Generative Adversarial Diffusion

U-Chae Jun Jaeeun Ko Jiwoo Kang*
Sookmyung Women's University, South Korea
wjsdbco@sookmyung.ac.kr rhwodms1223@sookmyung.ac.kr jwkang@sookmyung.ac.kr

Abstract

We introduce a novel generative framework that unifies adversarial and diffusion-based training to overcome the limitations of conventional models. Our approach, termed Generative Adversarial Diffusion (GAD), integrates an adversarial loss directly into each denoising step of a latent diffusion model. By employing a single U-Net as a unified generator and discriminator, our framework eliminates the need for a separate discriminator, thereby reducing memory overhead and mitigating common GAN issues such as mode collapse and training instability. This integrated adversarial regularizer promotes semantic information exchange across timesteps, enabling the model to better capture complex data distributions even when training data is scarce or biased. Extensive experiments on standard latent diffusion benchmarks demonstrate that GAD significantly enhances image quality and mode coverage in tasks including text-to-image and image-to-3D generation. Our results suggest that unifying adversarial and diffusion-based training in a single network offers a promising new direction for high-fidelity, stable image synthesis[1].

1. Introduction

Text-to-image [36, 37] and image-to-3D [22, 26, 27] generation tasks have recently attracted significant attention in computer vision, driven by rapid advances in model architectures and training techniques. Consequently, generative models [36, 51, 58] have evolved to meet these requirements of various generation tasks. Generative Adversarial Networks (GANs) [16] are among the most prominent frameworks, where a generator and discriminator engage in an adversarial game to produce high-quality images. GANs offer fast inference speeds and can be trained on relatively small datasets while also demonstrating impressive high-resolution image generation capabilities. However, GANs are notoriously unstable during training because the simultaneous update of the generator and discriminator can lead to issues such as mode collapse and non-convergence. For instance, if the discriminator becomes too strong early in the training, the generator may not receive meaningful gradients, causing it to stagnate. Similarly, the generator might converge to a limited subset of the data distribution that consistently deceives the discriminator, further exacerbating mode collapse. Despite various improvements, such as architecture modifications [3, 18, 30], revised objectives [4, 7, 61], and regularization strategies [13, 31], completely resolving these issues within the standard GAN framework remains challenging.

In contrast, diffusion models [36, 37, 58] have recently shown great promise across a wide range of generative tasks. In particular, latent diffusion models (LDMs) [36] reduce computational complexity by performing denoising in a low-dimensional latent space rather than directly in image space, while still producing high-quality images. Diffusion models also exhibit stable convergence and robust distribution coverage when trained on large datasets, making them well-suited for tasks such as text-to-image [36, 37] and image-to-3D [22, 26, 27] generation. Although diffusion models require an iterative denoising process, which can be computationally expensive, the primary focus of our work is not on accelerating inference but on enhancing generation quality and training stability. In practice, the conventional diffusion process optimizes each timestep independently under a Markov assumption, thereby limiting the semantic information exchange between steps. This can result in suboptimal performance when training on complex or limited datasets.

To overcome these limitations, researchers have explored combining the stability and diverse data coverage of diffusion models with the high-fidelity image generation of GANs [2, 46, 49, 51, 53]. A common method employs a diffusion model as the generator and a separate discriminator for adversarial feedback [2, 46, 49]. Although this can improve the level of detail, it also increases memory and computational costs and can reintroduce GAN-specific instabilities such as mode collapse.

In this paper, we propose a novel generation framework that adversarially optimizes each denoising step in the dif-

*Corresponding author.
[1] project page: https://github.com/u-chae/gad/

fusion process without relying on a separate discriminator. Our approach integrates an adversarial loss directly into each denoising stage of the diffusion model, effectively acting as a regularizer. By using a single U-Net as both generator and discriminator, we eliminate memory overhead from a separate discriminator and address common GAN issues. Moreover, this integrated adversarial loss promotes semantic information flow across timesteps, enabling the model to better capture complex data distributions even when training data is limited or biased. Consequently, our method synergistically combines the high-fidelity generation capabilities of GANs with the robust, stable convergence of diffusion models, resulting in high-quality and reliable image synthesis without additional memory costs.

We validate our framework through extensive experiments on standard latent diffusion baselines [36], demonstrating that our adversarial training strategy significantly improves image quality and mode coverage. Furthermore, our approach outperforms existing methods in various tasks, including conditional text-to-image generation [15, 24] and 2D-to-3D generation [22, 26]. These results highlight the efficacy of unifying adversarial and diffusion-based training within a single network.

2. Related Work

2.1. Diffusion Models

Diffusion models, such as Denoising Diffusion Probabilistic Models (DDPM) [19], are probabilistic generative models that learn the data distribution $p(x)$ by reversing a Markov noising process. DDPM achieves high generalization by leveraging large-scale training datasets and has demonstrated state-of-the-art performance on various generation tasks. However, DDPM requires sampling at every step along the Markov chain to generate high-quality images in image space, leading to slow generation speeds and high computational cost.

To address these issues, several efficient sampling methods [28, 29, 45] have been proposed. For example, Denoising Diffusion Implicit Models (DDIM) [45] significantly improve sampling speed by adopting a non-Markovian chain, thereby simplifying the computation between timesteps. Nonetheless, both DDPM and DDIM generate the next sample solely based on the current one, which introduces discretization errors and can destabilize the denoising process [23]. To alleviate these problems, researchers have explored high-order solvers [28, 29]. Despite these advances, diffusion models remain computationally expensive and suffer from low inference speed due to pixel-based learning in image space.

