Text-to-3D Generation using Jensen-Shannon Score Distillation

Khoi Do Binh-Son Hua

Trinity College Dublin, Ireland {dokh, binhson.hua}@tcd.ie

Abstract

Score distillation sampling is an effective technique to generate 3D models from text prompts, utilizing pre-trained large-scale text-to-image diffusion models as guidance. However, the produced 3D assets tend to be oversaturated, oversmoothed, and have limited diversity. These issues are a result of a reverse Kullback–Leibler (KL) divergence objective, which makes the optimization unstable and results in mode-seeking behavior. In this paper, we derive a bounded score distillation objective based on Jensen-Shannon divergence (JSD), which stabilizes the optimization process and produces high-quality 3D generation. JSD can match the generated and target distributions well, therefore mitigating mode seeking. We provide a practical implementation of JSD by utilizing the theory of generative adversarial networks to define an approximate objective function for the generator, assuming the discriminator is well-trained. By assuming the discriminator follows a log-odds classifier, we propose a minority sampling algorithm to estimate the gradients of our proposed objective, providing a practical implementation for JSD. We conduct both theoretical and empirical studies to validate our method. Experimental results on T3Bench demonstrate that our method can produce high-quality and diversified 3D assets. The code is available at https://github.com/KhoiDOO/jsddreamer.

1. Introduction

Text-to-3D generation has become an impactful and leading research field in computer vision, contributing to various applications. Creating high-quality 3D content with view consistency and diversity is resource-intensive, making automated 3D generation a crucial research goal. The developments in the neural radiance field [48, 91], Gaussian splatting [32, 37], and multimodal latent diffusion models (LDM) [58, 92], have driven substantial advancements in generating imaginative 3D content from text prompts [40, 44].

A straightforward but expensive approach to generating 3D assets given a single text prompt is to train a large-scale generative model on a large-scale 3D shape dataset. Another approach is to learn 3D assets by distilling from a large pre-trained model. Score distillation sampling (SDS) utilizes a pre-trained model to learn a neural network (particle), which can synthesize different views of an object [56]. SDS optimizes a KL divergence [68] (KLD) between Gaussian distributions in the forward and backward processes in LDM. Owing to the integration of a 2D pre-trained model, optimizing the KLD leads to an improved 3D representation with more consistent views.

Nevertheless, as highlighted in [44, 78, 80], SDS tends to produce 3D assets that are oversaturating, oversmoothing, and lack diversity. Other existing approaches aim to produce higher quality 3D assets by utilizing variational score distillation [78, 80], multi-stage training strategies [11, 31, 45]. However, those methods are highly costly in computation, which requires fine-tuning pre-trained models or performing mesh extraction or texture fine-tuning.

Mode-seeking in SDS can be attributed to the asymmetry of the KLD [68], which matches the Gaussians in the forward process to modes of the score functions in a diffusion process [56]. To mitigate this problem, one potential idea is to symmetrize the objective function, as inspired by the literature of Generative Adversarial Networks (GANs) [3, 21, 22]. We propose to use Jensen-Shannon divergence [41] (JSD), a bounded divergence, as our objective function. Multiple variants of JSD exist, but it remains challenging to perform optimizations using JSD as it requires estimating a mixture of the probability density in JSD. Some variants of JSD allow such estimation, such as Geometric JSD, the mixture density of which can be logarithmically derived. However, this objective is unbounded and, therefore, unstable to optimize. To implement JSD in practice, we aim to approximate JSD instead. Our key insight is to define a discriminator based on log-odds classifiers [7] and use minority sampling [73] to perform distillation with an approximated JSD derived via GAN criterion. We show that this technique can well approximate JSD, leading to score distillation with improved stability, enabling the optimization to converge to different modes on the latent manifold.

To support our theoretical derivations, we conduct empirical analysis on a toy dataset by training a toy diffusion model

Figure 1. Our method can improve quality and diversity generation with multiple seeds starting from a fixed initialization point.

and performing score distillation toward a specific cluster. The obtained results illustrate that the gradient of our proposed method is more stable than SDS owing to an effective control variate, which is positively correlated with the esti-

mated noise. These experiments also show that our method can enhance the diversity of the generated 3D objects.

We evaluate our method on the recently introduced T3Bench [24], including a wide range of prompts for the 3D generation task. Our quantitative comparisons with state-of-the-art text-to-3D methods show that our method can generate high-quality 3D assets with a strong alignment with the given prompt. Our contribution can be summarized as follows:

- We use Jensen-Shannon divergence (JSD) for score distillation and estimate JSD by leveraging GAN theories and a minority sampling technique.
- We validate our theory with empirical experiments, training a toy diffusion model on a toy Gaussian dataset and comparing SDS with our method, confirming the stability and diversity of the generated samples.
- We conduct evaluations on common benchmarks and show that our proposed method can generate high-fidelity and diverse 3D assets.

2. Related Works

Text-to-image generation. Early text-to-image generation is based on Generative Adversarial Networks (GAN) [21], where images are generated conditioned on a textual inputs. The objective criterion in ordinary GAN is Jensen-Shannon Divergence [41] (JSD), approximated by a min-max game optimization. This approach results in mode collapse due to the discontinuous region between the generator and discriminator distribution [2]. To address this problem, alternative divergences [3, 22, 54, 81], multiple generators [4, 6, 12, 20, 28, 55], manifold learning [15, 42, 50], and score matching [79, 84] are proposed to match the fake and real distributions. Particularly, approaches that bridge ordinary GAN and score generative models [27, 62] achieve SOTA results by diffusing all data points to the same manifold, thus making JSD continuous everywhere [79], while metric-based distance is not always converged [46, 82]. Diffusion models [58, 60, 62, 92], otherwise learn the relationship between image and textual distribution via a stochastic denoising process, improving high fidelity and diversified generation. Recently, there has been research aiming to improve diversity by guiding the estimated score toward low-density regions in diffusion models [33, 72, 73].

