Distilling Diffusion Models into Conditional GANs

Minguk Kang[1,2], Richard Zhang[2], Connelly Barnes[2],
Sylvain Paris[2], Suha Kwak[1], Jaesik Park[3],
Eli Shechtman[2], Jun-Yan Zhu[4], and Taesung Park[2]

Pohang University of Science and Technology[1] Adobe Research[2]
Seoul National University[3] Carnegie Mellon University[4]

Abstract. We propose a method to distill a complex multistep diffusion model into a single-step conditional GAN student model, dramatically accelerating inference, while preserving image quality. Our approach interprets diffusion distillation as a *paired image-to-image translation* task, using noise-to-image pairs of the diffusion model's ODE trajectory. For efficient regression loss computation, we propose E-LatentLPIPS, a perceptual loss operating directly in diffusion model's latent space, utilizing an ensemble of augmentations. Furthermore, we adapt a diffusion model to construct a multi-scale discriminator with a text alignment loss to build an effective conditional GAN-based formulation. E-LatentLPIPS converges more efficiently than many existing distillation methods, even accounting for dataset construction costs. We demonstrate that our one-step generator outperforms cutting-edge one-step diffusion distillation models – SDXL-Turbo and SDXL-Lightning – on the COCO benchmark.

1 Introduction

Diffusion models [23, 81, 86] have demonstrated unprecedented image synthesis quality on challenging datasets, such as LAION [79]. However, producing high-quality results requires dozens or hundreds of sampling steps. As a result, most existing diffusion-based image generation models, such as DALL·E 2 [65], Imagen [73], and Stable Diffusion [69], suffer from high latency, often exceeding 10 seconds and hindering real-time interaction. If our model only requires *one* inference step, it will not only improve the user experience in text-to-image synthesis, but also expand its potential in 3D and video applications [22, 64]. But how can we build a one-step text-to-image model?

One simple solution is to just train a one-step model from scratch. For example, we can train a GAN [15], a leading one-step model for simple domains [33]. Unfortunately, training text-to-image GANs on large-scale and diverse datasets is still challenging, despite recent advances [30, 76]. The challenge lies in GANs needing to tackle *two* difficult tasks all at once without any supervision: (1) finding correspondence between noises and natural images, and (2) effectively optimizing a generator model to perform the mapping from noises to images.

This "unpaired" learning is often considered more ill-posed, as mentioned in CycleGAN [101], compared to paired learning, where conditional GANs [27] can learn to map the input to output, given ground truth correspondences.

Our key idea is to tackle the above tasks one by one. We first find the correspondence between noises and images by simulating the ODE solver with a pre-trained diffusion model. Given the established corresponding pairs, we then ask a conditional GAN to map noises to images in a paired image-to-image translation framework [27,61]. This disentangled approach allows us to leverage two types of generative models for separate tasks, achieving the benefits of both: finding high-quality correspondence using diffusion models, while achieving fast mapping using conditional GANs.

In this work, we collect a large number of noise-to-image pairs from a pre-trained diffusion model and treat the task as a paired image-to-image translation problem [27], enabling us to exploit tools such as perceptual losses [11, 28, 97] and conditional GANs [15,27,56]. In doing so, we make a somewhat unexpected discovery. Collecting a large database of noise-image pairs and training with a regression loss without the GAN loss can already achieve comparable results to more recent distillation methods [54,84], at a significantly lower compute budget, if the regression loss is designed carefully.

First, in regression tasks, using perceptual losses (such as LPIPS [97]) better preserves perceptually important details over point-based losses (such as L2). However, perceptual losses are fundamentally incompatible with Latent Diffusion Models [69], as they require an expensive decoding from latent to pixel space. To overcome this, we propose LatentLPIPS, showing that perceptual losses can directly work in latent space. This enables a fourfold increase in batch size, compared to computing LPIPS in pixel space. Unfortunately, we observe that the latent-based perceptual loss has more blind spots than its pixel counterparts. While previous work has found that ensembling is helpful for pixel-based LPIPS [35], we find that it is critical for the latent-based version. Working in latent space with our Ensembled-LatentLPIPS, we demonstrate strong performance with just a regression loss, comparable to guided progressive distillation [54]. Additionally, we employ a discriminator in the training to further improve performance. We develop a multi-scale conditional diffusion discriminator, leveraging the pre-trained weights and using our new single-sample R1 loss and mix-and-match augmentation. We name our distillation model *Diffusion2GAN*.

Using the proposed Diffusion2GAN framework, we distill Stable Diffusion 1.5 [69] into a single-step conditional GAN model. Our Diffusion2GAN can learn noise-to-image correspondences inherent in the target diffusion model better than other distillation methods. It also outperforms recently proposed distillation models, UFOGen [93] and DMD [94], on the zero-shot one-step COCO2014 [45] benchmark. Furthermore, we perform extensive ablation studies and highlight the critical roles of both E-LatentLPIPS and multi-scale diffusion discriminator. Beyond the distillation of Stable Diffusion 1.5, we demonstrate the effectiveness of Diffusion2GAN in distilling a larger SDXL [63], exhibiting superior FID [20] and CLIP-score [19] over one-step SDXL-Turbo [77] and SDXL-Lightning [44].

Fig. 1: Visual comparison to SDXL teacher [63] with a classifier-free guidance scale [24] of 7 and selected distillation student models, including SDXL-Turbo [77], SDXL-Lightning [44], and our SDXL-Diffusion2GAN. All images in a given row were generated using the same noise input, except for SDXL-Turbo, which requires a distinct noise size of $4 \times 64 \times 64$. Compared to other distillation models, our SDXL-Diffusion2GAN more closely adheres to the original ODE trajectory. Please visit our website for more results.

