UFOGen: You Forward Once Large Scale Text-to-Image Generation via Diffusion GANs

Yanwu Xu*[1,2]†, Yang Zhao[1]†, Zhisheng Xiao[1]†, Tingbo Hou[1]

[1] Google

{yanwuxu,yzhaoeric,zsxiao,tingbo}@google.com

[2] Department of Electrical Computer Engineering, Boston University

yanwuxu@bu.edu

Abstract

Text-to-image diffusion models have demonstrated remarkable capabilities in transforming text prompts into coherent images, yet the computational cost of the multi-step inference remains a persistent challenge. To address this issue, we present UFOGen, a novel generative model designed for ultra-fast, one-step text-to-image generation. In contrast to conventional approaches that focus on improving samplers or employing distillation techniques for diffusion models, UFOGen adopts a hybrid methodology, integrating diffusion models with a GAN objective. Leveraging a newly introduced diffusion-GAN objective and initialization with pre-trained diffusion models, UFOGen excels in efficiently generating high-quality images conditioned on textual descriptions in a single step. Beyond traditional text-to-image generation, UFOGen showcases versatility in applications. Notably, UFOGen stands among the pioneering models enabling one-step text-to-image generation and diverse downstream tasks, presenting a significant advancement in the landscape of efficient generative models.

1. Introduction

Diffusion models [18, 56, 58] has recently emerged as a powerful class of generative models, demonstrating unprecedented results in many generative modeling tasks [6, 20, 29, 49, 51, 63]. In particular, they have shown the remarkable ability to synthesize high-quality images conditioned on texts [1, 43, 47, 49, 51, 66]. Beyond the text-to-image synthesis tasks, large-scale text-to-image models serve as foundational building blocks for various downstream applications, including personalized generation [9, 13, 30, 50], controlled generation [42, 70] and image editing [5, 15, 67]. Yet, despite their impressive generative quality and wide-ranging utility, diffusion models have a notable limitation: they rely on iterative denoising to generate final samples, which leads to slow generation speeds. The slow inference and the consequential computational demands of large-scale diffusion models pose significant impediments to their deployment.

In the seminal work by Song *et al*. [58], it was revealed that sampling from a diffusion model is equivalent to solving the probability flow ordinary differential equation (PF-ODE) associated with the diffusion process. Presently, the majority of research aimed at enhancing the sampling efficiency of diffusion models centers on the ODE formulation. One line of work seeks to advance numerical solvers for the PF-ODE, with the intention of enabling the solution of the ODE with greater discretization size, ultimately leading to fewer requisite sampling steps [2, 37, 38, 57]. However, the inherent trade-off between step size and accuracy still exists. Given the highly non-linear and complicated trajectory of the PF-ODE, it would be extremely difficult to reduce the number of required sampling steps to a minimal level. Even the most advanced solvers [37, 38] can generate images within 10 to 20 sampling steps, and further reduction leads to a noticeable drop in image quality. An alternative approach seeks to distill the PF-ODE trajectory from a pre-trained diffusion model. For instance, progressive distillation [31, 41, 52] tries to condense multiple discretization steps of the PF-ODE solver into a single step by explicitly aligning with the solver's output. Similarly, consistency distillation [39, 59] works on learning consistency mappings that preserve point consistency along the ODE trajectory. These methods have demonstrated the potential to significantly reduce the number of sampling steps. However, due to the intrinsic complexity of the ODE trajectory, they still struggle in the extremely small step regime.

The pursuit of developing ultra-fast large-scale diffusion models that require just one or two sampling steps, remains

*Work done as a student researcher of Google, † indicates equal contribution.

Figure 1. Images generated by our UFOGen model with **1 sampling step**.

a challenging open problem. We assert that to achieve this ambitious objective, fundamental adjustments are necessary in the formulation of diffusion models, as the current ODE-based approaches seem intrinsically constrained for very few steps sampling. In this work, we introduce a novel one-step text-to-image generative model, representing a fusion of GAN and diffusion model elements. Our inspiration stems from previous work that successfully incorporated GANs into the framework of diffusion models [61, 62, 65, 71], which have demonstrated the capacity to generate images in as few as four steps when trained on small-scale datasets. These models diverge from the traditional ODE formulation by leveraging adversarial loss for learning the denoising distribution, rather than relying on KL minimization.