To overcome these limitations, the Latent Diffusion Model (LDM) [36] was proposed. LDM maps the high-dimensional image space to a low-dimensional latent space using an autoencoder, extracting semantic information while significantly reducing training cost. This approach has enabled applications in various fields, including text-to-image [32, 35, 58], text-to-3D [9, 33, 52], image-to-3D [22, 26, 27], and more efficient generation models.

Our proposed method is applicable to both pixel-based diffusion models (*e.g*., Imagen [37]) and latent-based diffusion models (*e.g*., Stable diffusion [36]), independent of the sampling method. In this paper, experiments are conducted using Stable diffusion [36] to verify the general usability, scalability, and memory efficiency of the proposed framework.

2.2. Diffusion with GANs

With the emergence of latent diffusion models [36], which enable efficient computational processing, various studies have focused on training diffusion models with large datasets. However, diffusion models require an iterative denoising process to generate the final sample, resulting in slow generation speeds. Furthermore, diffusion models typically need a large training dataset to learn the complex data distribution for stable convergence. To mitigate these issues, recent research [2, 46, 49, 51, 53] has focused on combining the advantages of GANs and diffusion models to overcome their individual shortcomings.

This research can be broadly divided into two approaches. The first approach leverages adversarial training to improve the sampling speed of diffusion models while maintaining high generation quality [20, 39, 51, 54]. Diffusion models typically require several thousand iterative sampling steps under the assumption that the reverse diffusion distribution can be approximated by a Gaussian distribution when the added noise is small. However, if the noise is large, the reverse distribution becomes a non-Gaussian, multimodal distribution [44], and reducing the number of sampling steps degrades quality. Recent studies [51] address this by modeling the denoising distribution with conditional GANs that can capture multimodal distributions, thereby enabling faster sampling of high-quality images. Yet, their effectiveness is limited in high-dimensional settings, and finding an appropriate noise distribution remains challenging. To overcome this, recent work [49] employs a Gaussian mixture noise process on GANs to combine the high-resolution capabilities of GANs with the fast sampling speed and learning stability of diffusion models.

The second approach aims to improve the quality of diffusion models by incorporating adversarial training into the diffusion process itself. Although conventional diffusion models converge stably and generate diverse images, they often struggle to produce high-resolution images or capture fine details compared to GANs. To address these limitations, methods such as diffusion model distillation using GANs [10, 55] or adversarial learning with an independent discriminator [38, 40, 56] have been proposed. For example, recent work [40] employs large-scale off-the-shelf diffusion models as a teacher signal, combined with adversarial loss to train on generating high-quality samples. Although this approach can produce high-quality samples under limited conditions, it requires the discriminator to operate in image space, resulting in high memory consumption and computational cost. Other methods perform adversarial training in latent space [38], yet scalability remains an issue due to the difficulty in controlling discriminator feedback. An alternative is to improve the diffusion training process by incorporating an independently trained discriminator, such as a structure-guided discriminator [55], which aligns the generated images with the intrinsic structure of the training dataset. However, the simultaneous training of the generator and discriminator inherent to GANs makes it fundamentally challenging to resolve training instability issues such as mode collapse.

In this paper, we propose a novel generation framework that adversarially optimizes each learning step of the diffusion model to address these issues. Our method leverages the U-Net of the diffusion model to serve as both the generator and the discriminator simultaneously, thereby significantly reducing mode collapse probability and computational cost. This enhances stability and performance. Our approach yields improved consistency and quality over recent generation models [15, 22, 24, 26, 36], paving the way for advancements in generative modeling.

3. Methods

3.1. Preliminaries

Energy-based Generative Adversarial Networks. The energy-based Generative Adversarial Networks (GAN) [60] conceptualize the discriminator as an energy function, allowing for a variety of architectural designs and loss functions. This approach facilitates the stabilization of image generation network training by employing an auto-encoder network as the discriminator, which discriminates between real images from the training dataset and synthetic images produced by the generator. This stabilization is achieved by training the auto-encoder network to lower the error between input and output for the real data sample x while maintaining a higher error for the generated samples $G(z)$. With a moderately small value m, the discriminator loss L_D and generator loss L_G are formally defined as

$$L_D = D(x) + [m - D(G(z))]_+ \,, \tag{1}$$

$$L_G = D(G(z)), \tag{2}$$

where $[\cdot]_+ = \max(0, \cdot)$ and m is a positive margin.

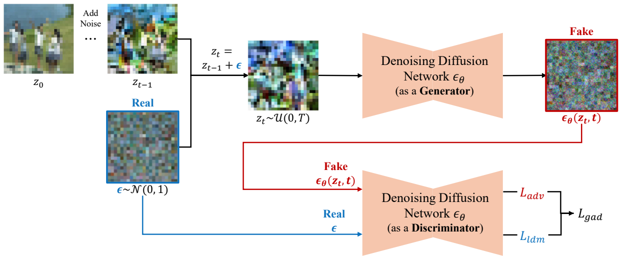

Add Noise
...
z_0
z_{t-1}
Real
$\epsilon \sim \mathcal{N}(0,1)$
$z_t = z_{t-1} + \epsilon$
$z_t \sim \mathcal{U}(0,T)$
Denoising Diffusion Network ϵ_θ (as a **Generator**)
Fake
$\epsilon_\theta(z_t, t)$
Fake
$\epsilon_\theta(z_t, t)$
Real
ϵ
Denoising Diffusion Network ϵ_θ (as a **Discriminator**)
L_{adv}
L_{ldm}
L_{gad}

Figure 1. High-level overview of the proposed GAD framework.