2 Related Work

Diffusion models. Diffusion models (DMs) [23, 81, 86] are a family of generative models consisting of the diffusion process and denoising process. The diffusion process progressively diffuses high-dimensional data from data distribution to easy-to-sample Gaussian distribution, while the denoising process aims to reverse the process using a deep neural network trained on a score-matching objective [85–87]. Once trained, these models can generate data from random Gaussian noise, using numerical integrators [1, 2, 31]. Diffusion models have enabled numerous vision and graphics applications, such as image editing [6, 18, 53], controllable image synthesis [58, 96], personalized generation [13, 39, 72], video synthesis [4, 17, 22], and 3D content creation [43, 64]. However, the sampling typically requires tens of sampling steps, leading to slower image generation speed than other generative models, such as GANs [15] and VAEs [37]. In this work, our goal is to accelerate the model's inference while maintaining image quality.

Diffusion distillation. Accelerating the sampling speed of diffusion models is crucial for enhancing practical applications, as well as reducing energy costs for inference. Several works have proposed reducing the number of sampling steps using fast ODE solvers [31, 49, 50] or reducing the computational time per step [8, 41, 42]. Another effective method for acceleration is knowledge distillation [3, 16, 21, 47, 48, 51, 54, 74, 84, 91, 100]. In this approach, multiple steps of a teacher diffusion model is distilled into a fewer-step student model. Luhman *et al.* [51] propose L_p regression training between student's output from a Gaussian noise $\mathbf{x}_T$ and its corresponding ODE solution $\mathbf{x}_0$. Despite its simplicity, such direct regression produces blurry outputs and does not match the image synthesis capabilities exhibited by other generative models. To enhance image quality, InstaFlow [48] straightens high-curvature ODE trajectory via ReFlow [47] and distills the linearized ODE trajectory to the student model. Consistency Distillation (CD) [52, 84] trains a student model to predict a consistent output for any noisy sample $\mathbf{x}_{t+1}$ and its single-step denoising $\mathbf{x}_t$. Recently, several studies have proposed using a GAN discriminator to enhance distillation performance. For example, CTM [36] and SDXL-Turbo [77] utilize an improved StyleGAN [76, 78] discriminator to train a one-step image generator. In addition, UFOGen [93], SDXL-Lightning [44], and LADD [75] adopt a pre-trained diffusion model as a strong discriminator, demonstrating their abilities in one-step text-to-image synthesis. Although these works are concurrent, we will compare our method with SDXL-Turbo and SDXL-Lightning for a more comprehensive comparison.

Conditional Generative Adversarial Networks [27, 56] have been a commonly-used framework for conditional image synthesis. The condition could be an image [9, 27, 46, 60, 68, 101], class-label [5, 29, 34, 57, 59], and text [30, 67, 76, 92, 95]. In particular, cGANs have shown impressive performance when helped by a regression loss to stabilize training, as in image translation [27, 61, 62, 88, 98, 101]. Likewise, we approach diffusion model distillation by employing the image-conditional GAN, along with a perceptual regression loss [97]. Early works [90, 91] combine GANs with the forward diffusion process, but do not aim at distilling a pre-trained diffusion model into a GAN.

3 Method

Our goal is to distill a pre-trained text-to-image diffusion model into a one-step generator. That is, we want to learn a mapping $\mathbf{x} = G(\mathbf{z}, \mathbf{c})$, where the one-step generator network G takes input noise $\mathbf{z}$ and text $\mathbf{c}$, and maps them to the output image $\mathbf{x}$ produced by the diffusion model. We assume that the student and teacher share the same architecture, so that we can initialize the student model G using weights of the teacher model. For our method section, we assume Latent Diffusion Models [69] with $\mathbf{x}, \mathbf{z} \in \mathbb{R}^{4\times64\times64}$. Later, we also adopt our method to the SDXL model [63].

In the rest of this section, we will elaborate on the design and training principles of our framework. We begin by describing distillation as a paired image-to-image translation problem in Section 3.1. Then, we introduce our Ensembled Latent LPIPS regression loss (E-LatentLPIPS) in Section 3.2. Just using this regression loss improves training efficiency and significantly improves distillation performance for latent diffusion models. Lastly, we present an improved discriminator design that reuses a pre-trained diffusion model (Section 3.3). It is worth noting that our findings extend beyond the specific type of latent space diffusion models [69, 70] and apply to a pixel space model [31] as well.

3.1 Paired Noise-to-Image Translation for One-step Generation

With the emergence of diffusion probabilistic models [23, 86], Luhman *et al.* [51] suggest that the multi-step denoising process of a pre-trained diffusion model can be reduced to a single step by minimizing the following distillation objective:

$$\mathcal{L}^{\text{ODE}}_{\text{distill}} = \mathbb{E}_{\{\mathbf{z},\mathbf{c},\mathbf{x}\}}\Big[d(G(\mathbf{z}, \mathbf{c}), \mathbf{x})\Big], \tag{1}$$

where $\mathbf{z}$ is a sample from Gaussian noise, $\mathbf{c}$ is a text prompt, G denotes a UNet generator with trainable weights, $\mathbf{x}$ is the output of the diffusion model simulating the ordinary differential equation (ODE) trajectory with the DDIM sampler [82], and $d(\cdot, \cdot)$ is a distance metric. Due to the computational cost of obtaining $\mathbf{x}$ for each iteration, the method uses pre-computed pairs of noise and corresponding ODE solutions before training begins. During training, it randomly samples noise-image pairs and minimizes the ODE distillation loss (Equation 1). While the proposed approach looks simple and straightforward, the direct distillation approach yields inferior image synthesis results compared to more recent distillation methods [48, 54, 74, 84].

In our work, we hypothesize that the full potential of direct distillation has not yet been realized. In our experiments on CIFAR10, we observe that we can significantly improve the quality of distillation by (1) scaling up the size of the ODE pair dataset and (2) using a perceptual loss [97] (as opposed to the pixel-space L2 loss in Luhman *et al.*). In Table 6, we show the training progression on the CIFAR10 dataset, and compare its performance to Consistency Model [84]. Surprisingly, the direct distillation with the LPIPS loss can achieve lower FID than the Consistency Model at smaller amount of total compute, even accounting for the extra compute to collect the ODE pairs.