Despite the promising outcomes of earlier diffusion GAN hybrid models, achieving one-step sampling and extending their utility to text-to-image generation remains a non-trivial challenge. In this research, we introduce innovative techniques to enhance diffusion GAN models, resulting in an ultra-fast text-to-image model capable of producing high-quality images in a single sampling step. In light of this achievement, we have named our model **UFOGen**, an acronym denoting "You Forward Once" Generative model. Our UFOGen model excels at generating high-quality images in just one inference step. Notably, when initialized with a pre-trained Stable Diffusion (SD) model [49], our method efficiently transforms SD into a one-step inference model while largely preserving the quality of generated content. See Figure 1 for a showcase of text-conditioned images generated by UFOGen. To the best of our knowledge, our model stands among the pioneers to achieve a reduction in the number of required sampling steps for text-to-image diffusion models to just one.

Our work presents several significant contributions:

1. We introduce UFOGen, a powerful generative model capable of producing high-quality images conditioned on text descriptions in a single inference step.
2. We present an efficient and simplified training process, enabling the fine-tuning of pre-existing large-scale diffusion models to operate as one-step generative models.
3. Our model's versatility extends to applications such as image-to-image and controllable generation, thereby unlocking the potential for one-step inference across various generative scenarios.

2. Related Works

Text-to-image Diffusion Models Denoising diffusion models [18, 56, 58] are trained to reconstruct data from corrupted inputs. The simplicity of the training objective makes denoising diffusion models well-suited for scaling up generative models. Researchers have made numerous efforts to train diffusion models on large datasets containing image-text pairs [55] for the text-to-image generation task [1, 43, 47, 49, 51, 66]. Among them, latent diffusion models, such as the popular SD models [44, 49], have gained substantial attention in the research community due to their simplicity and efficiency compared to pixel-space counterparts.

Accelerating Diffusion Models The notable issue of slow generation speed has motivated considerable efforts towards enhancing the sampling efficiency of diffusion models. These endeavors can be categorized into two primary approaches. The first focuses on the development of improved numerical solvers [2, 26, 37, 38, 57]. The second approach explores the concept of knowledge distillation [17], aiming at condensing the sampling trajectory of a numerical solver into fewer steps [3, 31, 39, 41, 52, 59]. However, both of these approaches come with significant limitations, and thus far, they have not demonstrated the ability to substantially reduce the sampling steps required for text-to-image diffusion models to a truly minimal level.

Text-to-image GANs As our model has GAN [14] as one of its component, we provide a brief overview of previous attempts of training GANs for text-to-image generation. Early GAN-based text-to-image models were primarily confined to small-scale datasets [48, 60, 64, 69]. Later, with the evolution of more sophisticated GAN architectures [24, 25, 53], GANs trained on large datasets have shown promising results in the domain of text-to-image generation [22, 54, 72]. Comparatively, our model has several distinct advantages. Firstly, to overcome the well-known issues of training instability and mode collapse, text-to-image GANs have to incorporate multiple auxiliary losses and complex regularization techniques, which makes training and parameter tuning extremely intricate. This complexity is particularly exemplified by GigaGAN [22], currently regarded as the most powerful GAN-based models. In contrast, our model offers a streamlined and robust training process, thanks to the diffusion component. Secondly, our model's design allows us to seamlessly harness pre-trained diffusion models for initialization, significantly enhancing the efficiency of the training process. Lastly, our model exhibits greater flexibility when it comes to downstream applications (see Section 5.3), an area in which GAN-based models have not explored.

Recent Progress on Few-step Text-to-image Generation While developing our model, we noticed some concurrent work on few-step text-to-image generation. Latent Consistency Model [39] extends the idea of consistency distillation [59] to SD, leading to 4-step sampling with reasonable quality. However, further reducing the sampling step results in significant quality drop. InstaFlow [35] achieves text-to-image generation in a single sampling step. Similar to our model, InstaFlow tackles the slow sampling issue of diffusion models by introducing improvements to the model itself. Notably, they extend Rectified Flow models [33, 34] to create a more direct trajectory in the diffusion process. In direct comparison to InstaFlow, our model outperforms in terms of both quantitative metrics and visual quality. Moreover, our approach presents the added benefits of a streamlined training pipeline and improved training efficiency. InstaFlow requires multiple stages of fine-tuning, followed by a subsequent distillation stage. In contrast, our model only needs one single fine-tuning stage with a minimal number of training iterations.

3. Background

Diffusion Models Diffusion models [18, 56] is a family of generative models that progressively inject Gaussian noises into the data, and then generate samples from noise via a reverse denoising process. Diffusion models define a forward process that corrupts a data sample $x_0 \sim q(x_0)$ in T steps with variance schedule β_t: $q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$. The parameterized reverse diffusion process aims to gradually recover cleaner data from noisy observations: $p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I})$.