Latent Diffusion Models. In latent diffusion models, a real image x is first encoded into a latent representation $z_0 = \mathcal{E}(x)$ via an encoder $\mathcal{E}$. The diffusion process gradually perturbs z_0 by mixing it with Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ using a noise schedule defined by $\sqrt{\bar{\alpha}_t}$ and $\sqrt{1 - \bar{\alpha}_t}$. A U-Net $\epsilon_\theta(\cdot, t)$ is then employed to predict the added noise, resulting in the training objective

$$L_{ldm} = \mathbb{E}_{\mathcal{E}(x),\, t,\, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon,\, t)\| \right], \tag{3}$$

where $\bar{\alpha}_t$ denotes the cumulative noise scaling factor at time t. This formulation enables efficient training in the latent space, reducing computational cost while preserving essential image details [36].

3.2. Generative Adversarial Diffusion

In our proposed Generative Adversarial Diffusion (GAD) model, we integrate the energy-based GAN framework with latent diffusion models. The key observation is that both methods rely on an auto-encoder-like network: energy-based GANs use an auto-encoder as the discriminator, and latent diffusion models employ a U-Net for noise prediction. However, while conventional autoencoders minimize reconstruction loss, the U-Net in latent diffusion is specifically trained to predict noise, a task that has been shown to yield superior convergence and finer details in generated images. Based on this insight, we replace the traditional reconstruction error with a noise prediction loss and subsequently introduce an adversarial regularizer. This regularizer, which enforces a margin between the noise predictions of real and generated latent representations, is designed under the assumption that the generator and the discriminator share the same network. By embedding the adversarial constraint directly into the latent diffusion loss, our approach eliminates the need for alternating updates between separate networks, thereby enhancing training stability and reducing mode collapse. Fig. 1 shows an overview pipeline of the proposed framework. In particular, we used a shared network for the generator and discriminator. Thus, the network simultaneously minimizes the noise prediction loss for perturbed latent samples and enforces a margin-based separation between real and generated (*i.e.*, denoised) latents, mitigating instability issues such as mode collapse.

It is important to note that this definition is motivated by the intrinsic nature of diffusion models. While diffusion training operates on a single time step, the denoising result from a subsequent time step (*e.g.*, $t+1$) can be mathematically related to that at time t (as established in formulations like DDIM [36]). Practically, using the next step to define the fake sample ensures the adversarial regularizer enforces margin-based separation between the current noisy latent (real) and the denoised latent (fake), enhancing both fidelity and training stability.

We now define the generator and the discriminator in our unified framework. Given a real image x, let $z_0 = \mathcal{E}(x)$ be its latent representation. At diffusion step t, the latent is perturbed as

$$z = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{4}$$

with $\epsilon \sim \mathcal{N}(0, I)$. The shared U-Net $\epsilon_\theta(\cdot, t)$ is employed in two roles:

$$G(z_0, t) = \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\, t), \tag{5}$$

$$D(z, t) = \|\epsilon_\theta(z, t) - \epsilon\|, \tag{6}$$

where $G(z_0, t)$ outputs the predicted (denoised) latent and $D(z, t)$ measures the discrepancy between the predicted noise and the true noise.

Loss Function Definitions. By substituting the definitions of the generator and discriminator functions in (5) and (6) into the energy-based GAN loss expressions in (2) and (1), we obtain the following forms:

$$L_G = \mathbb{E}_{z_0,t,\epsilon}[\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\, t) - \epsilon\|], \tag{7}$$

$$\begin{aligned} L_D = {} & \mathbb{E}_{\mathcal{E}(x),t,\epsilon}[\|\epsilon_\theta(z, t) - \epsilon\|] \\ & + \mathbb{E}_{z_0,t,\epsilon}[[m - \|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\, t) - \epsilon\|]_+]. \end{aligned} \tag{8}$$

From these expressions, we see that the generator's loss in (7) is identical to the latent diffusion loss L_{ldm}, as it minimizes the error between the predicted noise and the true noise. In contrast, the discriminator loss in (8) contains an additional adversarial term that enforces a margin between the noise prediction for fake samples (*i.e.*, the denoised latent from a subsequent time step) and the true noise. This adversarial term acts as a regularizer, ensuring that the shared network not only achieves accurate noise prediction but also maintains a clear separation between real and fake latents.

For adversarial training, the two loss functions in (8) and (7) are typically back-propagated alternately in conventional GAN training–a procedure known to induce instability and mode collapse [5, 16]. However, given that the generator's loss is inherently a subset of the discriminator's denoising loss, we incorporate the adversarial term directly as a regularizer. Thus, the overall loss for our Generative Adversarial Diffusion model is defined as

$$L_{gad} = L_{ldm} + \lambda_{adv}\, L_{adv}, \tag{9}$$

where $L_{adv} = [m - \|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon,\, t) - \epsilon\|]_+$, and λ_{adv} is the balancing weight. In our framework, the latent diffusion loss L_{ldm} treats noisy latents at time step t as real, while the adversarial loss L_{adv} enforces a margin on the denoised latents (from step $t+1$), thereby unifying the training without the need for alternating updates. For a more detailed theoretical analysis, please refer to Sec. **??** in the *supplemental material*.