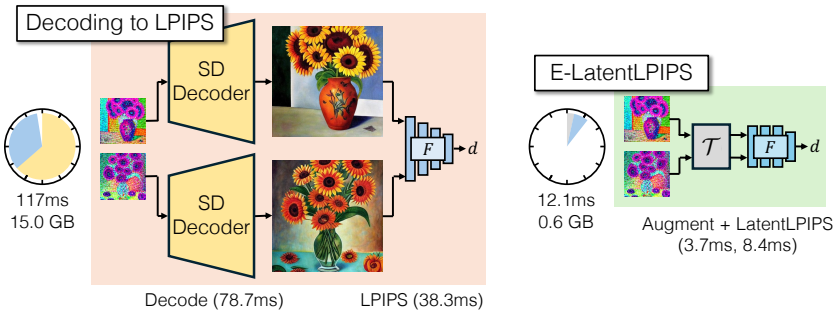

Fig. 2: E-LatentLPIPS for latent space distillation. Training a single iteration with LPIPS [97] takes 117ms and 15.0GB extra memory on NVIDIA A100, whereas our E-LatentLPIPS requires 12.1ms and 0.6GB on the same device. Consequently, E-latentLPIPS accelerates the perceptual loss computation time by 9.7× compared to LPIPS, while simultaneously reducing memory consumption.

3.2 Ensembled-LatentLPIPS for Latent Space Distillation

The original LPIPS [97] observes that the features from a pretrained classifier can be calibrated well enough to match human perceptual responses. Moreover, LPIPS serves as an effective regression loss across many image translation applications [61,89]. However, LPIPS, built to be used in the *pixel* space, is unwieldy to use with a *latent* diffusion model [69]. As shown in Figure 2, the latent codes must be decoded into the pixel space (e.g., 64 → 512 resolution) before computing LPIPS with a feature extractor F and a distance metric ℓ.

$$d_{\text{LPIPS}}(\mathbf{x}_0, \mathbf{x}_1) = \ell\big(F(\text{Decode}^{8\times}(\mathbf{x}_0)), F(\text{Decode}^{8\times}(\mathbf{x}_1))\big) \tag{2}$$

This defeats the primary motivator of LDMs, to operate in a more efficient latent space. As such, can we bypass the need to decode to pixels, and directly compute a perceptual distance in latent space?

Learning LatentLPIPS. We hypothesize that the same perceptual properties of LPIPS can hold for a function directly computed on latent space. Following the procedure from Zhang *et al.* [97], we first train a VGG network [80] on ImageNet, but in the latent space of Stable Diffusion. We slightly modify the architecture by removing the 3 max-pooling layers, as the latent space is already 8× downsampled, and change the input to 4 channels. We then linearly calibrate intermediate features using the BAPPS dataset [97]. This successfully yields a function that operates in the latent space: $d_{\text{LatentLPIPS}}(\mathbf{x}_0, \mathbf{x}_1) = \ell(F(\mathbf{x}_0), F(\mathbf{x}_1))$.

Interestingly, we observe that while ImageNet classification accuracy in latent space is slightly lower on latent codes than on pixels, the perceptual agreement is retained. This indicates that while compression to latent space destroys some of the low-level information that helps with classification [26], it keeps the perceptually relevant details of the image, which we can readily exploit. Additional details are in the Supplement C.

(a) Target (b) LPIPS (LPIPS = 0.016) (c) LatentLPIPS (0.490) (d) E-LatentLPIPS (0.001)

Fig. 3: Single image reconstruction. To gain insight into the loss landscape of our regression loss, we conduct an image reconstruction experiment by directly optimizing a single latent with different loss functions. Reconstruction with LPIPS roughly reproduces the target image, but at the cost of needing to decode into pixels. LatentLPIPS alone cannot precisely reconstruct the image. However, our ensembled augmentation, E-LatentLPIPS, can more precisely reconstruct the target while operating directly in the latent space.

Ensembling. We observe that the straightforward application of LatentLPIPS as the new loss function for distillation results in producing wavy, patchy artifacts. We further investigate this in a simple optimization setup, as shown in Figure 3, by optimizing a randomly-sampled latent code towards a single target image. Here we aim to recover the target latent using different loss functions: $\arg\min_{\hat{\mathbf{x}}} d(\hat{\mathbf{x}}, \mathbf{x})$, where $\mathbf{x}$ is the target latent, $\hat{\mathbf{x}}$ the reconstructed latent, and d either the original LPIPS or LatentLPIPS. We observe that the single image reconstruction does not converge under LatentLPIPS (Figure 3 (c)). We hypothesize this limitation is due to a suboptimal loss landscape formed by the latent version of the VGG network.

Inspired by E-LPIPS [35], we apply random differentiable augmentations [32, 99], general geometric transformations [32], and cutout [10], to both generated and target latents. At each iteration, a random augmentation is applied to both generated and target latents. When applied to single image optimization, the ensemble strategy nearly perfectly reconstructs the target image, as shown in Figure 3 (d). The new loss is named Ensembled-LatentLPIPS, or E-LatentLPIPS for short.

$$d_{\text{E-LatentLPIPS}}(\mathbf{x}_0, \mathbf{x}_1) = \mathbb{E}_{\mathcal{T}}\Big[\ell\big(F(\mathcal{T}(\mathbf{x}_0)), F(\mathcal{T}(\mathbf{x}_1))\big)\Big], \tag{3}$$

where $\mathcal{T}$ is a randomly sampled augmentation. Applying the loss function to ODE distillation:

$$\mathcal{L}_{\text{E-LatentLPIPS}}\big(G, \mathbf{z}, \mathbf{c}, \mathbf{x}\big) = d_{\text{E-LatentLPIPS}}(G(\mathbf{z}, \mathbf{c}), \mathbf{x}), \tag{4}$$

where $\mathbf{z}$ denotes a Gaussian noise, and $\mathbf{x}$ denotes its target latent. As illustrated in Figure 2 (right), compared to its LPIPS counterpart, the computation time is significantly lower, due to (1) not needing to decode to pixels (saving 79 ms