Text-to-3D generation. 3D content can be generated by several methods. Based on feed-forward inference, 3D content can be generated by a reconstruction model trained on large-scale datasets [39, 57, 66, 74, 85, 86], which comes at a cost of extensive annotated data and computational resources. Distillation-based methods, otherwise, optimize a 3D representation to learn an asset aligned with the text prompts from a pre-trained text-to-image model [30, 56]. However, these methods are per-prompt optimization, thus requiring a large amount of resources in time and computation. Amortized optimization [43, 45] trains a unified model on many prompts and 3D asset pairs. Besides, score distillation methods also focus on improving quality [40, 44, 70, 78, 80, 93], view consistency [9, 31, 36, 45, 59], diversity [44, 70, 78, 80], and faster distillation [19] of 3D asset generation. Recently, Adversarial Score Distillation [80] (ASD) bridges Variational Score Distillation [78] (VSD) and GAN theory to perform score distillation based on Wasserstein Probability Flow via alternative training. However, the ℓ_1 transport cost, which restricts the discriminator (e.g., LORA [29]) to be 1-Lipschitz [34], is shown not to always converge [46, 82]. This circumstance leads to uncontrollable artifacts and low-quality features in their results. Our approach also leverages GAN theories relevant to Jensen-Shannon divergence, enabling the generated and target distributions to lie on the same support space [79, 84], which returns more stable gradients for 3D generation. Our method is also related to variance reduction techniques for score distillation using control variates [64, 76, 77, 88], but our derivation is via the JSD objective and GAN theories.

3. Backgrounds

3.1. Score Distillation Sampling

Score distillation sampling (SDS) [56] has shown great promise in text-to-3D generation by distilling pre-trained large-scale text-to-image diffusion models. SDS optimizes a 3D model parameterized by $\theta \sim p(\Theta)$ by score distillation gradients derived from a large pre-trained model [58, 92]. Particularly, given text prompt y and the rendered image $\hat{x}_0 = g(\theta, c)$, where g and c are render function and camera pose, the SDS loss function can be written as:

$$\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\Big[w(t)\frac{\sigma_t}{\alpha_t}\text{KL}(q(\hat{x}_t|\hat{x}_0)\|p_\psi(\hat{x}_t|y)\Big]. \quad (1)$$

The gradient estimated through the score model is shown in Eq. (2), where $\hat{\epsilon}_\psi(\hat{x}_t, y)$ and ϵ are the estimated noise and the control variate.

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\Big[w(t)(\hat{\epsilon}_\psi(\hat{x}_t, y) - \epsilon)\frac{\partial \hat{x}_0}{\partial \theta}\Big]. \quad (2)$$

It is commonly known that SDS often suffers from over-saturation, over-smoothing, and low-diversity problems. Previous work [47, 78, 80] attributed these phenomena to the objective based on the reverse KLD. It is therefore beneficial to explore variants of KLD to seek improved convergence and stability in the generation process.

3.2. KL Divergence Symmetrization

KL Divergence [68] (KLD) or relative entropy is the most fundamental distance. KLD is an asymmetric distance (i.e.,

$\mathrm{KL}(p,q) \neq \mathrm{KL}(q,p)\ \forall p,q$), which is unbounded and may be infinite.

$$\mathrm{KL}(p,q) = \sum p \log(p/q), \tag{3}$$

where p and q are two arbitrary distributions. Jeffreys Divergence [51] (JD) is a symmetric divergence combining forward and reverse KLD:

$$\mathrm{JD}(p,q) = \sum \Big[p \log(p/q) + q \log(p/q) \Big]. \tag{4}$$

However, due to the upper-unbounded and numerical instability characteristics, optimizing JD objectives is challenging. Another popular symmetrization of the KLD is the Jensen-Shannon Divergence [41] (JSD), which can be defined as follows:

$$\mathrm{JSD}(p,q) = \frac{1}{2} \sum \Big[p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q} \Big]. \tag{5}$$

JSD is naturally lower-bounded and upper-bounded within $(0, \log_b 2)$ whose base is b. In the literature of generative adversarial networks, JSD was used to learn diversified generators [3, 21]. Given the boundedness of JSD, we aim to utilize JSD for our 3D generation. A visualization of all divergences is provided in the supplementary material.

4. Methodology

4.1. Jensen-Shannon Divergence Distillation

We propose to use JSD as the objective function for text-to-3D generation:

$$\mathcal{L}_{\mathrm{JSD}} := \mathbb{E}_{t,\epsilon} \Big[w(t) \frac{\sigma_t}{\alpha_t} \mathrm{JSD}(q(\hat{x}_t|\hat{x}_0) \| p_\psi(\hat{x}_t|y) \Big] \tag{6}$$

JSD differs from reverse KL in the following properties.