Overall, the proposed Generative Adversarial Diffusion framework leverages the inherent similarity between energy-based GANs and latent diffusion models. By replacing the reconstruction error with a noise prediction error and unifying the generator and discriminator into a single network, our approach provides an effective adversarial regularizer that not only prevents mode collapse, but also contributes to faster convergence and enhanced detail in generated images.

4. Experiments

In this section, we present our experiments to validate the flexibility and effectiveness of the proposed method. In Sec. 4.2.1, we demonstrate the performance improvement of the model utilizing GAD through a comparison experiment with baseline methods [44]. In addition, in Sec. 4.2.2 and in Sec. 4.2.3, we apply the proposed method to state-of-the-art methods of various generation tasks using diffusion models, such as conditional text-to-image generation [15, 24] and 2D-to-3D generation tasks [22, 26], to verify the significantly improved mode coverage ability.

4.1. Experimental Details

In all experiments, both the baseline methods and the comparative experiments for each generation task were trained using the same number of steps and with the same number of parameters to ensure a fair comparison. Specifically, we used the publicly available Stable diffusion 2.1 [36] model as the base diffusion architecture for all setups.

Also, we set the margin m by performing inference on the training set every 5 epochs and computing the average error. For detailed information on the rationale behind our margin selection, please refer to Sec. **??** in the *supplemental material*. Furthermore, we set $\lambda_{adv} = 0.01$ in all experiments, which shows the best empirical performance. A sensitivity analysis of the adversarial regularization weight λ_{adv} is provided in Sec. **??** of the *supplemental material*.

4.2. Performance Comparisons

4.2.1. Text-to-image Generation

To validate the benefits through adversarial training of the diffusion model, we compared the performance of Stable diffusion [36] with and without the proposed method. Following related studies [6, 58], to compare the perfor-

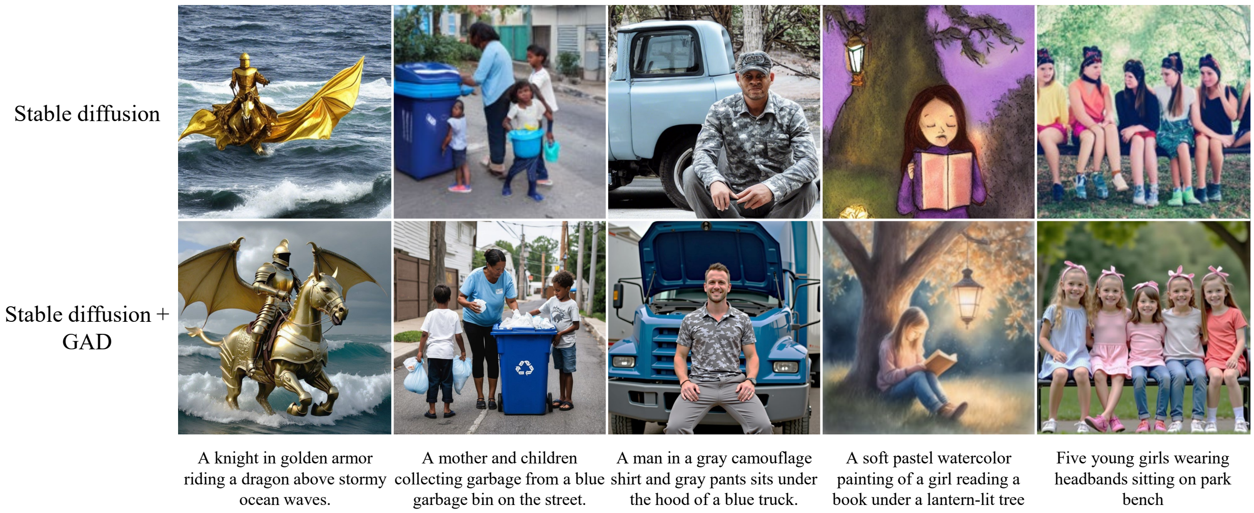

Stable diffusion
Stable diffusion + GAD
A knight in golden armor riding a dragon above stormy ocean waves.
A mother and children collecting garbage from a blue garbage bin on the street.
A man in a gray camouflage shirt and gray pants sits under the hood of a blue truck.
A soft pastel watercolor painting of a girl reading a book under a lantern-lit tree
Five young girls wearing headbands sitting on park bench

Figure 2. Qualitative comparisons with Stable diffusion.

Table 1. Quantitative comparisons with Stable diffusion.

Method	FID (↓)	CLIP score (↑)
Stable diffusion	13.52	0.3143
Stable diffusion + GAD (Ours)	**9.68**	**0.3471**

mance of image generation, we used Fréchet Inception Distance (FID) [17], which measures the distributional similarity between real and generated images, and Contrastive Language-Image Pre-training (CLIP) score [34], a normalized metric for evaluating the similarity between CLIP text-image embeddings. For a fair comparison of the experiments, both Stable diffusion with and without GAD were trained on the same LAION-5B [41] datasets. For evaluation, we randomly selected COCO2014 validation set [57] datasets.