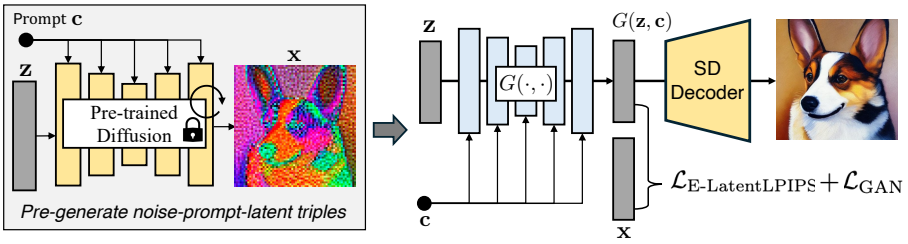

Fig. 4: Our Diffusion2GAN for one-step image synthesis. First, we collect diffusion model output latents along with the input noises and prompts. Second, the generator is trained to map noise and prompt to the target latent using the E-LatentLPIPS regression loss (Equation 4) and the GAN loss (Equation 6). While the output of the generator can be decoded by the SD latent decoder into RGB pixels, it is a compute intensive operation that is never performed during training.

for one image on an A100) and (2) (Latent)LPIPS itself operating at a lower-resolution latent code than in pixel space (38→8 ms). While augmentation takes some time (4 ms), in total, perceptual loss computation is almost 10× cheaper (117→ 12 ms) with our E-LatentLPIPS. In addition, memory consumption is greatly reduced (15→0.6 GB).

Experimental results of Table 1 demonstrate that learning the ODE mapping with E-LatentLPIPS leads to better convergence, exhibiting lower FID compared to other metrics such as MSE, Pseudo Huber loss [25,83], and the original LPIPS loss. For additional details regarding the toy reconstruction experiment and differentiable augmentations, please refer to the Supplement C.

3.3 Conditional Diffusion Discriminator

In Sections 3.1 and 3.2, we have elucidated that diffusion distillation can be achieved by framing it as a paired noise-to-latent translation task. Motivated by the effectiveness of conditional GAN for paired image-to-image translation [27], we employ a conditional discriminator. The conditions for this discriminator include not only the text description $\mathbf{c}$ but also the Gaussian noise $\mathbf{z}$ provided to the generator. Our new discriminator incorporates the aforementioned conditioning while leveraging the pre-trained diffusion weights. Formally, we optimize the following minimax objective for the generator G and discriminator D:

$$\min_G \max_D \; \mathbb{E}_{\mathbf{c},\mathbf{z},\mathbf{x}}[\log(D(\mathbf{c},\mathbf{z},\mathbf{x}))] + \mathbb{E}_{\mathbf{c},\mathbf{z}}[\log(1 - D(\mathbf{c},\mathbf{z},G(\mathbf{z},\mathbf{c})))]. \tag{5}$$

For the generator, we minimize the following non-saturating GAN loss [14].

$$\mathcal{L}_{\text{GAN}}(G,\mathbf{z},\mathbf{c},\mathbf{x}) = -\mathbb{E}_{\mathbf{c},\mathbf{z}}\big[\log(D(\mathbf{c},\mathbf{z},G(\mathbf{z},\mathbf{c})))\big]. \tag{6}$$

The final loss for the generator is $\mathcal{L}_G = \mathcal{L}_{\text{E-LatentLPIPS}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}$. We provide more details on the discriminator and loss functions.

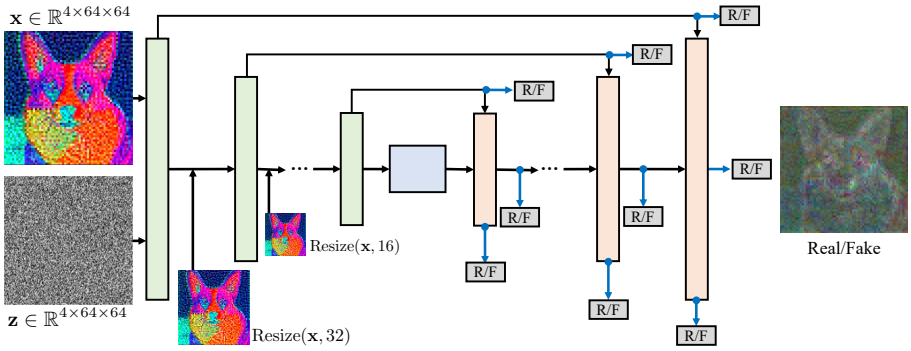

Fig. 5: Our multi-scale conditional discriminator design. We reuse the pre-trained weights from the teacher model's U-Net and augment it with multi-scale input and output branches. Concretely, we feed the resized version of input latents to each downsampling block of the encoder. For the decoder part, we enforce the discriminator to make real/fake predictions at three places at each scale: before, at, and after the skip connection. This multi-scale adversarial training further improves image quality.

Initialization from a pre-trained diffusion model. We demonstrate that initializing the discriminator weights with a pre-trained diffusion model is effective for diffusion distillation. Compared to the implementation of GigaGAN discriminator [30], using a pre-trained Stable Diffusion 1.5 U-Net [71] and fine-tuning the model as the discriminator in the latent space results in superior FID in Table 2. The adversarial loss is computed independently at each location of the U-Net discriminator output. Note that the original U-Net architecture conditions on text but not on the input noise map $\mathbf{z}$. We further modify the discriminator architecture to support $\mathbf{z}$ conditioning, simply by adding the input with $\mathbf{z}$ processed through a single convolution layer with zero initialization in the channel dimension. Note that the text conditioning for the diffusion discriminator is naturally carried out by the built-in cross-attention layers in the Stable Diffusion U-Net. We observe moderate improvement across all metrics.

Single-sample R1 regularization. While the conditional U-Net discriminator from pre-trained diffusion weights already achieves competitive results on the zero-shot COCO2014 [45] benchmark, we have noticed considerable training variance across different runs, likely due to the absence of regularization and unbounded gradients from the discriminator. To mitigate this, we introduce R1 regularization [55] on each mini-batch for training the diffusion discriminator. However, introducing R1 regularization increases GPU memory consumption, posing a practical challenge, especially when the discriminator is a high-capacity U-Net. To minimize memory consumption and accelerate training, we not only adopt lazy regularization [34] with an interval of 16, but also apply R1 regularization only to a single sample of each mini-batch. In addition to improved stability, we also observe that the single-sample R1 regularization results in better convergence, as shown in Table 2.