The model $p_\theta(x_{t-1}|x_t)$ is parameterized as a Gaussian distribution, because when the denoising step size from t to $t-1$ is sufficiently small, the true denoising distribution $q(x_{t-1}|x_t)$ is a Gaussian [10]. To train the model, one can minimize the negative ELBO objective [18, 27]:

$$\mathcal{L} = \mathbb{E}_{t,q(x_0)q(x_t|x_0)}\mathrm{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)), \tag{1}$$

where $q(x_{t-1}|x_t, x_0)$ is Gaussian posterior distribution derived in [18].

Diffusion-GAN Hybrids The idea of combining diffusion models and GANs is first explored in [62]. The main motivation is that, when the denoising step size is large, the true denoising distribution $q(x_{t-1}|x_t)$ is no longer a Gaussian. Therefore, instead of minimizing KL divergence with a parameterized Gaussian distribution, they parameterized $p_\theta(x'_{t-1}|x_t)$ as a conditional GAN to minimize the adversarial divergence between model $p_\theta(x'_{t-1}|x_t)$ and $q(x_{t-1}|x_t)$:

$$\min_\theta \mathbb{E}_{q(x_t)}\Big[D_{adv}(q(x_{t-1}|x_t)||p_\theta(x'_{t-1}|x_t))\Big]. \tag{2}$$

The objective of Denoising Diffusion GAN (DDGAN) in

[62] can be expressed as:

$$\min_{\theta} \max_{D_\phi} \mathbb{E}_{q(x_t)} \Big[\mathbb{E}_{q(x_{t-1}|x_t)} [\log(D_\phi(x_{t-1}, x_t, t))] + \mathbb{E}_{p_\theta(x'_{t-1}|x_t)} [\log(1 - D_\phi(x'_{t-1}, x_t, t))] \Big], \quad (3)$$

where D_ϕ is the conditional discriminator network, and the expectation over the unknown distribution $q(x_{t-1}|x_t)$ can be approximated by sampling from $q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})$. The flexibility of a GAN-based denoising distribution surpasses that of a Gaussian parameterization, enabling more aggressive denoising step sizes. Consequently, DDGAN successfully achieves a reduction in the required sampling steps to just four.

Nonetheless, the utilization of a purely adversarial objective in DDGAN introduces training instability, as documented by the findings in [65]. In response to this challenge, the authors in [65] advocated matching the joint distribution $q(x_{t-1}, x_t)$ and $p_\theta(x_{t-1}, x_t)$, as opposed to the conditional distribution as outlined in Equation 2. [65] further demonstrated that the joint distribution matching can be disassembled into two components: matching marginal distributions using adversarial divergence and matching conditional distributions using KL divergence:

$$\min_{\theta} \mathbb{E}_{q(x_t)} \Big[D_{adv}(q(x_{t-1})||p_\theta(x_{t-1})) + \lambda_{KL} \mathrm{KL}(q(x_t|x_{t-1})||p_\theta(x_t|x_{t-1})) \Big]. \quad (4)$$

The objective of adversarial divergence minimization is similar to Equation 3 except that the discriminator does not take x_t as part of its input. The KL divergence minimization translates into a straightforward reconstruction objective, facilitated by the Gaussian nature of the diffusion process. This introduction of a reconstruction objective plays a pivotal role in enhancing the stability of the training dynamics. As observed in [65], this approach led to markedly improved results, especially on more intricate datasets.

4. Methods

In this section, we present a comprehensive overview of the enhancements we have made in our diffusion-GAN hybrid models, ultimately giving rise to the UFOGen model.

4.1. Enabling One-step Sampling for UFOGen

Diffusion-GAN hybrid models are tailored for training with a large denoising step size. However, attempting to train these models with just a single denoising step effectively reduces the training to that of a conventional GAN. Consequently, prior diffusion-GAN models were unable to achieve one-step sampling. In light of this challenge, we conducted an in-depth examination of the SIDDM [65] formulation and implemented specific modifications in the generator parameterization and the reconstruction term within the objective. These adaptations enabled UFOGen to perform one-step sampling, while retaining training with several denoising steps.

Parameterization of the Generator In diffusion-GAN models, the generator should produce a sample of x_{t-1}. However, instead of directly outputting x_{t-1}, the generator of DDGAN and SIDDM is parameterized by $p_\theta(x_{t-1}|x_t) = q(x_{t-1}|x_t, x_0 = G_\theta(x_t, t))$. In other words, first x_0 is predicted using the denoising generator $G_\theta(x_t, t)$, and then, x_{t-1} is sampled using the Gaussian posterior distribution $q(x_{t-1}|x_t, x_0)$ derived in [18, 62]. Note that this parameterization is mainly for practical purposes, as discussed in [62], and alternative parameterization would not break the model formulation.