Results with Stable Diffusion. The performance measurements in terms of FID and CLIP score are summarized in Table 1. It is shown that Stable diffusion combined with our proposed method (GAD) outperforms the baseline. In particular, the FID is significantly lower with GAD, indicating that the embedding of the generated images is more similar to the real dataset. Therefore, these results demonstrate that applying adversarial training in the training process of the diffusion model enables more effective learning of the data distribution and improves the quality of the generated images.

Figure 2 presents qualitative comparisons. Stable diffusion with GAD has been shown to generate images of higher quality than the standard Stable diffusion [36]. When the proposed method is applied, high-frequency details are better preserved for realistic and complex captions. For instance, in the second and fifth columns of Fig. 2, the images generated by the proposed method appear sharper and depict faces and body details more accurately than those produced by the baseline. In the fourth image, the proposed method produces samples that better capture the 'soft pastel' style. Furthermore, in the first and third images, the background and objects described in the captions are integrated more naturally. These observations confirm that the incorporation of GAD into the diffusion model enables a richer and more detailed representation of the generation.

4.2.2. Conditional Text-to-Image Generation

To validate the extensibility of GAD, we compared the performance of recent conditional text-to-image methods [15, 24] with and without GAD. Here, we used the state-of-the-art conditional diffusion models, GLIGEN [24] and Textual Inversion [15], as baseline models.

GLIGEN generates images using additional conditions, such as bounding boxes and key points, in addition to the standard text caption. We employed GLIGEN with bounding boxes to verify how effectively adversarial training can learn these conditions. For a fair comparison, we used the same COCO2014 [25] grounding dataset. For evaluation, we randomly selected 2K image-text pairs from CC3M [8] and obtained bounding boxes through Grounded Language-Image Pre-training (GLIP) [21]. We then measured performance using the FID [17] and CLIP score [34] score metrics described in Sect. 4.2.1.

Textual Inversion is a technique for learning word embeddings that represent novel concepts in text-to-image models. It leverages a small set of images to optimize a diffusion model, thereby learning embedding vectors that capture these new concepts. Once trained, the embedding vectors can be inserted into new scenes or applied to tasks such as style conversion. To verify the effectiveness of our proposed method in converging on a small dataset, we compared Textual Inversion [15] with and without GAD.

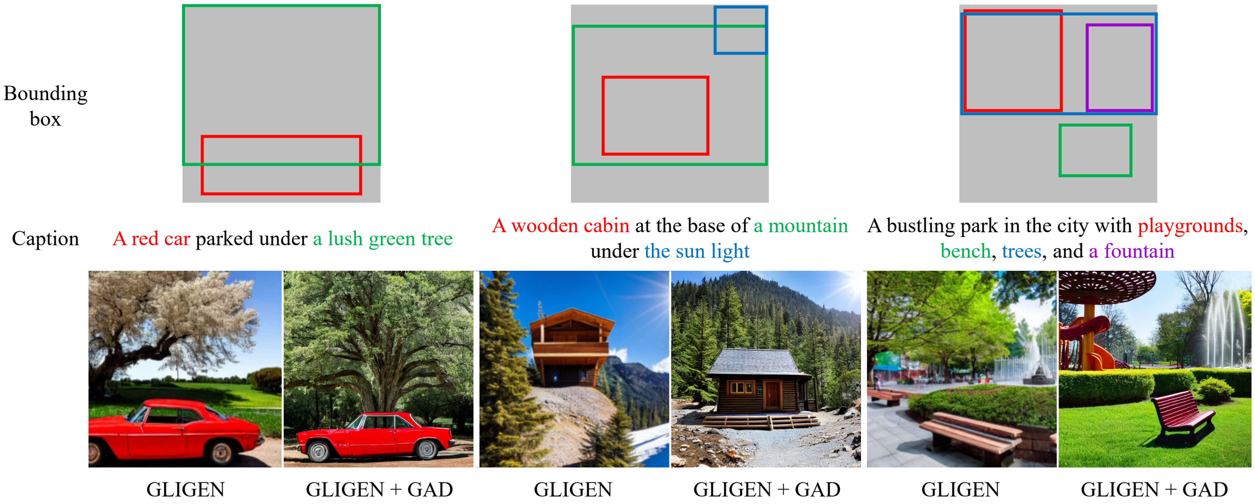

Bounding box
Caption
A red car parked under a lush green tree
A wooden cabin at the base of a mountain under the sun light
A bustling park in the city with playgrounds, bench, trees, and a fountain
GLIGEN
GLIGEN + GAD
GLIGEN
GLIGEN + GAD
GLIGEN
GLIGEN + GAD

Figure 3. Qualitative comparisons with GLIGEN.

Table 2. Comparisons with conditional text-to-image methods.

Method	FID (↓)	CLIP score (↑)	CLIP similarity (↑)
GLIGEN	12.26	0.292	-
GLIGEN + GAD (Ours)	**9.12**	**0.311**	-
Textual Inversion	13.74	-	0.712
Textual Inversion + GAD (Ours)	**8.71**	-	**0.764**

For a fair comparison, we used the same Google Scanned Object (GSO) [12] multi-view image dataset, selecting six 512×512 images as input. We then evaluated the methods using random text that was not included in the training set. Following previous work [1, 15], we employed FID and CLIP similarity [14], the cosine similarity between CLIP image embeddings, to measure how closely the generated images resembled the input images.