Multi-scale in-and-out U-Net discriminator. GigaGAN [30] observes that the GAN discriminator tends to focus on a particular frequency band, often overlooking high-level structures, and introduces a multi-scale discriminator to address this issue. Similarly, we propose a new U-Net discriminator design, as shown in Figure 5, which enforces independent real/fake prediction at various segments of the U-Net. Specifically, we modify the U-Net encoder to receive resized inputs at each downsampling layer and attach three readout layers at each scale of the U-Net decoder to make independent real/fake predictions, from the U-Net skip connection features, the upsampled features from the U-Net bottleneck, and the combined features. At a high level, the new design enforces that all U-Net layers participate in the final prediction, ranging from shallow skip connections to deep middle blocks. This design enhances low-frequency structural consistency and significantly increases FIDs, as observed in Table 2.

Mix-and-match augmentation. To further encourage the discriminator to focus on text alignment and noise conditioning, we introduce mix-and-match augmentation for discriminator training, similar to GigaGAN [30] and earlier text-to-image GAN works [66, 95]. During discriminator training, we replace a portion of the generated latents with random, unrelated latents from the target dataset while maintaining the other conditions unchanged. This categorizes the replaced latents as fake, since the alignments between the latent and its paired noise and text are incorrect, thereby fostering improved alignments. Additionally, we make substitutions to text and noise, contributing to the overall enhancement of the conditional diffusion discriminator.

4 Experiments

Here, we study the effectiveness of our algorithmic designs with a systematic ablation study in Section 4.1. Next, we compare our method with leading one-step generators using a standard benchmark regarding image quality, text alignment, and inference speed in Section 4.2. We then present human preference evaluation results in Section 4.3. Additionally, we report the training speed (Section 4.4).

Training details. We distill Stable Diffusion 1.5 into our one-step generator and train the model on two ODE datasets with different classifier-free guidance (CFG), namely, the SD-CFG-3 dataset with 3 million noise-latent pairs and the SD-CFG-8 dataset with 12 million pairs. We use the prompts from the LAION-aesthetic-6.25 and -6.0 datasets to create the SD-CFG-3 and SD-CFG-8 datasets, respectively, and simulate the ODE using 50 steps of DDIM [82]. To demonstrate the effectiveness of Diffusion2GAN for a larger text-to-image model, we distill SDXL-Base-1.0 [63] into Diffusion2GAN using 8 million noise-latent pairs named SDXL-CFG-7 dataset. These pairs were generated by SDXL-Base-1.0 using prompts from the LAION-aesthetic-6.0 dataset. We simulate the ODE of SDXL-Base-1.0 using 50 steps of DDIM. For further details on hyperparameters and evaluation details, please refer to the Supplement A and B. Notably, we did not use any real images from the LAION dataset.

Table 1: Ablation study on SD-CFG-3 dataset. We distill Stable Diffusion 1.5 [71] into one-step generators using ODE distillation loss (Equation 1). All models are trained with a batch size of 256 for 20k iterations using 8 A100-80GB GPUs, except for the LPIPS model and the larger batch-size model. For the LPIPS model, we use a batch size of 64 and accumulate gradients four times due to its 62GB GPU memory consumption per A100-80GB. The other models require nearly 68GB per GPU for 256 batch training. Our E-LatentLPIPS achieves stronger performance than traditional LPIPS without decoding to pixels.

Method (Loss function)	Loss space	img/sec (↑)	Batch size	FID (↓)	CLIP (↑)	Pre. (↑)	Rec. (↑)
ODE distillation (LPIPS [97])	Pixel	40.0	256	25.94	0.288	0.60	0.53
ODE distillation (MSE)	Latent	138.4	256	110.55	0.222	0.21	0.33
ODE distillation (Pseudo Huber [83])	Latent	144.2	256	87.60	0.230	0.29	0.40
ODE distillation (LatentLPIPS)	Latent	139.9	256	67.17	0.244	0.46	0.54
ODE distillation (E-LatentLPIPS)	Latent	127.5	256	**22.95**	**0.299**	**0.62**	**0.58**
→ larger batch-size (8× more GPUs)	Latent	128.0	2048	**14.72**	0.292	**0.66**	**0.65**

Table 2: Ablation study on SD-CFG-3 dataset. All models are initialized with the weights of a pre-trained ODE distillation model targeting Stable Diffusion 1.5 [71] and trained with a batch size of 256 using 16 A100-80GB GPUs. Each proposed component plays a crucial role in improving both FID [20] and CLIP-score [19].

Method	FID-30k (↓)	CLIP-30k (↑)	Precision (↑)	Recall (↑)
ODE distillation (E-LatentLPIPS)	14.72	0.292	0.66	0.65
+ GigaGAN D [30]	13.97	0.293	0.68	0.64
ODE distillation (E-LatentLPIPS)	14.72	0.292	0.66	0.65
+ Diffusion D	12.04	0.300	0.70	0.65
+ **z** conditional D	11.97	0.302	0.70	0.65
+ Single-sample R1	10.60	0.303	0.73	0.65
+ Multi-scale training	9.58	0.308	0.72	**0.66**
+ Mix-and-match augmentation	**9.45**	**0.310**	**0.73**	0.65
Stable Diffusion 1.5 [71] (Teacher)	8.74	0.312	0.72	0.67

4.1 Effectiveness of Each Component

In Table 1, we conduct an ablation study on the choice of distance metric for ODE distillation training. We consider L1, Pseudo Huber [83], LPIPS, LatentLPIPS, and our E-LatentLPIPS metrics. As shown in Table 1, ODE distillation using MSE [51] achieves worse results on large-scale text-to-image datasets. Also, introducing the Pseudo Huber metric improves FID significantly [84], but it remains insufficient. However, if we apply a perceptual loss, such as pixel space LPIPS and latent space E-LatentLPIPS, the ODE distillation presents FID near 20∼25, even trained using a small batch size. This suggests that the noise-to-image translation task holds promise, and it would give better results once we introduce a conditional discriminator to further improve the image quality.

