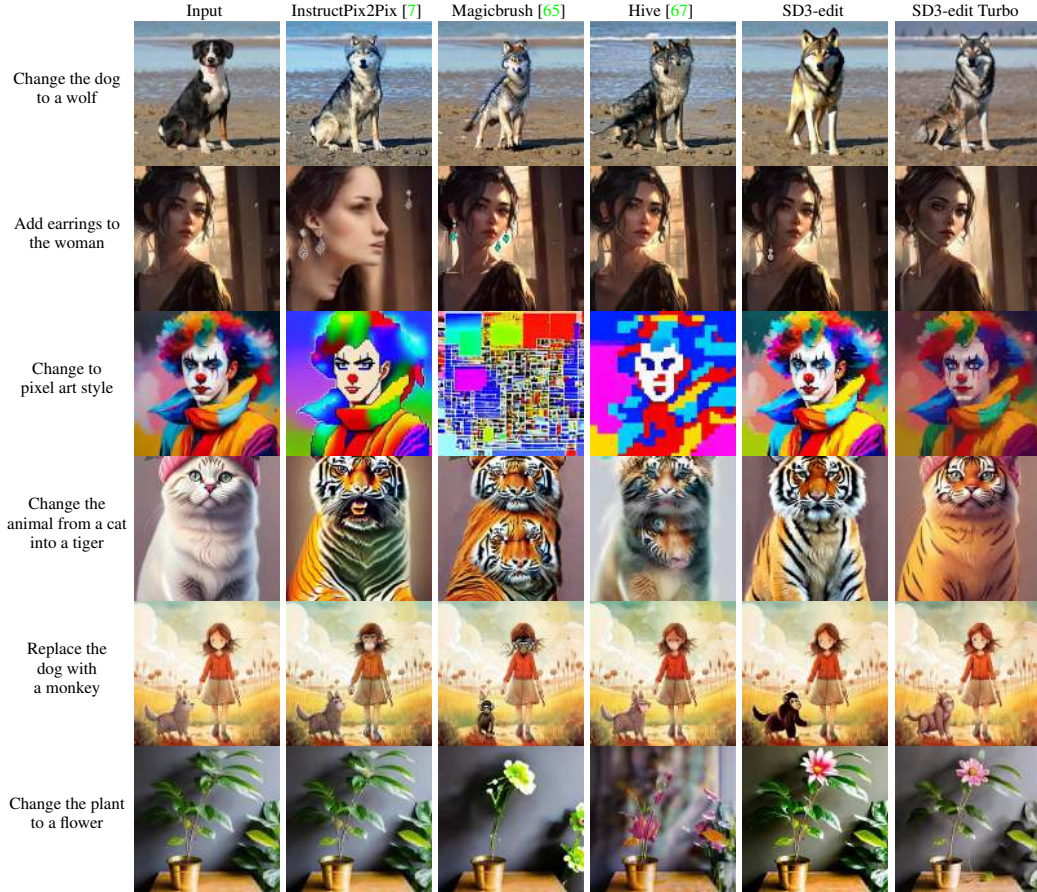
10

Figure 12: **Qualitative comparison for instruction-based editing.** For a given prompt and input image, we compare our distilled SD3-edit Turbo (1 step) to its teacher SD3-edit (50 steps) and several other baselines.

speed comes at the expense of controllability as the student does not allow to control the trade-off between image and text edit guidance strengths.

Image Inpainting For image inpainting, we condition on the masked input image for which we employ different masking strategies, ranging from narrow strokes, round cutouts and rectangular cutouts to outpainting masks. Furthermore, we always condition on the input image during training and inference, only omitting the text conditioning for the unconditional case. This configuration differs from that used in the editing task, where we employ the nested classifier-free guidance formulation. For distillation, we use the same LADD hyperparameters as for the editing model. Since we do not employ synthetic data for this task, we use an additional distillation loss to improve text-alignment. Our baselines are LaMa [59] and SD1.5-inpainting [2]. We sample LaMa and SD1.5-inpainting with the corresponding binary mask. SD3-inpainting is sampled for 50 steps with guidance strength 4, SD1.5 is sampled with the proposed default parameters, i.e., 50 steps, guidance scale 7.5.

Fig. 14 and Fig. 13 (Right) present qualitative and quantitative evaluations of the baselines and our model. Again, our distilled model performs on par with its teacher in a single step. LaMa beats all models on LPIPS, yet the high FID and qualitative comparisons show that LaMa lacks behind when large, non-homogeneous areas are masked.

