Diverse Text-to-3D Synthesis with Augmented Text Embedding

Uy Dieu Tran*[1], Minh Luu*[1], Phong Ha Nguyen[1], Khoi Nguyen[1], and Binh-Son Hua[1,2]

[1]VinAI Research [2]Trinity College Dublin

Abstract. Text-to-3D synthesis has recently emerged as a new approach to sampling 3D models by adopting pretrained text-to-image models as guiding visual priors. An intriguing but underexplored problem with existing text-to-3D methods is that 3D models obtained from the sampling-by-optimization procedure tend to have mode collapses, and hence poor diversity in their results. In this paper, we provide an analysis and identify potential causes of such a limited diversity, which motivates us to devise a new method that considers the joint generation of different 3D models from the same text prompt. We propose to use augmented text prompts via textual inversion of reference images to diversify the joint generation. We show that our method leads to improved diversity in text-to-3D synthesis qualitatively and quantitatively. Project page: `https://diversedream.github.io/`

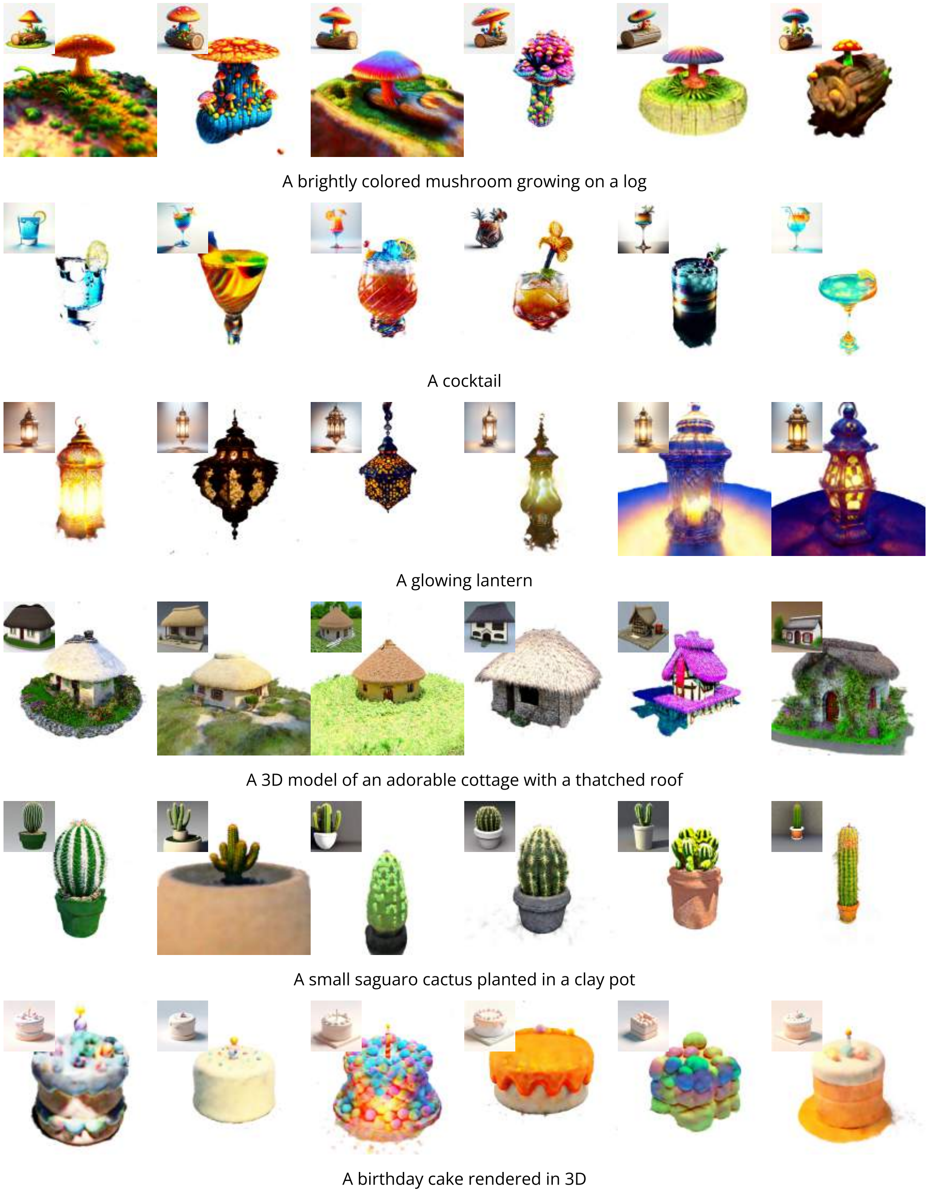
Keywords: Text-to-3D · 3D computer vision · Generative models