Table 2 presents the ablation study regarding each component of Diffusion2GAN's discriminator. All generators are initialized with the pre-trained weights of the best performing ODE distilled generator shown in Table 1. We

Table 3: Comparison to recent text-to-image models on COCO2014. We distill Stable Diffusion 1.5 [71] into Diffusion2GAN on the SD-CFG-3 dataset with a batch size of 1024 using 64 A100-80GB GPUs. Diffusion2GAN significantly outperforms the leading one-step diffusion distillation generators.

Multi-step generator	Type	# Param.	FID-30k (↓)	Inference time (s)
Stable Diffusion 1.5 [70]	Diffusion	0.9B	8.74	2.59
PIXART-α [7]	Diffusion	0.6B	10.65	-
One-step generator	**Type**	**# Param.**	**FID-30k (↓)**	**Inference time (s)**
GigaGAN [30]	GAN	1.0B	**9.09**	0.13
InstaFlow-0.9B [48]	Distillation	0.9B	13.10	0.09
UFOGen [93]	Distillation	0.9B	12.78	0.09
DMD [94]	Distillation	0.9B	11.49	0.09
Diffusion2GAN	Distillation	0.9B	9.29	0.09

Table 4: Comparison to recent text-to-image models on COCO2017. On the SD-CFG-3, Diffusion2GAN, distilled from Stable Diffusion 1.5 [71], demonstrates better performance over UFOGen [93]. While Diffusion2GAN presents slightly better FID [20] than ADD-M [77], it exhibits a lower CLIP-score [19].

Model	# Step	FID-5k (↓)	CLIP-5k (↑)	Inference time (s)
DPM solver [49]	25	20.1	0.318	0.88
Progressive distillation [54]	4	26.4	0.300	0.21
InstaFlow-0.9B [48]	1	23.4	0.304	**0.09**
UFOGen [93]	1	22.5	0.311	**0.09**
ADD-M [77]	1	19.7	**0.326**	**0.09**
Diffusion2GAN	1	**19.5**	0.311	**0.09**
Stable Diffusion 1.5 [71] (Teacher)	50	19.1	0.313	2.59

compare our diffusion-based discriminator to the state-of-the-art GigaGAN discriminator [30]. As shown in Table 2, each component of Diffusion2GAN plays a crucial role in enhancing FID and CLIP-score.

4.2 Comparison with Distilled Diffusion Models

Distilling Stable Diffusion 1.5. We compare Diffusion2GAN with leading diffusion distillation models on COCO2014 and COCO2017 benchmarks in Tables 3 and 4. InstaFlow-0.9B achieves an FID of 13.10 on COCO2014 and 23.4 on COCO2017, while Diffusion2GAN attains 9.29 and 19.5, respectively. Similar to our method, UFOGen [93], DMD [94], and ADD-M [77] use extra diffusion models for adversarial training or distribution matching. Although these models achieve lower FIDs compared to InstaFlow-0.9B, Diffusion2GAN still outperforms them, as Diffusion2GAN is trained to closely follow the original trajectory of the teacher diffusion model, thus mitigating the diversity collapse issue while maintaining high visual quality. Note that the concurrent work ADD-M exhibits

Table 5: Comparison to recent text-to-image models on COCO2017. On the SDXL-CFG-7 dataset, Diffusion2GAN, distilled from SDXL-Base-1.0 [63], demonstrates better FID and CLIP-score [19] over SDXL-Turbo [77] and SDXL-Lightning [44]. Our proposed diversity score, DreamDiv, confirms that SDXL-Diffusion2GAN generates more diverse images compared to SDXL-Turbo while exhibiting better text-to-image alignment compared to both SDXL-Turbo and SDXL-Lightning.

Model	# Step	FID-5k (↓)	CLIP-5k (↑)	DreamDiv-5k (↑)	DreamSim-5k (↓)
SDXL-Turbo [77]	1	28.10	0.342	0.232	0.368
SDXL-Lightning [44]	1	30.14	0.324	**0.315**	0.345
SDXL-Diffusion2GAN	1	**25.49**	**0.347**	0.268	**0.284**
SDXL-Base-1.0 (Teacher) [63]	50	25.56	0.346	0.338	0.0

a higher CLIP-score compared to Diffusion2GAN. We hypothesize this is because ADD-M conditions the discriminator using both image and text embeddings, as shown in Table 1(b) of the ADD-M

Distilling SDXL-Base-1.0. To demonstrate Diffusion2GAN's effectiveness for a larger text-to-image model, we distill SDXL-Base-1.0 [63] into Diffusion2GAN and evaluate its performance using FID and CLIP-score on COCO2017. Our empirical analysis shows that Recall [40] is inadequate for measuring image diversity. Instead, we generate 8 images per prompt and calculate the average pairwise perceptual distance using DreamSim [12], naming this metric *DreamDiv*. This metric captures diversity through perceptual dissimilarity within the same prompt. As shown in Table 5, SDXL-Diffusion2GAN achieves comparable FID and CLIP-scores to the teacher SDXL-Base-1.0 while exhibiting higher DreamDiv compared to SDXL-Turbo. SDXL-Lightning shows higher DreamDiv but a lower CLIP-score than SDXL-Diffusion2GAN, indicating its high diversity likely results from poor text-to-image alignment.

To quantify the ability to learn the diffusion teacher's ODE trajectory, we introduce *DreamSim-5k*. We simulate the ODE of both the target diffusion model and each one-step generator using 5k randomly sampled noises and COCO2017 prompts. DreamSim-5k is computed by averaging DreamSim [12] scores between image pairs generated from the same noise. A lower DreamSim-5k indicates better preservation of the teacher model's noise-image mapping. As shown in Table 5, SDXL-Diffusion2GAN outperforms SDXL-Turbo and SDXL-Lightning in learning the noise-image mapping of the SDXL-Base-1.0 teacher.