---

[2]https://huggingface.co/runwayml/stable-diffusion-inpainting
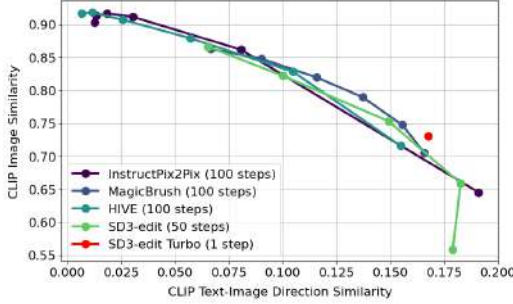
	FID ↓	LPIPS ↓
LaMa	27.21	**0.3137**
SD1.5-inpainting	10.29	0.3879
SD3-inpainting	**8.94**	0.3465
SD3-inpainting Turbo	9.44	0.3416

Figure 13: **Quantitative evaluation on image-to-image tasks.** Left: We plot CLIP Image Similarity measuring the fidelity to the input image over CLIP Direction Similarity measuring the fidelity to the edit prompt; higher is better for both metrics. We evaluate over varying image conditioning strengths on the PIE-Bench [23] dataset to compare SD3-edit Turbo and baselines. Right: Quantitative evaluation of image inpainting on COCO [29]; we report FID and LPIPS scores. The masks are created with different policies, ranging from narrow to wide masks and outpainting style masks.
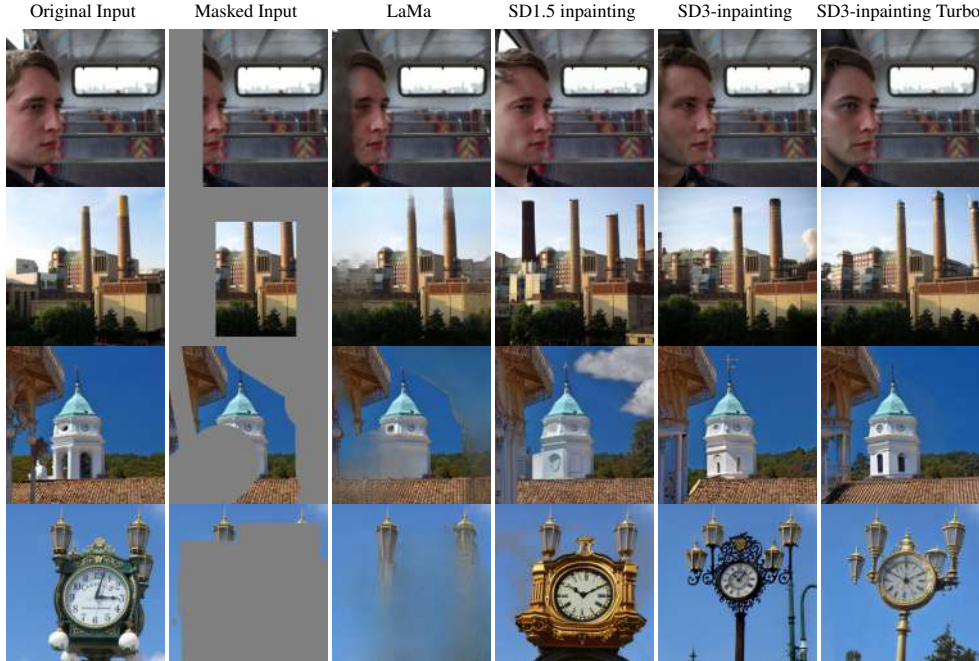

Figure 14: **Qualitative comparison for image inpainting editing.** For every masked input image, we compare our distilled SD3-edit inpainting Turbo (1 step) to its teacher SD3-inpainting (50 steps) and other baselines.

6 Limitations

In the human preference study detailed in Section 5.1, we demonstrate that while SD3 Turbo maintains the teacher's image quality within just four steps, it does so at the expense of prompt alignment. This trade-off introduces common text-to-image synthesis challenges such as object duplication and merging, fine-grained spatial prompting, and difficulties with negation. These issues, while not unique to our model, underscore a fundamental trade-off between model capacity, prompt alignment, and inference speed; exploring and quantifying this trade-off constitutes an exciting future research direction.

In our evaluation of image editing capabilities, we observe a lack of control due to the absence of adjustable image and text guidance strengths found in comparative methods [7]. A potential solution is deliberately adjusting these parameters during the training phase, coupled with model conditioning

A black dog sitting on a wooden chair. A white cat with black ears is standing up with its paws on the chair.

A set of 2x2 emoji icons with happy, angry, surprised and sobbing faces. The emoji icons look like dogs. All of the dogs are wearing blue turtlenecks.

a subway train with no cows in it.