1 Introduction

The realm of 3D content creation has persistently posed intricate challenges within the domains of computer vision and computer graphics. Over time, various methodologies have emerged to address this challenge. Traditional techniques in generating 3D models often necessitate user interaction, involving meticulous shaping of scene geometry and appearance through software like Blender [2]. Another prevalent avenue revolves around scene reconstruction using multi-view geometry principles, extensively explored in literature such as [31]. These approaches have garnered substantial adoption, particularly within industries like interior design and computer animation, revolutionizing their workflows and creative possibilities.

The rise of deep learning has led to increased interest in developing data-driven techniques to automate 3D modeling. Several efforts have been made to generate 3D models by learning directly from 3D data [51]. However, due to the scarce availability of 3D data, it has been of great interest to explore the generation of 3D data by learning from different modalities such as images and natural languages. It has been shown that pretrained text-to-image diffusion

[*] These authors contributed equally to this work



A brightly colored mushroom growing on a log

A cocktail

A glowing lantern

A 3D model of an adorable cottage with a thatched roof

A small saguaro cactus planted in a clay pot

A birthday cake rendered in 3D

Fig. 1. We address the intriguing low-diversity issue in text-to-3D synthesis by reconsidering the text prompt used by variational score distillation [57]. We propose to use reference images to sample augmented text prompts via textual inversion and use these augmented text prompts to condition the particles in the variational inference of text-to-3D optimization to learn more diverse 3D representations. Thanks to the diversity in the reference images (top-left inline images), we obtain diverse 3D models that inherit certain structures from their references.

models can serve as a strong prior to guiding the optimization of a 3D model represented by a neural radiance field in DreamFusion with SDS loss [38], from which text-to-3D synthesis emerges as a promising research direction.

While text-to-3D synthesis has shown promises, challenges persist in fidelity, diversity, convergence, and scalability of generated models. Efforts to address these issues include enhanced loss functions like ProlificDreamer [57], generalized across prompts in ATT3D [33], ATOM [39], ET3D [9], and personalized generation in DreamBooth3D [42]. Diversity, however, remains largely unexplored in current text-to-3D methods, with limited insight into its mechanisms.

In this paper, we explore methods to enhance the diversity of 3D model generation in text-to-3D systems. We posit that the diversity of model outputs is influenced by the objective function used, such as SDS [38] and VSD [57], when conditioning 3D model generation on a text prompt. Motivated by this insight, we propose a method to diversify text-to-3D generation results by augmenting the original text prompt through textual inversion techniques [11, 14]. Our approach involves sampling reference images from a pretrained text-to-image diffusion model and extracting the corresponding text features via Textual Inversion. These text features are then combined with the features of the original text prompts to guide the optimization process for sampling 3D models. Experimental results (Fig. 1) demonstrate a significant improvement in the diversity of generated 3D models compared to state-of-the-art methods, both quantitatively and qualitatively.

In summary, our contributions are:

- An empirical analysis of the diversity of existing text-to-3D methods;
- A general technique based on augmented text embedding acquired from textual inversion of 2D reference images to improve the diversity and speed of the optimization process;
- Extensive experiments and ablation studies to demonstrate the validity and robustness of our method, which is applicable to different text-to-3D methods.