We propose another plausible parameterization for the generator: $p_\theta(x_{t-1}) = q(x_{t-1}|x_0 = G_\theta(x_t, t))$. The generator still predicts x_0, but we sample x_{t-1} from the forward diffusion process $q(x_{t-1}|x_0)$ instead of the posterior. As we will show later, this design allows distribution matching at x_0, paving the path to one-step sampling.

Improved Reconstruction Loss at x_0 We argue that with the new generator parameterization, the objective of SIDDM in Equation 4 indirectly matches the distribution at x_0. To see this, we analyze the adversarial objective and KL objective in Equation 4 separately. The first term minimizes adversarial divergence $D_{adv}\big(q(x_{t-1})||p_\theta(x'_{t-1})\big)$, where $q(x_{t-1})$ and $p_\theta(x'_{t-1})$ can both be seen as the corruption of a distribution at x_0 by the same Gaussian kernel. Specifically, since $q(x_{t-1}) = \mathbb{E}_{q(x_0)}[q(x_{t-1}|x_0)]$, given a sample $x_0 \sim q(x_0)$, we have $q(x_t) = \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})$, according to the forward diffusion formulation [18]. Similarly, $p_\theta(x'_{t-1})$ has the same form except that x_0 is produced by the generator. As a result, adversarial distribution matching on $q(x_{t-1})$ and $p_\theta(x'_{t-1})$ will also encourage the matching between $q(x_0)$ and $p_\theta(x'_0)$, which is the distribution over x_0 produced by the generator. A formal explanation can be found in the supplementary file.

The second term in the objective minimizes the KL divergence between $q(x_t|x_{t-1})$ and $p_\theta(x_t|x'_{t-1})$, which can be simplified to the following reconstruction term:

$$\mathbb{E}_{q(x_t)} \Big[\frac{(1 - \beta_t)||x'_{t-1} - x_{t-1}||^2}{2\beta_t} \Big]. \quad (5)$$

Based on above analysis, it is easy to see that minimizing this reconstruction loss will essentially matches x_0 and x'_0 as well (a straightforward derivation is provided in the supplementary file).

Per our analysis, both terms in the SIDDM objective in Equation 4 implicitly matches the distribution at x_0, which suggests that one-step sampling is possible. However, empirically we observe that one-step sampling from SIDDM

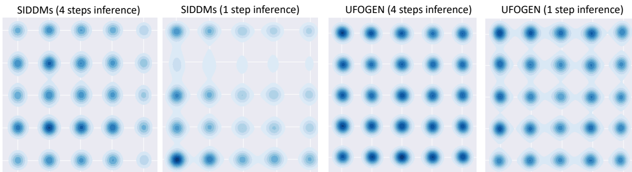

Figure 2. Results of training with UFOGen loss versus the original loss of SIDDM on 25-Gaussian toy data. With the modified objective, UFO enables one-step sampling.

does not work well even on 2-D toy dataset (See Figure 2). We conjecture that this is due to the variance introduced in the additive Gaussian noise when sampling x_{t-1} with x_0. To reduce the variance, we propose to replace the reconstruction term in Equation 5 with the reconstruction at clean sample $||x_0 - x'_0||^2$, so that the matching at x_0 becomes explicit. We observe that with this change, we can obtain samples in one step, as shown in Figure 2.

Training and Sampling of UFOGen To put things together, we present the complete training objective and strategy for the UFOGen model. UFOGen is trained with the following objective:

$$\min_{\theta} \max_{D_\phi} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0),p_\theta(x'_0)p_\theta(x'_{t-1}|x'_0)} \Big[[\log(D_\phi(x_{t-1},t))] + [\log(1 - D_\phi(x'_{t-1},t))] + \lambda_{KL}\gamma_t \left\|x_0 - x'_0\right\|^2 \Big], \tag{6}$$

where γ_t is a time-dependent coefficient. The objective consists of an adversarial loss to match noisy samples at time step $t-1$, and a reconstruction loss at time step 0. Note that the reconstruction term is essentially the training objective of diffusion models [18, 58], and therefore the training of UFOGen model can also be interpreted as training a diffusion model with adversarial refinement. The training scheme of UFOGen is presented in Algorithm 1.

Despite the straightforward nature of the modifications to the training objective, these enhancements have yielded impressive outcomes, particularly evident in the context of one-step sampling, where we simply sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ and produce sample $x'_0 = G_\theta(x_T)$.