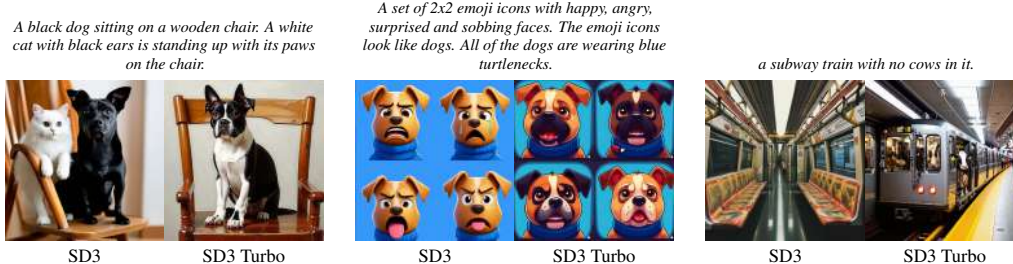| SD3 | SD3 Turbo | SD3 | SD3 Turbo | SD3 | SD3 Turbo |

Figure 15: **Failure cases.** While SD3-Turbo retains the image quality of its teacher, prompt alignment can suffer. Notably, we observe issues such as the merging of distinct entities, diminished accuracy in detailed spatial descriptions, and overlooked negations in prompts, though not universally across different random seeds.

on these parameters as proposed in [33]. Lastly, ins some cases the model exhibits rigidity, i.e., it adheres too closely to the input, rendering large changes challenging.

Acknowledgments

References

[1] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2, 5

[2] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, Y. Li, T. Michaeli, O. Wang, D. Sun, T. Dekel, and I. Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2

[3] D. Berthelot, A. Autef, J. Lin, D. A. Yap, S. Zhai, S. Hu, D. Zheng, W. Talbott, and E. Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023. 4

[4] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[5] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023. 2

[6] F. Boesel and R. Rombach. Improving image editing models with generative data refinement, 2024. to appear. 10

[7] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 10, 11, 12

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[9] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, M. Yu, A. Kadian, F. Radenovic, D. Mahajan, K. Li, Y. Zhao, V. Petrovic, M. K. Singh, S. Motwani, Y. Wen, Y. Song, R. Sumbaly, V. Ramanathan, Z. He, P. Vajda, and D. Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023. 2

[10] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[11] T. Dockhorn, A. Vahdat, and K. Kreis. Genie: Higher-order denoising diffusion solvers, 2022. 2

[12] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 2

[13] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2, 4, 5, 7, 9

[14] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2018. 5

[15] J. Heek, E. Hoogeboom, and T. Salimans. Multistep consistency models. *arXiv preprint arXiv:2403.06807*, 2024. 4

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 17

[17] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6, 10

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. 2

[19] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans. Imagen video: High definition video generation with diffusion models, 2022. 2

[20] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022. 2

[21] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 4

[22] P. Jaini, K. Clark, and R. Geirhos. Intriguing properties of generative classifiers. *ICLR*, 2023. 5

[23] X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 12

[24] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. 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. 2

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. 2

[26] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 4, 5

[27] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 17

[28] S. Lin, A. Wang, and X. Yang. Sdxl-lightning: Progressive adversarial diffusion distillation, 2024. 2, 4

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 12

[30] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 2

[31] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 2, 4

[32] G. Luo, L. Dunlap, D. H. Park, A. Holynski, and T. Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[33] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 4, 7, 13

[34] S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 2, 7

[35] C. Meng, R. Rombach, R. Gao, D. P. Kingma, S. Ermon, J. Ho, and T. Salimans. On distillation of guided diffusion models, 2023. 2, 4

[36] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5

[37] W. Peebles and S. Xie. Scalable diffusion models with transformers, 2023. 2, 7

[38] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 4, 9, 10, 17

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. 6, 10

[40] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv:2305.18290*, 2023. 8

[41] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[43] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 2022. 2, 6

[44] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models, 2022. 2, 4

[45] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans, 2016. 17

[46] A. Sauer, K. Chitta, J. Müller, and A. Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 5

[47] A. Sauer, K. Schwarz, and A. Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 6

[48] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023. 2, 6

[49] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 2, 4, 6

[50] J. Schmidhuber. Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991), 2020. 2

[51] N. Shaul, J. Perez, R. T. Q. Chen, A. Thabet, A. Pumarola, and Y. Lipman. Bespoke solvers for generative flow models, 2023. 2

[52] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023. 10

[53] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 2

[54] J. N. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015. URL https://api.semanticscholar.org/CorpusID:14888175. 2

[55] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2022. 2

[56] Y. Song and P. Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023. 4, 7