2 Related Work

Text-to-image synthesis has seen significant advancements, with methods relying on Generative Adversarial Networks (GANs) [3, 23, 47] and auto-regressive models like DALL-E [43], Parti [61], and MUSE [6]. While GANs offer fast and realistic image generation, they are prone to mode collapse. Recently, diffusion models such as Stable Diffusion [44], DALL-E 3 [50], and Imagen [46] have shown promise in synthesizing high-quality images. In this study, we utilize diffusion-based models, particularly Stable Diffusion, as the pretrained 2D prior to supervise our 3D model generation.

3D representation serves as the foundation for various 3D tasks like novel view synthesis and content creation. Neural Radiance Fields (NeRFs) [34] have gained traction for their volumetric rendering approach, learning 3D scenes from 2D images alone. Despite NeRF's widespread use [55], its optimization process is time-intensive [12]. To address this, researchers have explored hybrid scene representations like voxel grids [52, 60], hash-grids [35], tri-planes [4, 7], and Gaussian splatting [25], aiming to improve speed and view synthesis performance. Among these, hash-grids [35] are favored for text-to-3D tasks [38, 57] due to their fast training and superior performance compared to NeRFs. In this paper, we leverage hash grids to learn diverse 3D scene representations from a single text prompt using our proposed textual score distillation loss.

Image-to-3D generation is a crucial aspect of conditional generative 3D models. Early methods like SynSin [59] and Free View Synthesis [36] rely on differentiable neural renderers for single view synthesis but are limited by pose distances and struggle with full 360°reconstructions from a single input. Recent advancements have seen models like Zero-1-to-3 [31] pioneering open-world single-image-to-3D conversion through zero-shot novel view synthesis, yet face challenges with geometric consistency. Works such as One-2-3-45 [30], SyncDreamer [32], LRM [17], LGM [53], and Consistent123 [58] address this by adding geometry-constraint layers to improve consistency. However, these methods typically require extensive 3D model datasets like ShapeNet [5] or large-scale multiview datasets like Objaverse [10] for training. In contrast, our approach solely relies on a pretrained text-to-image model for supervision, making it more accessible.

Text-to-3D generation has seen remarkable progress recently, leveraging pretrained text-to-image models like Stable Diffusion [44]. Early works like DreamField [21] use CLIP [41] to align rendered images with input text but often compromise on model quality due to CLIP's limited semantic feature capture. DreamFusion [38] substitutes CLIP loss with Score Distillation Sampling (SDS) and introduces efficient gradient calculation for neural radiance field learning. However, it tends to produce oversmooth surfaces and saturated colors. Subsequent methods aim to address these limitations by enhancing resolution [28], appearance [8, 24, 62, 64], geometry [19, 27, 40, 48, 49], speed [20, 27, 54], and photorealism [24, 26, 57, 64]. Despite this progress, diversity in text-to-3D synthesis remains underexplored, motivating our work.

Textual inversion. While recent text-to-image diffusion models like Stable Diffusion [44] and DALL-E 3 [50] produce high-quality 2D images, they may not preserve the subject's shape or identity, known as "personalization". Techniques such as Textual Inversion [11] and DreamBooth [45] aim to maintain subject identity in reference images by introducing a virtual token whose embedding can be optimized to manipulate the generated images. HiPer inversion [14], building upon Textual Inversion, enhances inversion by using a single reference image and optimizing textual tokens in the text prompt to store object identity. Inspired by this, we apply textual inversion to diversify generated 3D content.

3 Background

A typical approach to text-to-3D synthesis is to leverage the 2D prior from a pretrained text-to-image model such as Stable Diffusion (SD) [44], to guide the training of a 3D model represented by a neural radiance field (NeRF). In particular, a NeRF parameterized by θ is optimized so that its rendered images $x = g(\theta, c)$, with g as the volumetric rendering function and c as the camera pose, look realistic and conform to the text prompt y.