4.2. Leverage Pre-trained Diffusion Models

Our objective is developing an ultra-fast text-to-image model. However, the transition from an effective UFOGen recipe to web-scale data presents considerable challenges. Training diffusion-GAN hybrid models for text-to-image generation encounters several intricacies. Notably, the discriminator must make judgments based on both texture and semantics, which govern text-image alignment. This challenge is particularly pronounced during the initial stage of training. Moreover, the cost of training text-to-image models can be extremely high, particularly in the case of GAN-based models, where the discriminator introduces additional parameters. Purely GAN-based text-to-image models [22, 54] confront similar complexities, resulting in highly intricate and expensive training.

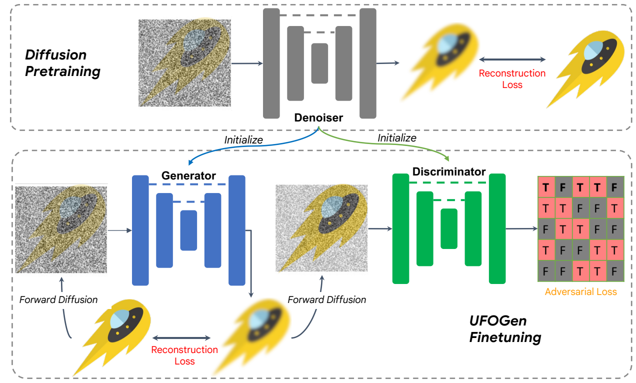

Figure 3. Illustration of UFOGen training.

Algorithm 1 UFOGen Training

Require: Generator G_θ, discriminator D_ϕ, loss coefficient λ_{KL}

repeat
  Sample $x_0 \sim q(x_0), t-1 \sim$ Uniform$(0, ..., T-1)$.
  Sample $x_{t-1} \sim q(x_{t-1}|x_0), x_t \sim q(x_t|x_{t-1})$
  Sample $x'_{t-1} \sim q(x_{t-1}|x'_0)$, where $x'_0 = G_\theta(x_t, t)$
  Update D_ϕ with gradient
  $\nabla_\phi \left(\log\left(D_\phi(x_{t-1}, t-1)\right) + \log\left(1 - D_\phi(x'_{t-1}, t-1)\right)\right)$
  Update G_θ with gradient
  $\nabla_\theta \left(\log(1 - D_\phi(x'_{t-1}, t-1) + \lambda_{KL}\gamma_t ||x_0 - x'_0||_2^2\right)$
until converged

To surmount the challenges of scaling-up diffusion-GAN hybrid models, we propose the utilization of pre-trained text-to-image diffusion models, notably the SD model [49]. Specifically, our UFOGen model is designed to employ a consistent UNet structure for both its generator and discriminator. This design enables seamless initialization with the pre-trained SD model. We posit that the internal features within the SD model contain rich information of the intricate interplay between textual and visual data. This initialization strategy significantly streamlines the training of UFOGen. Upon initializing UFOGen's generator and discriminator with the SD model, we observe stable training dynamics and remarkably fast convergence. The complete training strategy of UFOGen is illustrated in Figure 3.

5. Experiments

In this section, we evaluate our proposed UFOGen model with text-to-image generation, ablation studies, and downstream applications.

Method	#Steps	Time (s)	FID-5k	CLIP
DPM Solver [37]	25	0.88	**20.1**	0.318
	8	0.34	31.7	**0.320**
Progressive Distillation [41]	1	0.09	37.2	0.275
	2	0.13	26.0	0.297
	4	0.21	26.4	0.300
CFG-Aware Distillation [31]	8	0.34	24.2	0.30
InstaFlow-0.9B	1	0.09	23.4	0.304
InstaFlow-1.7B	1	0.12	**22.4**	0.309
UFOGen	1	0.09	22.5	**0.311**

Table 1. Comparison of FID on MSCOCO-2017 5k and CLIP score. All models are based on SD. Numbers of progressive distillation and InstaFlow are cited from [35].

5.1. Text-to-image Generation

Configuration for Training and Evaluation For experiments on text-to-image generation, we follow the scheme proposed in Section 4.2 to initialize both the generator and discriminator with the pre-trained SD 1.5[1] model [49]. We train our model on the LAION-Aesthetics-6+ subset of LAION-5B [55]. For evaluation, we adopt the common practice that uses zero-shot FID [16] on MS-COCO [32], and CLIP score with ViT-g/14 backbone [45].