[57] Y. Song, J. N. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020. URL https://api.semanticscholar.org/CorpusID:227209335. 2, 4

[58] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *International conference on machine learning*, 2023. 2, 4, 7, 9

[59] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 11

[60] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 4

[61] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion Model Alignment Using Direct Preference Optimization. *arXiv:2311.12908*, 2023. 8

[62] Y. Xu, Y. Zhao, Z. Xiao, and T. Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans, 2023. 2

[63] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park. One-step diffusion with distribution matching distillation, 2023. 2

[64] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 6, 17

[65] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 11

[66] Q. Zhang and Y. Chen. Fast sampling of diffusion models with exponential integrator, 2023. 2

[67] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. *arXiv preprint arXiv:2303.09618*, 2023. 11

[68] J. Zheng, M. Hu, Z. Fan, C. Wang, C. Ding, D. Tao, and T.-J. Cham. Trajectory consistency distillation. *arXiv preprint arXiv:2402.19159*, 2024. 4

Appendix A   Details on Human Preference Assessment

Since standard performance metrics for generative models [45, 16] measure practically relevant quality image aspects as aesthetics and scene composition only insufficiently [27, 38] we use human evaluation to compare our model against the current state-of-the-art in text-to-image synthesis. Both for ours and the competing models, we generate samples based on prompts from the commonly used PartiPrompts benchmark [64]. To focus on more complex prompts, we omit the *Basic* category, which predominantly includes single-word prompts. We then selected every fourth prompt from the remaining categories, resulting in a total of 128 prompts, detailed in Sec. A.1. For each prompt, we generate four samples per model.

For the results presented in Figures 9 and 10 we compare our model with each competing one in a 1v1 comparison, where a human evaluator is presented samples from both models for the same text prompt and forced to pick one of those. To prevent biases, evaluators are restricted from participating in more than one of our studies. For the prompt following task, we show the respective prompt above the two images and ask, "Which image looks more representative of the text shown above and faithfully follows it?" When probing for visual quality we ask "Given the prompt above, which image is of higher quality and aesthetically more pleasing?". When comparing two models we choose the number of human evaluators such that each prompt is shown four times on average for each task

A.1   List of Prompts used for Human Evaluation

```
                              ┌─────────────────────┐
──────────────────────────────│ Selected Parti Prompts │──────────────────────────────
                              └─────────────────────┘
```

```
"A city in 4-dimensional space-time"
"Pneumonoultramicroscopicsilicovolcanoconiosis"
"A black dog sitting on a wooden chair.
A white cat with black ears is standing up with its paws on the chair."
"a cat patting a crystal ball with the number 7 written on it in black marker"
"a barred owl peeking out from dense tree branches"
"a cat sitting on a stairway railing"
"a cat drinking a pint of beer"
"a bat landing on a baseball bat"
"a black dog sitting between a bush and a pair of green pants standing up with nobody
inside them"
"a close-up of a blue dragonfly on a daffodil"
"A close-up of two beetles wearing karate uniforms and fighting, jumping over a
waterfall."
"a basketball game between a team of four cats and a team of three dogs"
"a bird standing on a stick"
"a cat jumping in the air"
"a cartoon of a bear birthday party"
"A cartoon tiger face"
"a dog wearing a baseball cap backwards and writing BONEZ on a chalkboard"
"a basketball to the left of two soccer balls on a gravel driveway"
"a photograph of a fiddle next to a basketball on a ping pong table"
"a black baseball hat with a flame decal on it"
"a doorknocker shaped like a lion's head"
"a coffee mug floating in the sky"
"a bench without any cats on it"
"a harp without any strings"
"long shards of a broken mirror reflecting the eyes of a great horned owl"
"view of a clock tower from above"
"a white flag with a red circle next to a solid blue flag"
"a bookshelf with ten books stacked vertically"
"a comic about two cats doing research"
"a black baseball hat"
"A castle made of cardboard."
"a footprint shaped like a peanut"
"a black t-shirt with the peace sign on it"
"a book with the words 'Don't Panic!' written on it"
"A raccoon wearing formal clothes, wearing a tophat and holding a cane.
The raccoon is holding a garbage bag. Oil painting in the style of abstract cubism."
```