Table 2 shows the performance of GLIGEN in terms of FID [17] and CLIP score [34]. The results indicate that adding GAD leads to notable improvements. In particular, the FID significantly decreases when GAD is applied, indicating that adversarially training the diffusion model integrates the text and bounding box conditions more effectively, resulting in a distribution closer to that of the real dataset. Following the previous studies [1, 15], we used FID and CLIP-similarity [14], which is the cosine similarity between CLIP image embeddings to measure similarity with input images and generated images.

Results with GLIGEN. The comparison results are illustrated in Fig. 3. By employing the proposed method, bounding boxes and text captions are combined more naturally, with bounding box positions maintained accurately while generating high-resolution images. For example, in the first sample of Fig. 3, our approach accurately reflects the caption "lush green tree" and produces a sharper image. In the second sample, our method can capture the bounding boxes for "the cabin" and "a mountain" more precisely, representing details, such as the wooden cabin's texture and the mountain's foliage, more faithfully than the baseline. In the third sample, where multiple bounding boxes overlap, GLIGEN struggles to depict "playground" correctly when its bounding box overlaps with "trees," whereas our approach properly arranges the foreground and background to incorporate all given conditions with clarity. These findings demonstrate that GAD contributes to the generation of better images that effectively combine text and bounding-box information in a conditional diffusion setting.

We observed that GAD's benefits become more pronounced as the number of bounding boxes or key points increases. This suggests that adversarial regularization helps the diffusion model manage multiple constraints while maintaining object boundaries and distinct features. This likely prevents the collapse of individual object representations. When evaluating more elaborate scenes involving three or more bounding boxes, our method consistently maintained object boundaries and distinct features, whereas the baseline often showed partial overlaps or blurred transitions between adjacent objects.

Results with Textual Inversion. The comparison performance results of Textual Inversion are also represented in Table 2 in terms of FID and CLIP similarity. Textual Inversion with the proposed method is shown to outperform the baseline method, indicating that our method can be effec-

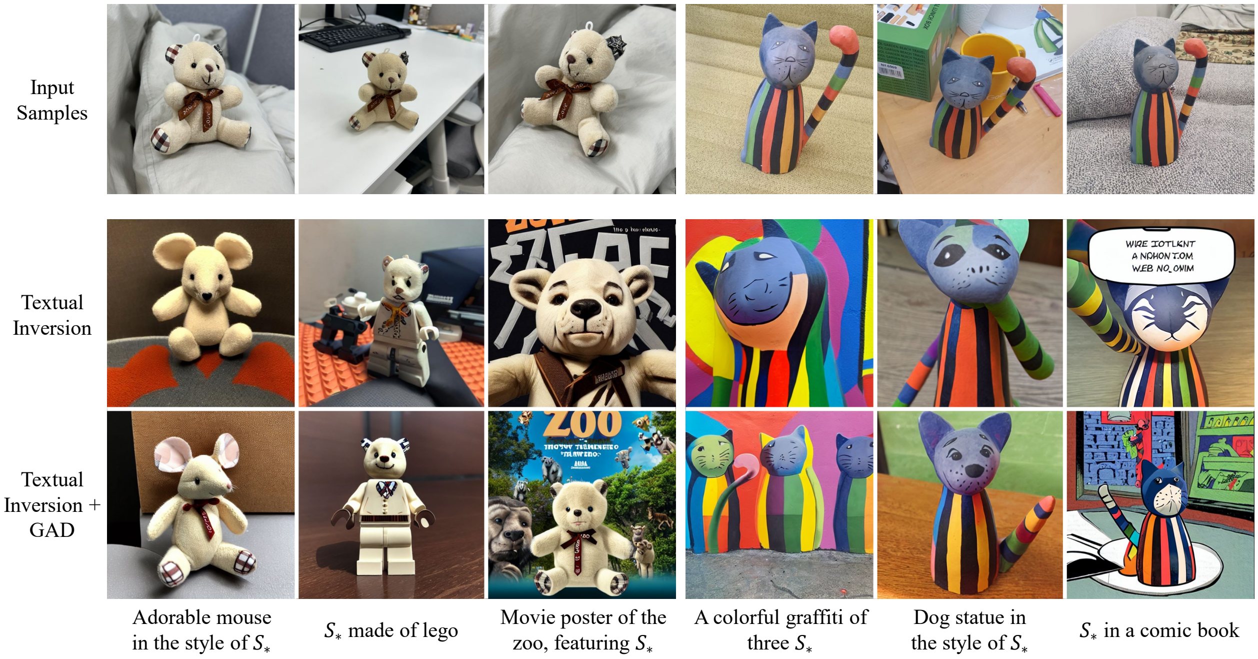

Input Samples
Textual Inversion
Textual Inversion + GAD
Adorable mouse in the style of S_*
S_* made of lego
Movie poster of the zoo, featuring S_*
A colorful graffiti of three S_*
Dog statue in the style of S_*
S_* in a comic book

Figure 4. Qualitative comparisons with Textual Inversion.

tively extended to conditional diffusion models to produce more natural images through appropriate training of conditions and text. Figure 4 illustrates the results. Textual Inversion with GAD yields higher output diversity and better image generation quality than the version without GAD. For instance, in the first sample of Fig. 4, our approach learns complex teddy bear patterns, such as ears and feet, more efficiently, accurately reflecting them across various generated styles. In the fifth column, the proposed method balances the given caption with the input sample, demonstrating that GAD stably learns the concept and generates images consistently in different styles.