Main Results To kick-start our evaluation, we perform a comparative analysis in Table 1, bench-marking UFOGen against other few-step sampling models that share the same SD backbone. Our baselines include Progressive Distillation [41] and its variant [31], which are previously the state-of-the-art for few-step sampling of SD, as well as the concurrent work of InstaFlow [35]. Latent Consistency Model (LCM) [39] is excluded, as the metric is not provided in their paper. Analysis of the results presented in Table 1 reveals the superior performance of our single-step UFOGen when compared to Progressive Distillation across one, two, or four sampling steps, as well as the CFG-Aware distillation [31] in eight steps. Furthermore, our method demonstrates advantages in terms of both FID and CLIP scores over the single-step competitor, InstaFlow-0.9B, which share the same network structure of SD with us. Impressively, our approach remains highly competitive even when compared to InstaFlow-1.7B with stacked UNet structures, which effectively doubles the parameter count.

The results depicted in Table 1 may suggest that InstaFlow remains a strong contender in one-step generation alongside UFOGen. However, we argue that relying solely on the MS-COCO zero-shot FID score for evaluating visual quality might not be the most reliable metric, a concern highlighted in prior research such as [28, 44] and discussed by [4]. Consequently, we believe that qualitative assessments can provide more comprehensive insights. We present qualitative comparisons involving InstaFlow and LCM[2] in Table 2. The comparisons allow for a clear-cut conclusion: UFOGen's one-step image generation surpasses InstaFlow by a substantial margin in terms of image quality. Notably, UFOGen also demonstrates significant advantages when contrasted with the 2-step LCM, as showed by the evident blurriness present in LCM's samples. Furthermore, even when compared to the samples generated by the 4-step LCM, our generated images exhibit distinct characteristics, including sharper textures and finer details. We do not present results of single-step LCM, as we observe that it fail to generate any textures.Additional examples of the comparison are provided in the supplementary file, where we display multiple images generated by each model for different prompts.

For completeness, we extend our comparison to encompass a diverse array of text-to-image generative models in Table 3. While the results in Table 3 are not directly comparable due to substantial variations in model architecture, parameter count, and training data, it is noteworthy that UFOGen is a competitive contender among the contemporary landscape of text-to-image models, offering the advantage of remarkable speed over auto-regressive or diffusion models, thanks to its inherent one-step generation capability.

Based on both quantitative and qualitative assessments, we assert that UFOGen stands as a powerful text-to-image generative model, capable of producing sharp and visually appealing images that align well with the provided text conditioning, all in a single step. Our evaluation underscores its capacity to produce superior sample quality when contrasted with competing diffusion-based methods designed for a few-step generation process.

5.2. Ablation Studies

Ablation studies have been conducted to offer deeper insights into the effectiveness of our training strategies. As outlined in Table 4, we compare the training of diffusion-GAN hybrid models using the SIDDM objective [65] against the proposed UFOGen objective in Section 4.1. The results validate our assertions, demonstrating that the modifications in the UFOGen objective facilitate one-step sampling. We additionally provide qualitative samples, and an supplementary ablation study on the denoising step size during training in the supplementary file.

5.3. Applications

A promising aspect of text-to-image diffusion models is their versatility as foundational components for various applications, whether fine-tuned or utilized as is. In this sec-

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

[2]InstaFlow (https://huggingface.co/spaces/XCLiu/InstaFlow) and LCM (https://huggingface.co/spaces/SimianLuo/Latent_Consistency_Model)