Boundedness. KLD is unstable and less robust to noise due to its unboundedness [17, 67]. KLD can reach extreme values, producing unstable gradients during training, thus preventing θ from converging to an optimal solution [75]. JSD [41], otherwise, has a bounded and thus more numerically stable loss landscape than KLD, which stabilizes the training procedure [1], encouraging θ to reach an optimal solution [35]. We provided a boundedness analysis in the supplementary material, which shows that JSD is in fact a lower bound of the reverse KL objective:

$$\mathcal{L}_{\mathrm{JSD}} \leq \mathbb{E}_{t,\epsilon} \Big[w(t) \frac{\sigma_t}{\alpha_t} \mathrm{KL}(q(\hat{x}_t|\hat{x}_0) \| p_\psi(\hat{x}_t|y) \Big].$$

Mode coverages. SDS faces a problem of low-diversity generation due to mode collapses from the reverse KLD formulation [78, 80]. Unlike reserve KLD, JSD can deal with null mass probability and satisfies the triangle inequality [16], therefore being a good metric distance to match two insignificantly non-overlapping distributions. JSD is, therefore, a potential divergence to mitigate mode collapses, enabling the learning of diversified 3D representations.

Challenges. Despite JSD's benefits, applying JSD for score distillation remains a challenge because sampling from the mixture distribution in JSD is not straightforward, as it is the arithmetic mean $(q(\hat{x}_t|x) + p_\psi(\hat{x}_t|y))/2$ that cannot be logarithmically derived. An alternative way is to use geometric mean to perform mixture distribution [14, 52, 53] $\sqrt{q(\hat{x}_t|x) p_\psi(\hat{x}_t|y)}$, which can be derived easily by the logarithmic function. However, this geometric JSD is unbounded, being an upper-bounded version of the ordinary JSD [52]. In this paper, we instead derive a new objective approximating JSD for score distillation.

4.2. A Discriminator-based Objective

We leverage the GAN [2, 21] training strategy to approximate JSD. In GAN learning theory, there are two steps of training, including training a discriminator $\mathcal{D}$ (a binary classifier) and a generator $\mathcal{G}$, which are both parameterized models. This learning process is a minimax two-player game:

$$V(\mathcal{G},\mathcal{D}) = \int_x \Big[p_{data}(x) \log \mathcal{D}(x) + p_{\mathcal{G}}(x) \log(1 - \mathcal{D}(x)) \Big] dx. \tag{7}$$

When $\mathcal{D}$ reaches the optimal solution, the criterion $V(\mathcal{G},\mathcal{D})$ with respect to $\mathcal{G}$ is equivalent to a JSD objective function [2, 21]. Following the vanilla GAN [21], to train the generator, instead of minimizing $\log(1 - D(x))$, we minimize $-\log D(x)$, resulting in the following objective:

$$\mathcal{L}_{\mathcal{G}}(\theta) = \mathbb{E}_{t,\epsilon} \Big[-\log D(\hat{x}_t; y) \Big] \tag{8}$$

To define the discriminator, we drew inspiration from the fact that our pre-trained text-to-image diffusion model can be regarded as a robust classifier [10, 13, 69], making the model naturally a discriminator. We assume that our discriminator is optimal, and formulate it such that the discriminator follows a log-odds binary classifier [7]:

$$\mathcal{D}(\hat{x}_t; y) = \frac{p_\psi(y|\hat{x}_t)}{1 - p_\psi(y|\hat{x}_t)}, \tag{9}$$

where $p_\psi(y|\hat{x}_t)$ is the likelihood of $\hat{x}_t$ classified as text prompt y. In this discriminator, we note that the term $(1 - p_\psi(y|\hat{x}_t))$ cannot be directly computed. For simplicity, we denote this term by a density function $p(\phi|\hat{x}_t) \approx 1 - p_\psi(y|\hat{x}_t)$. We assume the input prompt $y \in \Omega$ (space of all prompts). Given a prompt y, we assume multiple solutions $\hat{x}_t$, aligned to y due to diversity. Such solutions can be captioned by an extended set of prompts $\mathcal{Y}$ such that $y \in \mathcal{Y} \subset \Omega$. The complementary prompt ϕ can be defined by $\phi \in \Omega \setminus \mathcal{Y}$, capturing irrelevant descriptions of $\hat{x}_t$. We

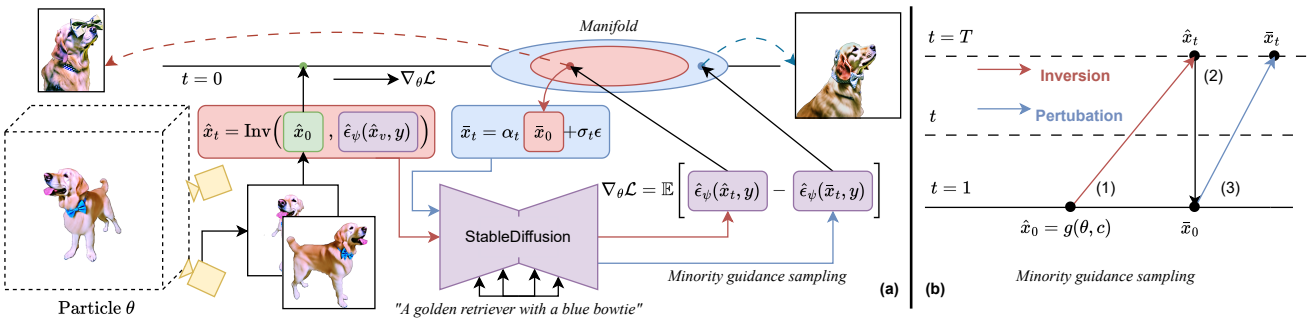


Figure 2. **a) Methodology overview**. Initially, an image $\hat{x}_0$ is generated via a render function $g(\theta, c)$. We obtain a noisy latent of common mode $\hat{x}_t$ conditioned on a text prompt y (■) by using the inversion technique. To gather low-density (■) samples, we acquire $\bar{x}_0$ via the reverse process, then diffuse it by using SDE. The estimated score of both high- and low-density samples guides θ toward convergence. **b) Minority guidance sampling**. The process of obtaining minority sample $\bar{x}_t$ in three steps: (1) DDIM inversion to map $\hat{x}_0$ to $\hat{x}_t$, (2) reverse sampling from $\hat{x}_t$ to obtain $\bar{x}_0$, and (3) a random diffusion to obtain $\bar{x}_t$.