4.3 Human Preference Evaluation

We conduct human preference evaluations following the LADD human study procedure [75]. For Stable Diffusion 1.5 distillation, Diffusion2GAN shows better human preferences for both image realism and text-to-image alignment compared to InstaFlow-0.9B (Figure 6). For SDXL-Base-1.0 distillation, SDXL-Diffusion2GAN demonstrates comparable or superior image realism and text-to-image alignment compared to SDXL-Turbo and SDXL-Lightning (Figure 6).

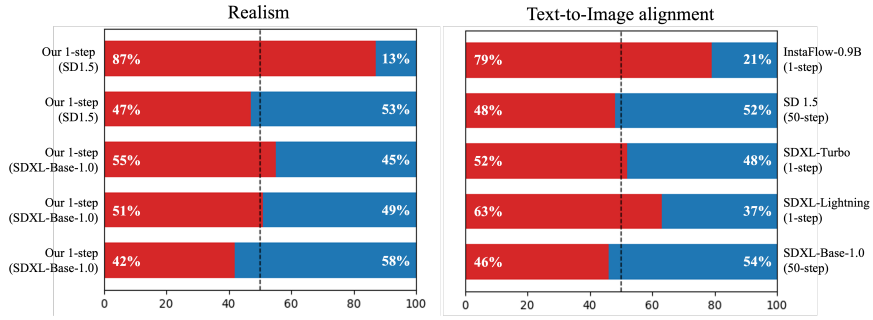

Fig. 6: We evaluate human preferences for image realism and text-to-image alignment.

Model	Total # NFE	# saved images	FID-50k (↓)
Consist. Distill. [84]	819.2M	50k	3.67
ODE distillation (LPIPS [97])	78.6M	50k	8.51
	80.3M	100k	5.62
	83.8M	200k	3.85
	94.3M	500k	3.25
	111.8M	1.0M	**3.16**

Table 6: LPIPS regression achieves better FID [20] than Consistency Distillation [84] on CIFAR10 [38], while needing fewer number of function evaluations (NFE) for both ODE pair generation and model training.

Model	A100 days	FID-30k (↓)
InstaFlow-0.9B [48]	183.2	13.10
ODE distillation	36.0	15.94
Diffusion2GAN	43.6	**9.29**

Table 7: Diffusion2GAN requires fewer A100 GPU days for training and attains a significantly lower FID compared to InstaFlow [48]. The number of A100 days and FID for InstaFlow are obtained from the original paper. We train the ODE distillation model and Diffusion2GAN using a batch size of 256.

4.4 Training Speed

Even with the cost of preparing the ODE dataset, Diffusion2GAN converges more efficiently than existing distillation methods. On the CIFAR10 dataset, we compare the total number of function evaluations of the generator during training. Training with the LPIPS loss on 500k teacher outputs surpasses the FID of Consistency Distillation [84] with a fraction of the compute budget (Table 6). In text-to-image synthesis, our full version of Diffusion2GAN achieves superior FID compared to InstaFlow while using significantly fewer GPU days (Table 7).

5 Conclusion

We have proposed a new framework *Diffusion2GAN* for distilling a pre-trained multi-step diffusion model into a one-step generator trained with conditional GAN and perceptual losses. Our study shows that separating generative modeling into two tasks—first identifying correspondences and then learning a mapping—allows us to use different generative models to improve the performance-runtime tradeoff. Our one-step model is not only beneficial for interactive image generation but also offers the potential for efficient video and 3D applications.

Acknowledgments. We would like to thank Tianwei Yin, Seungwook Kim, and Sungyeon Kim for their valuable feedback and comments. Part of this work was done while Minguk Kang was an intern at Adobe Research. Minguk Kang and Suha Kwak were supported by the NRF grant and IITP grant funded by Ministry of Science and ICT, Korea (NRF-2021R1A2C3012728, AI Graduate School (POSTECH): RS-2019-II191906). Jaesik Park was supported by the IITP grant funded by the Korea government (MSIT) (AI Graduate School (SNU): RS-2021-II211343 and AI Innovation Hub: RS-2021-II212068). Jun-Yan Zhu was supported by the Packard Fellowship.

References

1. Ascher, U.M., Petzold, L.R.: Computer methods for ordinary differential equations and differential-algebraic equations. Siam (1998) 4
2. Atkinson, K.: An introduction to numerical analysis. John wiley & sons (1991) 4
3. Berthelot, D., Autef, A., Lin, J., Yap, D.A., Zhai, S., Hu, S., Zheng, D., Talbot, W., Gu, E.: TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation. arXiv preprint arXiv:2303.04248 (2023) 4
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 4
5. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: International Conference on Learning Representations (ICLR) (2019) 4
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
7. Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations (ICLR) (2024) 12
8. Chen, Y.H., Sarokin, R., Lee, J., Tang, J., Chang, C.L., Kulik, A., Grundmann, M.: Speed is all you need: On-device acceleration of large diffusion models via gpu-aware optimizations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
9. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
10. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 7
11. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2016) 2
12. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In: Conference on Neural Information Processing Systems (NeurIPS) (2023) 13

13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In: International Conference on Learning Representations (ICLR) (2023) 4
14. Goodfellow, I.: Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016) 8
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Conference on Neural Information Processing Systems (NeurIPS) (2014) 1, 2, 4
16. Gu, J., Zhai, S., Zhang, Y., Liu, L., Susskind, J.M.: Boot: Data-free distillation of denoising diffusion models with bootstrapping. In: ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling (2023) 4
17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In: International Conference on Learning Representations (ICLR) (2024) 4
18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control. In: International Conference on Learning Representations (ICLR) (2023) 4
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 2, 11, 12, 13
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Conference on Neural Information Processing Systems (NeurIPS) (2017) 2, 11, 12, 14
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. In: Advances in Neural Information Processing Systems Deep Learning and Representation Learning Workshop (2015) 4
22. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 1, 4
23. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Conference on Neural Information Processing Systems (NeurIPS) (2020) 1, 4, 5
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: Conference on Neural Information Processing Systems (NeurIPS) Workshop (2022) 3
25. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics: Methodology and distribution (1992) 8
26. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Conference on Neural Information Processing Systems (NeurIPS) (2019) 6
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2, 4, 8
28. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (ECCV) (2016) 2
29. Kang, M., Shim, W., Cho, M., Park, J.: Rebooting ACGAN: Auxiliary Classifier GANs with Stable Training. In: Conference on Neural Information Processing Systems (NeurIPS) (2021) 4

30. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 4, 9, 10, 11, 12
31. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the Design Space of Diffusion-Based Generative Models. In: Conference on Neural Information Processing Systems (NeurIPS) (2022) 4, 5
32. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Conference on Neural Information Processing Systems (NeurIPS) (2020) 7
33. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1
34. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4, 9
35. Kettunen, M., Härkönen, E., Lehtinen, J.: E-lpips: robust perceptual image similarity via random transformation ensembles. arXiv preprint arXiv:1906.03973 (2019) 2, 7
36. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion. In: International Conference on Learning Representations (ICLR) (2023) 4
37. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4
38. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Ph.D. thesis, University of Toronto (2012) 14
39. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
40. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved Precision and Recall Metric for Assessing Generative Models. In: Conference on Neural Information Processing Systems (NeurIPS) (2019) 13
41. Li, M., Lin, J., Meng, C., Ermon, S., Han, S., Zhu, J.Y.: Efficient spatially sparse inference for conditional gans and diffusion models. In: Conference on Neural Information Processing Systems (NeurIPS) (2022) 4
42. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In: Conference on Neural Information Processing Systems (NeurIPS) (2023) 4
43. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
44. Lin, S., Wang, A., Yang, X.: SDXL-Lightning: Progressive Adversarial Diffusion Distillation. arXiv preprint arXiv:2402.13929 (2024) 2, 3, 4, 13
45. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014) 2, 9
46. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-Shot Unsupervised Image-to-Image Translation. In: IEEE International Conference on Computer Vision (ICCV) (2019) 4

47. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022) 4
48. Liu, X., Zhang, X., Ma, J., Peng, J., qiang liu: InstaFlow: One Step is Enough for High-Quality Diffusion-Based Text-to-Image Generation. In: International Conference on Learning Representations (ICLR) (2024) 4, 5, 12, 14
49. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Conference on Neural Information Processing Systems (NeurIPS) (2022) 4, 12
50. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022) 4
51. Luhman, E., Luhman, T.: Knowledge distillation in iterative generative models for improved sampling speed. arXiv preprint arXiv:2101.02388 (2021) 4, 5, 11
52. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv preprint arXiv:2310.04378 (2023) 4
53. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2022) 4
54. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 4, 5, 12
55. Mescheder, L., Nowozin, S., Geiger, A.: Which Training Methods for GANs do actually Converge? In: International Conference on Machine Learning (ICML) (2018) 9
56. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets. arXiv preprint arXiv 1411.1784 (2014) 2, 4
57. Miyato, T., Koyama, M.: cGANs with Projection Discriminator. In: International Conference on Learning Representations (ICLR) (2018) 4
58. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023) 4
59. Odena, A., Olah, C., Shlens, J.: Conditional Image Synthesis with Auxiliary Classifier GANs. In: International Conference on Machine Learning (ICML) (2017) 4
60. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive Learning for Unpaired Image-to-Image Translation. In: European Conference on Computer Vision (ECCV) (2020) 4
61. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2, 4, 6
62. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. In: Conference on Neural Information Processing Systems (NeurIPS) (2020) 4
63. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations (ICLR) (2024) 2, 3, 5, 10, 13
64. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: International Conference on Learning Representations (ICLR) (2023) 1, 4

65. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 1
66. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning (ICML) (2016) 10
67. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Conference on Neural Information Processing Systems (NeurIPS) (2016) 4
68. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
69. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1, 2, 5, 6
70. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: Stable Diffusion. https://github.com/CompVis/stable-diffusion, accessed: 2022-11-06 5, 12
71. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: Stable Diffusion 1.5. https://github.com/runwayml/stable-diffusion, accessed: 2022-11-06 9, 11, 12
72. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
73. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: Conference on Neural Information Processing Systems (NeurIPS) (2022) 1
74. Salimans, T., Ho, J.: Progressive Distillation for Fast Sampling of Diffusion Models. In: International Conference on Learning Representations (ICLR) (2022) 4, 5
75. Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P., Rombach, R.: Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. arXiv preprint arXiv:2403.12015 (2024) 4, 13
76. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. In: International Conference on Machine Learning (ICML) (2023) 1, 4
77. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 2, 3, 4, 12, 13
78. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings (2022) 4
79. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: Conference on Neural Information Processing Systems (NeurIPS) (2022) 1
80. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 6
81. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML) (2015) 1, 4

82. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: International Conference on Learning Representations (ICLR) (2021) 5, 10
83. Song, Y., Dhariwal, P.: Improved Techniques for Training Consistency Models. In: International Conference on Learning Representations (ICLR) (2024) 8, 11
84. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency Models. In: International Conference on Machine Learning (ICML) (2023) 2, 4, 5, 11, 14
85. Song, Y., Garg, S., Shi, J., Ermon, S.: Sliced score matching: A scalable approach to density and score estimation. In: Uncertainty in Artificial Intelligence. PMLR (2020) 4
86. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: International Conference on Learning Representations (ICLR) (2021) 1, 4, 5
87. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation (2011) 4
88. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
89. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 6
90. Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M.: Diffusion-GAN: Training GANs with Diffusion. In: International Conference on Learning Representations (ICLR) (2023) 4
91. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion GANs. In: International Conference on Learning Representations (ICLR) (2022) 4
92. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 4
93. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023) 2, 4, 12
94. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step Diffusion with Distribution Matching Distillation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2, 12
95. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017) 4, 10
96. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (ICCV) (2023) 4
97. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2, 4, 5, 6, 11, 14
98. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In: International Conference on Learning Representations (ICLR) (2021) 4

99. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. In: Conference on Neural Information Processing Systems (NeurIPS) (2020) 7
100. Zheng, H., Nie, W., Vahdat, A., Azizzadenesheli, K., Anandkumar, A.: Fast sampling of diffusion models via operator learning. In: International Conference on Machine Learning (ICML) (2023) 4
101. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017) 2, 4