Score distillation sampling (SDS): DreamFusion [38] introduced the SDS loss whose gradient is computed as:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} \triangleq \mathbb{E}_{t,\epsilon,c}\left[\omega(t)(\epsilon_{\text{SD}}(x_t, t, y) - \epsilon)\frac{\partial g(\theta, c)}{\partial \theta}\right], \tag{1}$$

where $\omega(t)$ is a time-dependent weighting function, ϵ_{SD} is the predicted noise of SD given the noisy input image $x_t = \alpha_t x + \sigma_t \epsilon$ created by adding Gaussian noise ϵ to the rendered image x at timestep t with noise scheduling coefficients α_t, σ_t. However, the SDS loss often suffers from over-saturation, over-smoothing, and low-diversity issues as empirically analyzed in [57]. The low diversity issue in SDS becomes apparent when multiple runs yield similar results empirically. Therefore, we advocate the use of the more sophisticated variational score distillation loss [57] for our exploration of the diversity of text-to-3D synthesis, which we briefly describe below.

Variational score distillation (VSD): ProlificDreamer [57] mitigates the limitations of DreamFusion by introducing a variational form of score distillation. Their VSD loss aims to tackle the low-diversity issue by modeling the distribution μ of 3D models θ generated from a single text prompt y as $\mu(\theta|y)$. It's worth noting that SDS is a special case of VSD, where $\mu(\theta|y)$ simplifies to a Dirac distribution $\delta(\theta - \theta^1)$, resulting in only a single 3D model θ^1 for each text prompt.

To optimize VSD, the distribution μ is approximated by K learnable particles where each particle i corresponds to a 3D representation parameterized by θ_i which is sampled from a set of K particles $\{\theta_i\}_{i=1}^K$ for each training iteration, following the particle-based variational inference framework. The gradient of the VSD loss is as follows:

$$\nabla_{\theta_i} \mathcal{L}_{\text{VSD}} \triangleq \mathbb{E}_{t,\epsilon,c}\left[\omega(t)(\epsilon_{\text{SD}}(x_t^i, t, y) - \epsilon_\phi(x_t^i, t, c, y))\frac{\partial g(\theta_i, c)}{\partial \theta_i}\right], \tag{2}$$

where ϵ_ϕ is a fine-tuned version of the original SD using the LoRA [18] parameterization ϕ on the rendered images of in-progress learning NeRFs. LoRA can be regarded as the *domain adaptation* of SD to noisy images rendered from NeRFs since SD is not originally trained on noisy images. Specifically, ϕ is trained with the following objective:

$$\min_\phi \mathbb{E}_{t,\epsilon,c}\left[\|\omega(t)(\epsilon_\phi(x_t^i, t, c, y) - \epsilon)\|_2^2\right]. \tag{3}$$

Beyond its theoretical modeling of 3D representations as a distribution, an empirical observation on why the VSD loss enhances the diversity compared to the SDS loss is that the objective of VSD for each particle is different from each other. Notably, the second term $\epsilon_\phi(x_t^i, t, c, y)$ (in Eq. (2)) dynamically changes due to the learning progression of ϕ and the input image x_t^i rendered from the current particle θ_i.

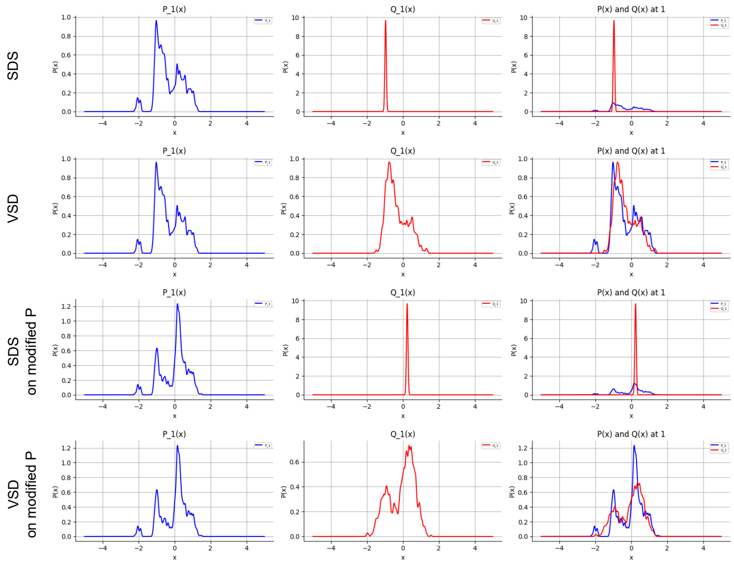


Fig. 2. We present a simulation of SDS (first row) and VSD (second row) in KL form on a 1D toy dataset, where the ground truth distribution $p_{\text{SD}}(x_t|y)$ is a 7-component Gaussian mixture model. Results are shown at $t = 1$ (low noise data). In the third row and forth row, varying $p_{\text{SD}}(x_t|y')$ with a new text prompt y' leads to diverse outcomes across different runs with SDS/VSD loss, motivating our approach.