"a young badger delicately sniffing a yellow rose, richly textured oil painting"
"A tornado made of bees crashing into a skyscraper. painting in the style of Hokusai."
"a drawing of a house on a mountain"
"a dutch baroque painting of a horse in a field of flowers"
"a glass of orange juice to the right of a plate with buttered toast on it"
"a bloody mary cocktail next to a napkin"
"a can of Spam on an elegant plate"
"A bowl of soup that looks like a monster made out of plasticine"
"A castle made of tortilla chips, in a river made of salsa.
There are tiny burritos walking around the castle"
"a close-up of a bloody mary cocktail"
"a close-up of an old-fashioned cocktail"
"three green peppers"
"A bowl of Beef Pho"
"a painting of the food of china"
"Two cups of coffee, one with latte art of a heart. The other has latte art of stars."
"Two cups of coffee, one with latte art of yin yang symbol. The other has latte art of
a heart."
"A set of 2x2 emoji icons with happy, angry, surprised and sobbing faces.
The emoji icons look like dogs. All of the dogs are wearing blue turtlenecks."
"a kids' book cover with an illustration of white dog driving a red pickup truck"
"a drawing of a series of musical notes wrapped around the Earth"
"a triangle with a smiling face"
"a black background with a large yellow circle and a small red square"
"A green heart with shadow"
"three small yellow boxes"
"A green heart"
"A heart made of water"
"a cute illustration of a horned owl with a graduation cap and diploma"
"a flag with a dinosaur on it"
"G I G G L E painted in thick colorful lettering as graffiti on a faded red brick wall
with a
splotch of exploding white paint."
"A high resolution photo of a donkey in a clown costume giving a lecture at the front
of a lecture hall.
The blackboard has mathematical equations on it. There are many students in the
lecture hall."
"An empty fireplace with a television above it. The TV shows a lion hugging a giraffe."
"a ceiling fan with an ornate light fixture"
"a small kitchen with a white goat in it"
"a ceiling fan with five brown blades"
"two pianos next to each other"
"a kitchen with a large refrigerator"
"A glass of red wine tipped over on a couch, with a stain that writes "OOPS" on the
couch."
"the words 'KEEP OFF THE GRASS' written on a brick wall"
"a white bird in front of a dinosaur standing by some trees"
"a beach with a cruise ship passing by"
"a chemtrail passing between two clouds"
"a grand piano next to the net of a tennis court"
"a peaceful lakeside landscape with migrating herd of sauropods"
"a marina without any boats in it"
"a view of the Earth from the moon"
"an aerial photo of a sandy island in the ocean"
"a tennis court with three yellow cones on it"
"a basketball hoop with a large blue ball stuck in it"
"a horse in a field of flowers"
"a cloud in the shape of a elephant"
"a painting of a white country home with a wrap-around porch"
"a store front with 'Grassy Meadow' written on it"
"a black dog jumping up to hug a woman wearing a red sweater"
"a grandmother reading a book to her grandson and granddaughter"
"a child eating a birthday cake near some palm trees"
"a cricket team walking on to the pitch"
"a man riding a cat"

"A boy holds and a girl paints a piece of wood."
"A man gives a woman a laptop and a boy a book."
"a three quarters view of a man getting into a car"
"two people facing the viewer"
"a family of four posing at the Grand Canyon"
"a boy going to school"
"a father and a son playing tennis"
"a cartoon of a man standing under a tree"
"a comic about a family on a road trip"
"a baby daikon radish in a tutu walking a dog"
"a plant growing on the side of a brick wall"
"a flower with a cat's face in the middle"
"a smiling banana wearing a bandana"
"several lily pads without frogs"
"two red flowers and three white flowers"
"a flower with large yellow petals"
"A photo of a four-leaf clover made of water."
"A photo of a palm tree made of water."
"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 blue airplane taxiing on a runway with the sun behind it"
"a car with tires that have yellow rims"
"a bike rack with some bike locks attached to it but no bicycles"
"a friendly car"
"a subway train with no cows in it"
"an overhead view of a pickup truck with boxes in its flatbed"
"Three-quarters front view of a blue 1977 Ford F-150 coming around a curve
in a mountain road and looking over a green valley on a cloudy day."
"two motorcycles facing each other"
"A green train is coming down the tracks"
"a cardboard spaceship"
"a crayon drawing of a space elevator"
"a hot air balloon with chameleon logo. the sun is shining and puffy white clouds are
in the background."
"A close-up high-contrast photo of Sydney Opera House sitting next to Eiffel tower,
under a blue night sky of roiling energy, exploding yellow stars, and radiating swirls
of blue"
"a painting of the mona lisa on a white wall"
"A blue Porsche 356 parked in front of a yellow brick wall"
"a diplodocus standing in front of the Millennium Wheel"
"a herd of buffalo stampeding at the Kremlin"
"A smiling sloth is wearing a leather jacket, a cowboy hat, a kilt and a bowtie. The
sloth is holding a quarterstaff and a big book. The sloth is standing on grass a few
feet in front of a shiny VW van with flowers painted on it. wide-angle lens from below."