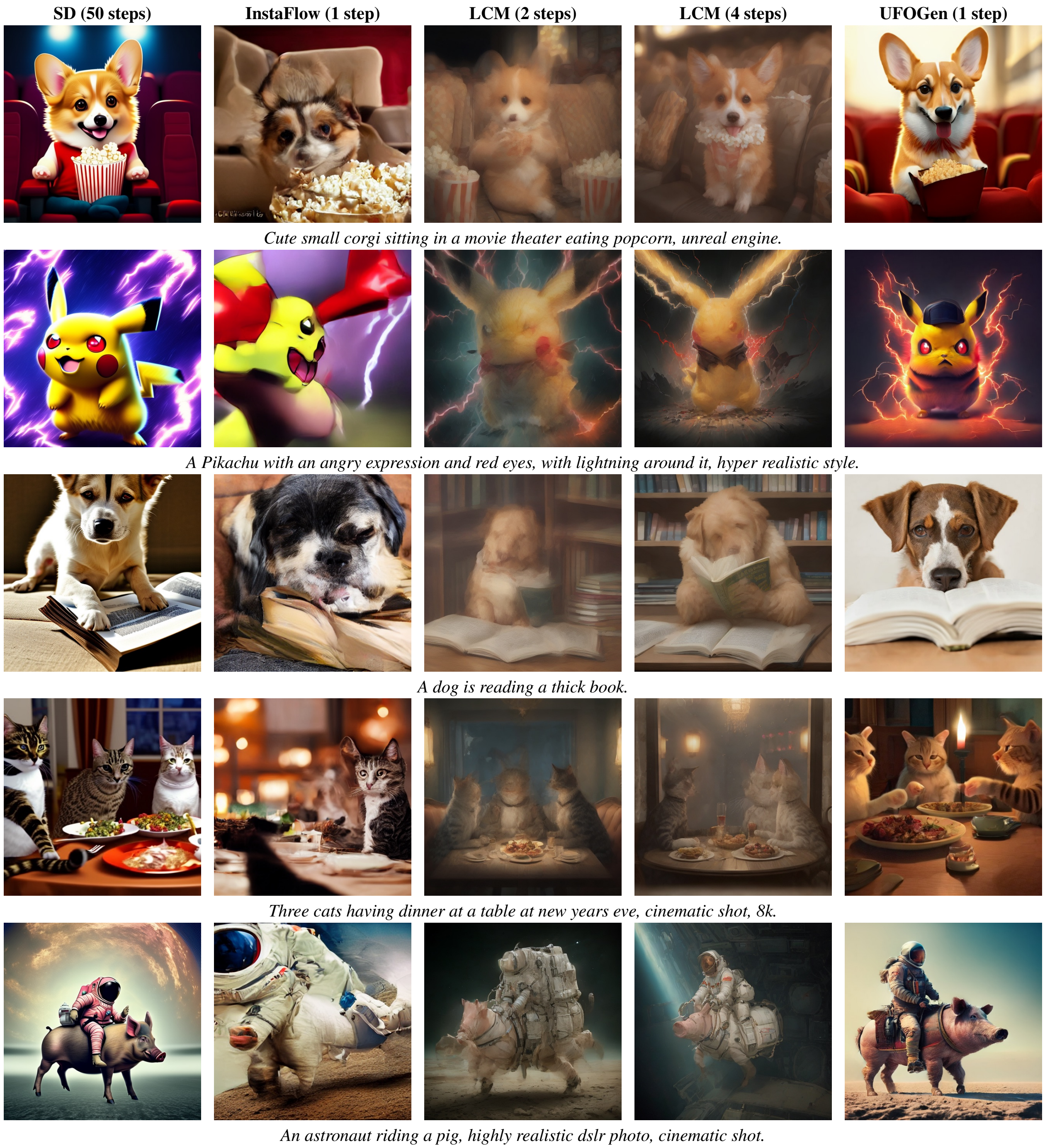

Table 2. Qualitative comparisons of UFOGen against competing methods and SD baseline. Zoom-in for better viewing.

tion, we showcase UFOGen's ability to extend beyond text-to-image generation, while benefiting from its unique advantage of single-step generation. Specifically, we explore two applications of UFOGen: image-to-image [40] generation and controllable generation [42, 70].

Table 5 showcases UFOGen's image-to-image generation outcomes. Following SDEdit [40], we introduce a suitable amount of noise to the input data, and let UFOGen to execute single-step generation based on the given prompt. Our observations affirm that UFOGen adeptly produces samples that adhere to the specified conditions of both the prompt and the input image.

To facilitate controllable generation, we conduct fine-tuning of UFOGen by incorporating an additional adapter

Method	Type	Time (s)	# Param.	FID-30k
DALLE [46]	AR	-	12B	27.5
Parti-20B [68]	AR	-	20B	7.23
Make-A-Scene [12]	AR	25.0	-	11.84
GLIDE [43]	Diff	15.0	5B	12.24
DALLE 2 [47]	Diff	-	5.5B	10.39
Imagen [19]	Diff	9.1	3B	7.27
eDiff-I [1]	Diff	32.0	9B	6.95
SD [49]	Diff	2.9	0.9B	9.62
LAFITE [72]	GAN	0.02	75M	26.94
StyleGAN-T [54]	GAN	0.10	1B	13.90
GigaGAN [23]	GAN	0.13	1B	9.09
Muse-3B [7]	-	1.3	3B	7.88
InstaFlow [35]	-	0.09	0.9B	13.10
UFOGen (Ours)	-	0.09	0.9B	12.78

Table 3. Comparison of FID on MSCOCO 2014 with 30k images. Numbers of other models are cited from [35]. Inference time measurement follows the setting of [21].

Method	#Steps	FID-5k	CLIP
SIDDM [65]	4	21.7	0.306
	1	28.0	0.289
UFOGen	4	22.1	0.307
	1	22.5	0.311

Table 4. Ablation study comparing the SIDDM objective with our UFOGen objective, incorporating the introduced modifications detailed in Section 4.1.