define such irrelevance between y and ϕ via a complement density $p(\phi|\hat{x}_t) = 1 - p(y|\hat{x}_t)$. Expanding the logarithm function and taking the derivative, the gradient of our objective function becomes:

$$\nabla_\theta \mathcal{L}_\mathcal{G} = \mathbb{E}_{t,\epsilon}\Big[\nabla_\theta \log p_\psi(\phi|\hat{x}_t) - \nabla_\theta \log p_\psi(y|\hat{x}_t)\Big]. \quad (10)$$

Let us now proceed to relate this gradient to the score function of the pre-trained text-to-image model below.

4.3. Gradient Approximation

Let us proceed to derive each term in the gradient. The right term can be factorized using the Bayesian theorem, such that $\nabla_\theta \log p_\psi(y|\hat{x}_t) \propto \nabla_\theta \log p_\psi(\hat{x}_t|y) - \nabla_\theta \log p_\psi(\hat{x}_t|\oslash)$. The left term can be derived via multiclass generalization of the logistic sigmoid [7] to yield:

$$\nabla_\theta \log p_\psi(\phi|\hat{x}_t) \propto \nabla_\theta[\log p_\psi(\bar{x}_t|y) - \log p_\psi(\bar{x}_t|\oslash)]. \quad (11)$$

This formulation means that instead of using the prompt ϕ and the noised image $\hat{x}_t$ for estimating the gradient, we estimate a minority sample $\bar{x}_t$ from $\hat{x}_t$, such that $p(\phi|\hat{x}_t) \approx p(y|\bar{x}_t)$ and therefore we use the same prompt y to approximate the gradients. The minority sampling process is illustrated in Fig. 2.

Our minority sampling has three steps. (1) We first apply DDIM inversion [61] to estimate $\hat{x}_t$ from the rendered image $\hat{x}_0$, preserving the conditioning on text prompt y. (2) We then estimate the denoised rendered image $\bar{x}_0$ from $\hat{x}_t$ by following the reverse process $\bar{x}_0 = (1/\alpha_t)\big(\hat{x}_t - \sigma_t \epsilon_\psi(\hat{x}_t, y)\big)$. (3) We then diffuse $\bar{x}_0$ to obtain the perturbed sample $\bar{x}_t = \alpha_t \bar{x}_0 + \sigma_t \epsilon$ where $\epsilon \sim \mathcal{N}(0, \mathbb{I})$ is random noise. As the perturbation to produce $\bar{x}_0$ is random, the noised sample $\bar{x}_t$ becomes less well aligned with the original prompt y, meaning that the density $p_\psi(y|\bar{x}_t)$ is low [73] and hence a good approximation to $p(\phi|\hat{x}_t)$. The full derivation can be found in the supplementary material.

The estimated gradient is therefore:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{t,\epsilon}\Big[w(t)\frac{\alpha_t}{\sigma_t}(\hat{\epsilon}_\psi(\hat{x}_t, y) - \hat{\epsilon}_\psi(\bar{x}_t, y))\frac{\partial \hat{x}_0}{\partial \theta}\Big], \quad (12)$$

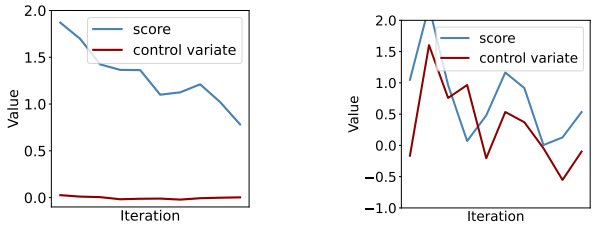
where $\hat{\epsilon}_\psi(\hat{x}_t, y)$ and $\hat{\epsilon}_\psi(\bar{x}_t, y)$ are the predicted scores. It is worth noting that it is natural to integrate classifier guidance scale [26], denoted by s, into this gradient by representing $\hat{\epsilon}_\psi(\hat{x}_t, y) = \epsilon_\psi(\hat{x}_t, \oslash) + s(\epsilon_\psi(\hat{x}_t, y) - \epsilon_\psi(\hat{x}_t, \oslash))$ and $\hat{\epsilon}_\psi(\bar{x}_t, y) = \epsilon_\psi(\bar{x}_t, \oslash) + s(\epsilon_\psi(\bar{x}_t, y) - \epsilon_\psi(\bar{x}_t, \oslash))$. It can be observed that the gradient formulated in Eq. 12 leads to increased diversity of the generated samples because the estimated noise $\hat{\epsilon}_\psi(\bar{x}_t, y)$ acts as an effective control variate. The detailed derivation is provided in the supplementary material. In the next section, we will present an empirical analysis of this gradient, connecting it to the control variate perspective on improved optimization.

5. Experimental Results

5.1. Empirical Analysis

We conducted empirical experiments using a toy diffusion model to analyze the optimization behavior of our proposed objective. We illustrate the convergence of the optimization, demonstrating its gradient stability and trajectory diversity.