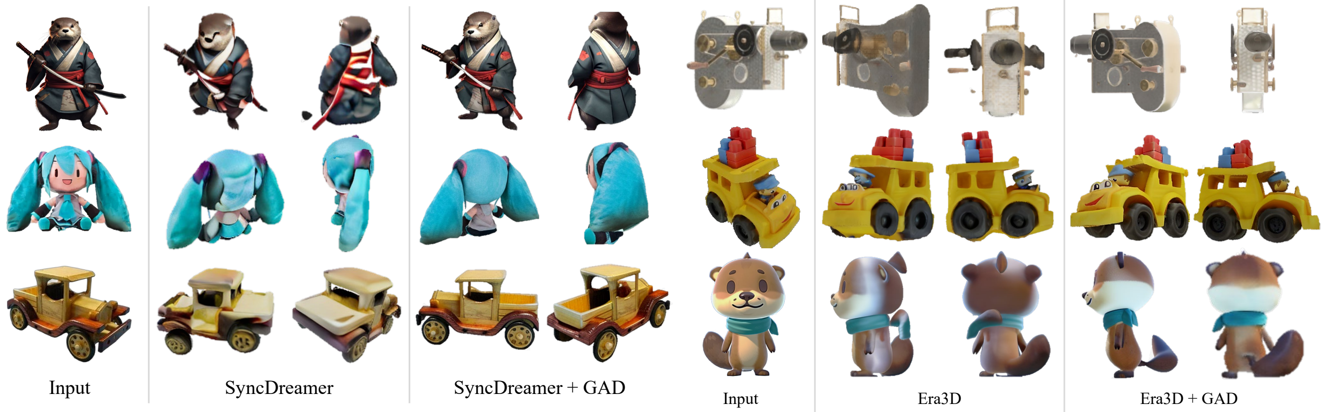
These observations suggest that adversarial training on a small dataset is more effective than relying solely on traditional noise-based training. With limited data, the diffusion model faces a higher risk of overfitting, which can reduce the diversity of generated images. However, incorporating GAD allows adversarial loss to act as a regularizer, stabilizing model training while boosting output diversity. Additionally, adversarial loss promotes information exchange across time steps, enabling better convergence under data-scarce conditions.

4.2.3. 2D-to-3D Generation

To validate the extensibility of the proposed method, we compared the performance of recent 2D-to-3D methods [22, 26] with and without GAD. In these experiments, we used SyncDreamer [26] and Era3D [22], as baseline methods, which represent state-of-the-art 2D-to-3D diffusion models.

SyncDreamer employs a synchronized multi-view diffusion model to ensure consistency across multiple views by processing each view simultaneously through a shared noise estimator and an attention mechanism. Era3D addresses camera prior mismatches and computational inefficiencies found in existing multi-view diffusion methods [42, 43, 47] by incorporating a diffusion-based camera prediction module and row-wise attention, thereby producing consistent multi-view images. For training all methods, we used a subset of the Objaverse [11] dataset. SyncDreamer with and without GAD was trained using 16 uniform views, each with a fixed elevation of 30° and a randomly sampled elevation. Similarly, to compare Era3D with and without GAD, we trained the model on 16 uniform views at a fixed elevation of 30°, also incorporating normal and depth maps from each view. For evaluation, we used 3D data from the Google Scanned Object (GSO) [12] and OmniObject3D (Omni3D) [50] datasets. To demonstrate the generalizability of the model, we randomly selected 50 objects from various categories, including everyday items and animals. For the 2D-to-3D generation task, we adopted three standard metrics to assess novel view synthesis performance: Peak Signal-to-Noise Ratio (PSNR) [48], Structural Similarity Index Measure (SSIM) [48], and Learned Perceptual Image Patch Similarity (LPIPS) [59].

Results with SyncDreamer. Table 3 summarizes the performance of SyncDreamer and Era3D in terms of PSNR, SSIM, and LPIPS. The results show that the use of the proposed method leads to notable performance improvements. In particular, when GAD is applied, the SSIM and LPIPS scores outperform those of the baseline methods, indicating that adversarial training enhances perceptual similarity in



(a) Performance comparison with SyncDreamer.

(b) Performance comparison with Era3D.

Figure 5. Qualitative comparisons with 2D-to-3D state-of-the-art methods.

Table 3. Comparison with 2D-to-3D state-of-the-art methods.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
SyncDreamer	20.25	0.798	0.146
SyncDreamer + GAD (Ours)	**22.81**	**0.858**	**0.119**
Era3D	22.74	0.837	0.126
Era3D + GAD (Ours)	**24.12**	**0.891**	**0.102**

multi-view images beyond basic reconstruction capabilities in 3D generation.

Figure 5 provides visual comparisons of the novel view synthesis. As illustrated in Fig. 5a, SyncDreamer exhibits insufficient view consistency and does not generate visually detailed images. For instance, in the first row of Fig. 5a, the novel views generated by our method appear sharper than those produced by SyncDreamer, capturing high-frequency details such as sword blades or clothing patterns more effectively. In the second row, our method consistently renders the character's hair and outfit from the rear view, and in the third row, it maintains coherent wheel shapes across different viewpoints.