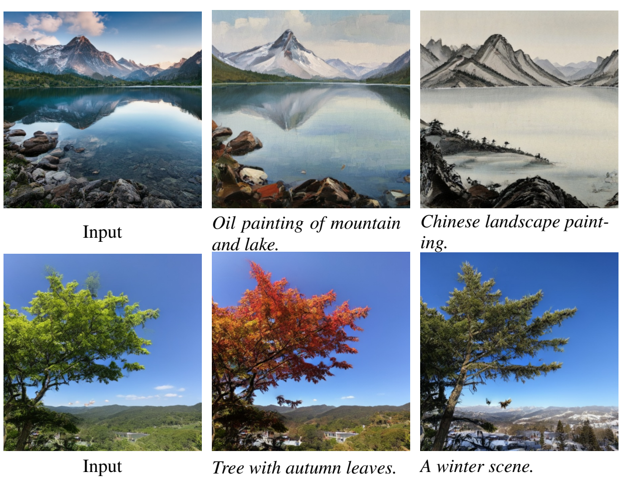

Table 5. Results of single-step image-to-image generation by UFOGen. Zoom in to view the details.

network, akin to the approach outlined in [42]. This adapter network takes control signals as input to guide the generation process. In our exploration, we employ two types of control signals: depth maps and canny edges. The results are presented in Table 6. Post fine-tuning, UFOGen exhibits the ability to generate high-quality samples that align with both the provided prompt and control signal.

Our results highlight UFOGen can work on diverse generation tasks in a single step, a distinctive feature that, to the best of our knowledge, sets our model apart. Unlike GAN-based text-to-image models [22, 54], which lack the ability to handle zero-shot image-to-image generation tasks as they do not generate samples through denoising, UFOGen excels in this context. Moreover, our model succeeds in controllable generation, a domain that earlier GAN-based models have not explored due to the complexities of fine-tuning and adding supplementary modules to the StyleGAN architecture. Consequently, the flexibility of our model in addressing various downstream tasks positions it uniquely among one-step text-to-image models. Additional results of the applications are provided in the supplementary file.

Table 6. Results of controllable generation by UFOGen.

6. Conclusions

In this paper, we present UFOGen, a groundbreaking advancement in text-to-image synthesis that effectively addresses the enduring challenge of inference efficiency. Our innovative hybrid approach, combining diffusion models with a GAN objective, propels UFOGen to achieve ultra-fast, one-step generation of high-quality images conditioned on textual descriptions. The comprehensive evaluations consistently affirm UFOGen's superiority over existing accelerated diffusion-based methods. Its distinct capability for one-step text-to-image synthesis and proficiency in downstream tasks underscore its versatility and mark it as a standout in the field. As a pioneer in enabling ultra-fast text-to-image synthesis, UFOGen paves the way for a transformative shift in the generative models landscape. The potential impact of UFOGen extends beyond academic discourse, promising to revolutionize the practical landscape of rapid and high-quality image generation.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 3, 8

[2] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 1, 3

[3] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbot, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023. 3

[4] Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*, 2022. 6

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[6] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 364–381. Springer, 2020. 1

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 8

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 17

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 1

[10] William Feller. Retracted chapter: On the theory of stochastic processes, with particular reference to applications. In *Selected Papers I*, pages 769–798. Springer, 2015. 3

[11] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 17

[12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 8

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3, 4, 5, 13

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 8

[20] Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022. 1

[21] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis, 2023. 8

[22] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 5, 8

[23] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 8

[24] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7799–7808, 2020. 3

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[26] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 3

[27] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[28] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 6

[29] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. 1

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1

[31] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023. 1, 3, 6

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[33] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. 3

[34] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[35] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023. 3, 6, 8

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 14

[37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 1, 3, 6

[38] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1, 3

[39] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1, 3, 6, 15, 17

[40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 7

[41] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 1, 3, 6

[42] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 7, 8

[43] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1, 3, 8

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 6

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 8

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1, 3, 8

[48] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5, 6, 8

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi.

Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 1, 3

[52] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 1, 3

[53] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 3

[54] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 3, 5, 8

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3, 6

[56] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1, 2, 3

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 3

[58] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1, 2, 5

[59] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. 1, 3, 15

[60] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022. 3

[61] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[62] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022. 2, 3, 4, 13, 14

[63] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022. 1

[64] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[65] Yanwu Xu, Mingming Gong, Shaoan Xie, Wei Wei, Matthias Grundmann, Tingbo Hou, et al. Semi-implicit denoising diffusion models (siddms). *arXiv preprint arXiv:2306.12511*, 2023. 2, 4, 6, 8, 12, 14

[66] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *arXiv preprint arXiv:2305.18295*, 2023. 1, 3

[67] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1

[68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 8

[69] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 7

[71] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[72] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 3, 8