Gradient Stability. We trained a simple diffusion model on a dataset of eight cluster samples drawn from a mixture of eight two-dimensional Gaussian distributions. We then performed optimization using JSD and SDS to sample a data point toward a specific cluster. Each cluster is considered a class, which will be used for classifier-free guidance during score distillation. We examine the value of the estimated score and the control variate (the left and right term in the gradient in Eq. 12). It can be observed that both terms are positively correlated, making the gradient values close to zero, hence reducing variances and stabilizing the optimization (Fig. 3).



(a) SDS (b) JSD

Figure 3. Quantitative comparison between estimated noise and control variate in SDS and JSD. It can be seen that our control variate is positively correlated to the estimated noise, hence reducing variances in gradient estimation.

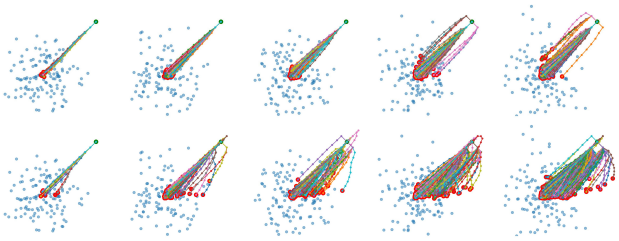
Gradient Trajectory Diversity. Figure 4 illustrates the gradient trajectory from the same initialization, comparing SDS and our gradients from JSD. It can be seen that our method has more diverse trajectories, resulting in different modes.

5.2. 3D Generation Results

Implementation details. We implement the proposed method on top of the Threestudio [23] framework. We perform optimization for 10,000 steps using StableDiffusion[58]. We set the DDIM inversion [49] steps to 10. Otherwise, we use a CFG of 13.5 for all experiments, providing a balance between quality and diversity.

Benchmark and metrics. In this paper, we provide results on T3Bench [24]. T3bench includes 300 prompts, classified into three categories: Single Objects (SO), Single Objects with Surrounding (SOS), and Multiple Objects (MO). To evaluate the fidelity, 3D samples are converted into a mesh form of the level-0 icosahedron and then scored by an ImageReward [87] model. For asset alignment, an image captioning model (BLIP [38]) is used to obtain captions across multiple views, which GPT4 scores. In the appendix, we also provided benchmark results from GPTEval3D [83].

Quantitative results. Table 1 presents the comparisons with SOTA methods. Our method's performance is competitive



(a) 10 (b) 100 (c) 200 (d) 500 (e) 1000

Figure 4. Optimization behavior of SDS (top row) and our method (bottom row). Each score distillation (random seed) is performed from a fixed starting point (•), converging to a result (•). The number of trajectories is from 10 to 1000.

with DreamMesh [89] and CompGS [19], especially in the SO and SOS categories. For text prompts involving multiple objects, our method achieved the highest alignment assessment score. We also include comparisons with Trellis, a state-of-the-art pretrained 3D generation model [85]. Trellis can rapidly generate 3D models, but tends to underperform score distillation methods.

Qualitative comparison. We provide examples to compare our method with SOTA methods, including Consistent Flow Distillation [88] (CFD), ScaleDreamer [45] (SD), DreamReward [90], JointDreamer [31] (JD), MVDream [59], ASD [80], ProlificDreamer [78] (VSD), HIFA [94], and LatentNerf [47] (see Figure 5). Overall, our approach can generate higher-quality 3D assets than previous methods. For fair comparison, we used prompts provided by the previous research. In ASD, many artifacts appear around the main objects, which is evidence of the instability of using the Wasserstein metric. For the object with surrounding prompts, ScaleDreamer and ASD are unstable in learning both the foreground and the background. Our approach can learn the main object, the surrounding objects, and the background with good quality. We provide more examples at Appendix.

Our method can produce 3D objects well aligned with the guidance prompt in the alignment assessment. We present various 3D objects generated with long, detailed prompts in Fig. 1 to indicate the strong alignment characteristic of our proposed method. Other methods tend to ignore information in the background, not focus on the surrounding objects, or place those objects in the wrong location.

Janus problem. Table 1 shows that our method exhibits fewer Janus problems compared to other methods (Fig. 7). However, for objects which has both concave and convex geometry, our method still suffers from the Janus problem, which can be alleviated by using multi-view guidance [59].

Diversity comparison. To evaluate diversity in our optimization, we initialize the representation θ with fixed values and perform score distillation in multiple runs, each using a different random seed. We follow DiverseDream [71] to compute the Inception Variance (IV) and Cosine Similarity metric. The formula of IV is $\mathrm{IV}(\theta) = \mathcal{H}[\mathbb{E}_{i,c}[p(y|g(\theta_i, c)]]$, where $p(y|x_i = g(\theta_i, c))$ is the pretrained classifier given the rendered images x_i from particles i. When the outputs of a pre-trained classification model are uniform, it means that the diversity is high, along with a higher IV. We leverage InceptionV3 [63] as a pre-trained classifier and DINO [8] as a feature extractor. InceptionV3 and DINO are used to obtain the likelihood and feature vector to compute the cosine similarity matrix. We utilize 120 views of an object in the evaluation process.

We compare the diversity of our method with DreamFusion [56], ProlificDreamer [78], and DiverseDream [71]. The result is shown in Table 2 using 70 prompts inherited

Table 1. Comparative results for the text-to-3D tasks in T3Bench. The best results are **bold** while the second best results are underlined.