Results with Era3D. Likewise, Fig. 5b shows that our method produces sharper and higher frequency details compared to Era3D. For example, in the first row, the proposed approach preserves fine details, such as handles or chains, as well as the object's color and overall geometric structure when generating novel views. In the second row, the novel views produced by the proposed method exhibit more consistent color representation of blocks and continuity in wheel size and shape. These findings show that our method captures fine details effectively and generalizes well to diverse generation tasks.

Additionally, we observed that GAD particularly benefits object boundaries when transitioning between different viewpoints. In several instances, such as the figure in the second row of Fig. 5a, our method maintained smoother object contours and more coherent color transitions than the baseline, indicating that adversarial regularization helps enforce consistency across angles. This was especially evident for objects with reflective or metallic textures, where slight inconsistencies in the diffusion process can lead to visually distracting artifacts in multi-view renderings.

5. Conclusion

Summary. In this work, we introduced Generative Adversarial Diffusion (GAD), a novel framework that seamlessly integrates diffusion models with adversarial training using a unified network for both generation and discrimination. Our approach leverages the inherent strengths of diffusion models, such as stable convergence and robust data coverage, while mitigating common GAN pitfalls like mode collapse. As a result, GAD achieves outstanding image synthesis quality compared to baseline methods.

Limitation and Future Work. Our unified framework offers significant benefits but has trade-offs. One limitation is that coupling the generator and the discriminator into a single network may restrict the flexibility to optimize each component independently. Although this design choice reduces memory overhead and simplifies training, it might also constrain the fine-tuning of individual contributions from generation and discrimination. In future work, we will refine the diffusion process by leveraging the insights gained from our adversarial regularization approach.

Acknowledgements. This research was supported by the Ministry of Education's 4th phase of the BK21 project (Grant No. 4120240215083) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00229451, Interoperable Digital Human (Avatar) Interlocking Technology Between Heterogeneous Platforms, 100%).

References

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics*, 42 (6):1–10, 2023.

[2] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2017.

[3] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2017.

[4] Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 214–223, 2017.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. In *Proceedings of the International Conference on Machine Learning*, 2017.

[6] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. EDIFF-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[7] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Remi Munos. The cramer distance as a solution to biased wasserstein gradients. In *Proceedings of the International Conference on Learning Representations*, 2018.

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[9] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22246–22256, 2023.

[10] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. AdvDiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023.

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *Proceedings of the International Conference on Robotics and Automation*, 2022.

[13] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *Proceedings of the International Conference on Learning Representations*, 2018.

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):1–13, 2022.

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the International Conference on Learning Representations*, 2023.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, 2014.

[17] 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. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[19] Jonathan Ho, Aravind Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 6840–6851, 2020.

[20] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *Proceedings of the International Conference on Learning Representations*, 2024.

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, and et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[22] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, and et al. Era3D: High-resolution multiview diffusion using efficient row-wise attention. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[23] Shengmeng Li, Luping Liu, Zenghao Chai, Runnan Li, and Xu Tan. ERA-Solver: Error-robust adams solver for fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2301.12935*, 2023.

[24] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22511–22521, 2023.

[25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision*, 2014.

[26] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *Proceedings of the International Conference on Learning Representations*, 2023.

[27] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, and et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.

[28] 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. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5775–5787, 2022.

[29] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. In *Proceedings of the International Conference on Learning Representations*, 2023.

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the International Conference on Machine Learning*, pages 3481–3490, 2018.

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

[32] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning aadapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024.

[33] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2D diffusion. In *Proceedings of the International Conference on Learning Representations*, 2023.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[35] Aditya Ramesh et al. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

[36] 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*, 2022.

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, and et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 36479–36494, 2022.

[38] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024.

[39] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of the European Conference on Computer Vision*, pages 87–103. Springer Nature Switzerland, 2024.

[40] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Proceedings of the European Conference on Computer Vision*, pages 87–103, 2024.

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 25278–25294, 2022.

[42] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, and et al. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

[43] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *Proceedings of the International Conference on Learning Representations*, 2024.

[44] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Learning Representations*, pages 2256–2265, 2015.

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021.

[46] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. In *Proceedings of the International Conference on Learning Representations*, 2017.

[47] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3D content creation. In *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2024.

[48] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs

with diffusion. In *Proceedings of the International Conference on Learning Representations*, 2023.

[50] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[51] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *Proceedings of the International Conference on Learning Representations*, 2022.

[52] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023.

[53] Yanwu Xu, Mingming Gong, Shaoan Xie, Wei Wei, Matthias Grundmann, Kayhan Batmanghelich, and Tingbo Hou. Semi-implicit denoising diffusion models (siddms). In *Proceedings of the Advances in Neural Information Processing Systems*, page 17383, 2024.

[54] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. UFOGen: You forward once large scale text-to-image generation via diffusion GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8196–8206, 2024.

[55] Ling Yang, Haotian Qian, Zhilong Zhang, Jingwei Liu, and Bin Cui. Structure-guided adversarial training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2024.

[56] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 47455–47487, 2025.

[57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[59] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[60] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2022.

[61] Huangjie Zheng and Mingyuan Zhou. Exploiting chain rule and bayes' theorem to compare probability distributions. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.