	Conference	Time	Single Object			Single Object with Surr			Multiple Objects		
Method		(mins)	**Qual. ↑**	**Align. ↑**	**Avg ↑**	**Qual. ↑**	**Align. ↑**	**Avg ↑**	**Qual. ↑**	**Align. ↑**	**Avg ↑**
LatentNeRF [47]	CVPR 2023	65	34.2	32.0	33.1	23.7	37.5	30.6	21.7	19.5	20.6
VSD [78]	NIPS 2023	240	51.1	47.8	49.4	42.5	47.0	44.8	45.7	25.8	35.8
MVDream [59]	ICLR 2024	30	53.2	42.3	47.8	36.3	48.5	42.4	39.0	28.5	33.8
DreamGaussian [65]	ICLR 2024	7	19.9	19.8	19.8	10.4	17.8	14.1	12.3	9.5	10.9
RichDreamer [57]	CVPR 2024	70	57.3	40.0	48.6	43.9	42.3	43.1	34.8	22.0	28.4
VP3D [11]	CVPR 2024	-	54.8	52.2	53.5	45.4	50.8	48.1	49.1	31.5	40.3
ModeDreamer [70]	Arxiv 2024	40	55.4	52.6	54.0	45.7	**59.0**	52.4	43.4	39.4	41.4
DreamReward* [90]	ECCV 2024	40	54.3	43.8	49.0	38.2	49.6	43.9	41.2	33.6	37.4
DreamMesh [89]	ECCV 2024	30	55.6	**53.8**	54.7	43.1	54.3	48.7	47.6	30.8	39.2
Trellis [85]	CVPR 2025	$< \mathbf{1}$	35.6	21.4	28.5	19.5	17.5	18.5	16.6	17.0	16.8
CompGS [19]	CVPR 2025	30	55.1	52.5	53.8	43.2	46.8	45.0	**54.2**	37.9	**46.1**
Ours		70	**58.7**	53.6	**56.1**	**47.4**	57.6	**52.5**	51.3	**40.2**	45.7

Figure 5. Qualitative comparison with available text-to-3D generation techniques.

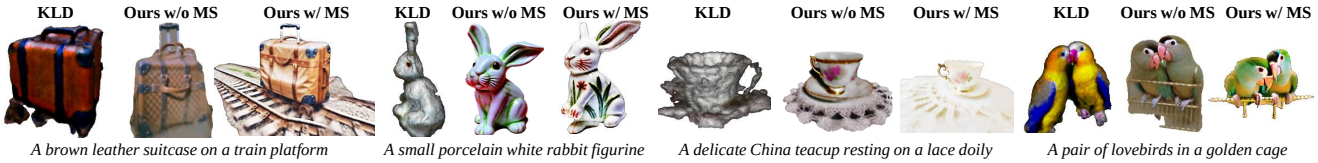


Figure 6. Visual comparison between KLD and our proposed method with and without minority sampling (MS).

Figure 7. Multi-view analysis. Our method produces objects with consistent views.

from DiverseDream [71]. While DiverseDream produces a wide range of 3D models given the same prompt, their object quality varies, including some over-saturated objects. Unlike DiverseDream, our method generates diverse and high-quality 3D objects solely from text prompts, not requiring any reference images or textual inversion.

Table 2. Diversity comparisons between DiverseDream and ours.

Method	SDS[56]	VSD[78]	DiverseDream[71]	Ours
IV↑	4.577	4.586	5.075	**5.826**
Cosine Sim↓	0.720	0.476	**0.380**	0.401

Inversion analysis. To prove the independence of our proposed method on the inversion technique, we compare the results of our method using DDIM [61] and Renoise [18] inversion (Tab. 3). As can be seen, our proposed method is not influenced much by the choice of the inversion method, either DDIM or Renoise, where the performance achieved by both inversion techniques are very close.

Ablation test with Minority Sampling To better understand the effect of minority sampling, we perform an ablation test to compare SDS [56], our proposed method, with and without minority sampling. According to Fig. 6, minority sampling plays a crucial role in our proposed method, which stabilizes the updating gradient (as discussed in Sec. 5.1) and thereby enhances the quality of the output 3D object.

6. Limitation and Conclusion

Limitation. Our method is not without limitations. First, although our approach can generate high-fidelity objects, it still faces common problems in text-to-3D generation, such as Janus [5], hollow face illusion [25], and diffusion anomalies [44], which might be solved by further regularization. Second, the diversity in our pipeline is fully automatic. Extending our method to include more controls would benefit downstream applications like 3D manipulation and editing.

Table 3. Quantitative comparison between DDIM and Renoise inversion conducted on T3Bench [24] benchmark.

		Qual.	Align	Avg
Single Object	*DDIM*	**58.7**	**53.6**	56.1
	Renoise	**58.7**	53.5	**56.2**
Single Object w. Surr	*DDIM*	47.4	57.6	52.5
	Renoise	**47.8**	**58.2**	**53.0**
Multiple Objects	*DDIM*	51.3	**40.2**	45.7
	Renoise	**51.6**	**40.2**	**45.9**

Conclusion. In this research, we propose to approximate the Jensen-Shannon Divergence (JSD) to improve the convergence and diversity of text-to-3D generation. Our method reformulates JSD using the theory of GAN training, leading to a minority sampling technique that effectively approximates JSD. For future work, we believe that there are more objective functions to explore in addition to JSD for 3D generation, including variants of the Wasserstein distances [3, 22, 81]. Extending our method to generate dynamic 3D objects would be an interesting future work as well.

Acknowledgments. This project is supported by Research Ireland under the Research Ireland Frontiers for the Future Programme - Project, award number 22/FFP-P/11522.

References

[1] Ali Akbari, Muhammad Awais, Manijeh Bashar, and Josef Kittler. How does loss function affect generalization performance of deep learning? application to human age estimation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 141–151. PMLR, 2021. 4

[2] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017. 3, 4

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 1, 3, 4, 8

[4] Mohammadreza Armandpour, Ali Sadeghian, Chunyuan Li, and Mingyuan Zhou. Partition-guided gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109, 2021. 3

[5] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm for 2d/3d diffusion, 2024. 8

[6] Duhyeon Bang and Hyunjung Shim. Mggan: Solving mode collapse using manifold-guided training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2347–2356, 2021. 3

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. 1, 4, 5

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 6

[9] Cheng Chen, Xiaofeng Yang, Fan Yang, Chengzeng Feng, Zhoujie Fu, Chuan-Sheng Foo, Guosheng Lin, and Fayao Liu. Sculpt3d: Multi-view consistent text-to-3d generation with sparse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10228–10237, 2024. 3

[10] Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. In *The Conference on Neural Information Processing Systems*, 2024. 4

[11] Yang Chen, Yingwei Pant, Haibo Yang, Ting Yao, and Tao Meit. Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 7

[12] Jinyoung Choi and Bohyung Han. Mcl-gan: Generative adversarial networks with multiple specialized discriminators. In *Advances in Neural Information Processing Systems*, pages 29597–29609. Curran Associates, Inc., 2022. 3

[13] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. In *Conference on Neural Information Processing Systems*, 2023. 4

[14] Jacob Deasy, Nikola Simidjievski, and Pietro Lió. Constraining variational inference with geometric jensen-shannon divergence. In *Advances in Neural Information Processing Systems*, pages 10647–10658. Curran Associates, Inc., 2020. 4

[15] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mggan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13138–13147, 2021. 3

[16] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49 (7):1858–1860, 2003. 4

[17] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 30284–30297. Curran Associates, Inc., 2021. 4

[18] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *Computer Vision – ECCV 2024*, 2025. 8

[19] Chongjian Ge, Chenfeng Xu, Yuanfeng Ji, Chensheng Peng, Masayoshi Tomizuka, Ping Luo, Mingyu Ding, Varun Jampani, and Wei Zhan. Compgs: Unleashing 2d compositionality for compositional text-to-3d via dynamically optimizing 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18509–18520, 2025. 3, 6, 7

[20] Arnab Ghosh, Viveka Kulharia, Vinay Namboodiri, Philip H.S. Torr, and Puneet K. Dokania. Multi-agent diverse generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8513–8521, 2018. 3

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 1, 3, 4

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1, 3, 8

[23] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 6

[24] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T^3bench: Benchmarking current progress in text-to-3d generation, 2023. 3, 6, 8

[25] Harold Hill and Alan Johnston. The hollow-face illusion: Object-specific knowledge, general assumptions or properties of the stimulus? *Perception*, 36(2):199–223, 2007. PMID: 17402664. 8

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information*

Processing Systems, pages 6840–6851. Curran Associates, Inc., 2020. 3

[28] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. MGAN: Training generative adversarial nets with multiple generators. In *International Conference on Learning Representations*, 2018. 3

[29] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[30] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[31] Chenhan Jiang, Yihan Zeng, Tianyang Hu, Songcun Xu, Wei Zhang, Hang Xu, and Dit-Yan Yeung. Jointdreamer: Ensuring geometry consistency and text congruence in text-to-3d generation via joint score distillation. In *Computer Vision – ECCV 2024*, Cham, 2025. 1, 3, 6

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 1

[33] Jeongsol Kim, Geon Yeong Park, and Jong Chul Ye. Dreamsampler: Unifying diffusion sampling and score distillation for image manipulation. In *Computer Vision – ECCV 2024*, pages 398–414, Cham, 2025. Springer Nature Switzerland. 3

[34] Anthony W Knapp. *Basic Real Analysis*. Birkhauser Boston, Secaucus, NJ, 2005 edition, 2005. 3

[35] Sungyoon Lee, Woojin Lee, Jinseong Park, and Jaewook Lee. Towards better understanding of training certifiably robust models against adversarial examples. In *Advances in Neural Information Processing Systems*, pages 953–964. Curran Associates, Inc., 2021. 4

[36] Bonan Li, Zicheng Zhang, Xingyi Yang, and Xinchao Wang. Coser: Towards consistent dense multiview text-to-image generator for 3d creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2880–2890, 2025. 3

[37] Haolin Li, Jinyang Liu, Mario Sznaier, and Octavia Camps. 3d-hgs: 3d half-gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10996–11005, 2025. 1

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 6

[39] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *The International Conference on Learning Representations*, 2024. 3

[40] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. LucidDreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6526, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1, 3

[41] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. 1, 3, 4

[42] Haozhe Liu, Bing Li, Haoqian Wu, Hanbang Liang, Yawen Huang, Yuexiang Li, Bernard Ghanem, and Yefeng Zheng. Combating mode collapse via offline manifold entropy estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8834–8842, 2023. 3

[43] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. *The International Conference on Computer Vision (ICCV)*, 2023. 3

[44] Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Campagnolo Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin Solomon. Score distillation via reparametrized DDIM. In *The Conference on Neural Information Processing Systems*, 2024. 1, 3, 8

[45] Zhiyuan Ma, Yuxiang Wei, Yabin Zhang, Xiangyu Zhu, Zhen Lei, and Lei Zhang. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *Computer Vision – ECCV 2024*, pages 1–19, Cham, 2025. 1, 3, 6

[46] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, pages 3481–3490. PMLR, 2018. 3

[47] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12663–12673, 2023. 3, 6, 7

[48] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. 1

[49] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 6

[50] Yao Ni, Piotr Koniusz, Richard Hartley, and Richard Nock. Manifold learning benefits gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11265–11274, 2022. 3

[51] Frank Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters*, 20(7):657–660, 2013. 4

[52] Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019. 4

[53] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards, 2011. 4

[54] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3

[55] David Keetae Park, Seungjoo Yoo, Hyojin Bahng, Jaegul Choo, and Noseong Park. Megan: Mixture of experts of generative adversarial networks for multimodal image generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 878–884. International Joint Conferences on Artificial Intelligence Organization, 2018. 3

[56] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The International Conference on Learning Representations*, 2023. 1, 3, 6, 8

[57] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 7

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 3, 6

[59] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The International Conference on Learning Representations*, 2024. 3, 6, 7

[60] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 2256–2265. JMLR.org, 2015. 3

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5, 8

[62] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 6

[64] Boshi Tang, Jianan Wang, Zhiyong Wu, and Lei Zhang. Stable score distillation for high-quality 3d generation, 2024. 3

[65] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *The International Conference on Learning Representations*, 2024. 7

[66] 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 *Computer Vision – ECCV 2024*, pages 1–18, Cham, 2025. Springer Nature Switzerland. 3

[67] Ponkrshnan Thiagarajan and Susanta Ghosh. Jensen–shannon divergence based novel loss functions for bayesian neural networks. *Neurocomputing*, 618:129115, 2025. 4

[68] Joy A. Thomas Thomas M. Cover. *Entropy, Relative Entropy, and Mutual Information*, chapter 2, pages 13–55. John Wiley and Sons, Ltd, 2005. 1, 3

[69] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3554–3563. IEEE Computer Society, 2024. 4

[70] Uy Dieu Tran, Minh Luu, Phong Ha Nguyen, Khoi Nguyen, and Binh-Son Hua. Modedreamer: Mode guiding score distillation for text-to-3d generation using reference image prompts, 2024. 3, 7

[71] Uy Dieu Tran, Minh Luu, Phong Ha Nguyen, Khoi Nguyen, and Binh-Son Hua. Diverse text-to-3d synthesis with augmented text embedding. In *European Conference on Computer Vision*, pages 217–235. Springer, 2025. 6, 8

[72] Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In *Computer Vision – ECCV 2024*, pages 414–430, Cham, 2025. Springer Nature Switzerland. 3

[73] Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *ICLR*, 2024. 1, 3, 5

[74] Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, and Leonidas Guibas. Cad : Photorealistic 3d generation via adversarial distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10194–10207, 2024. 3

[75] Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 4

[76] Peihao Wang, Zhiwen Fan, Dejia Xu, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Steindreamer: Variance reduction for text-to-3d score distillation via stein identity. *arXiv preprint arXiv:2401.00604*, 2023. 3

[77] Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, and Vikas Chandra. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[78] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan LI, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems*, pages 8406–8441. Curran Associates, Inc., 2023. 1, 3, 4, 6, 7, 8

[79] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. In *The International Conference on Learning Representations*, 2023. 3

[80] Min Wei, Jingkai Zhou, Junyao Sun, and Xuesong Zhang. Adversarial score distillation: When score distillation meets gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8131–8141, 2024. 1, 3, 4, 6

[81] Xiang Wei, Zixia Liu, Liqiang Wang, and Boqing Gong. Improving the improved training of wasserstein GANs. In *International Conference on Learning Representations*, 2018. 3, 8

[82] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[83] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[84] Mengfei Xia, Yujun Shen, Ceyuan Yang, Ran Yi, Wenping Wang, and Yong-Jin Liu. SMaRt: Improving GANs with score matching regularity. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54133–54155. PMLR, 2024. 3

[85] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3, 6, 7

[86] Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. Latte3d: Large-scale amortized text-to-enhanced3d synthesis. *The 18th European Conference on Computer Vision (ECCV)*, 2024. 3

[87] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023. 6

[88] Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *The International Conference on Learning Representations*, 2025. 3, 6

[89] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Zuxuan Wu, Yu-Gang Jiang, and Tao Mei. Dreammesh: Jointly manipulating and texturing triangle meshes for text-to-3d generation. In *Computer Vision – ECCV 2024*, pages 162–178, Cham, 2025. Springer Nature Switzerland. 6, 7

[90] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. In *European Conference on Computer Vision*, pages 259–276. Springer, 2024. 6, 7

[91] Silong Yong, Yaqi Xie, Simon Stepputtis, and Katia Sycara. Gl-nerf: Gauss-laguerre quadrature enables training-free nerf acceleration. In *Advances in Neural Information Processing Systems*, pages 120418–120442. Curran Associates, Inc., 2024. 1

[92] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3

[93] Chenxi Zheng, Yihong Lin, Bangzhen Liu, Xuemiao Xu, Yongwei Nie, and Shengfeng He. Recdreamer: Consistent text-to-3d generation via uniform score distillation. In *The International Conference on Learning Representations*, 2025. 3

[94] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. HIFA: High-fidelity text-to-3d generation with advanced diffusion guidance. In *The International Conference on Learning Representations*, 2024. 6