Although the VSD loss addressed the limitation of SDS loss and clearly improved the quality of the 3D representations, we empirically found that it still yields limited diversity in some particular prompts. To further improve diversity, we propose to use augmented text embedding guided by 2D reference images, which is presented in the next section.

4 Our Approach

4.1 Analysis

Let us first motivate our method by an empirical analysis on the diversity of SDS and VSD loss, the prevailing loss functions for generating 3D assets from a given text prompt using a text-to-image model as prior. A notable drawback of this technique is that the SDS loss often yields almost identical results across different runs, primarily due to the mode-seeking behavior exhibited by the KL divergence between a Gaussian distribution and a multi-modal landscape of the text-to-image prior. More precisely, as shown from [38], the SDS loss in the KL form is given by:

$$\mathcal{L}_{\text{SDS}} \triangleq \mathbb{E}_t \left[\text{KL}(q(x_t|x = g(\theta, c))||p_{\text{SD}}(x_t|y)) \right]. \tag{4}$$

At a given time step t, $q(x_t|x) = \mathcal{N}(\alpha_t x, \sigma_t^2 \mathbf{I})$ represents a Gaussian distribution characterizing the forward diffuse process of the rendered image, while $p_{\text{SD}}(x_t|y) =$

$\int p^0_{\mathrm{SD}}(x_0|y)q(x_t|x_0)dx_0$ denotes the marginal distribution of the diffusion model. It is reasonable to assume that the distribution of the diffusion model exhibits multimodality, particularly for lower values of t. Considering that $q(x_t|x)$ is a unimodal distribution, it tends to align with the closest mode of $p_{\mathrm{SD}}(x_t|y)$, as demonstrated by [1]. This behavior is particularly pronounced in 3D, as it necessitates initializing the 3D scene as a Gaussian blob in each run. Consequently, the initial parameters of NeRF θ tend to be closer to each other in successive runs, resulting in low-diversity outcomes for SDS.

Meanwhile, the KL form of VSD (Eq. (5)) is nearly identical to SDS, albeit the substitution of the unimodal Gaussian $q(x_t|x)$ with a more intricate, implicit distribution $q^{\mu}(x_t|c, y) = \mathbb{E}_{\mu(\theta|y)}[q(x_t|x = g(\theta, c))]$.

$$\mathcal{L}_{\mathrm{vsd}} \triangleq \mathbb{E}_t \left[\mathrm{KL}(q^{\mu}(x_t|c, y)||p_{\mathrm{SD}}(x_t|y))\right] \tag{5}$$

By increasing the complexity of q^{μ}, the optimized distribution $q^{\mu*}$ will possess a greater capacity to accurately fit the target distribution $p_{SD}(x_t|y)$. The VSD algorithm aims to draw data from $\mu^*(\theta|y)$ to minimize Eq. (5) through particle-based variational inference. As a result, the final 3D assets will exhibit a greater degree of variety.

We elucidate our intuition in Fig. 2 by performing a simulation for SDS and VSD on a 1D toy dataset, where $p_{\mathrm{SD}}(x_t|y)$ is a Gaussian mixture model (GMM) comprising 7 components. We select a random position x as the optimizable parameter for SDS, while the means and standard deviations of a 3-component GMM $\mu(\theta|y)$ serve as the optimizable parameters for VSD. Based on this analysis, instead of modeling $q(x_t|x)$ as done in SDS and VSD, we propose a new way to increase the diversity of the text-to-3D models by modifying the distribution $p_{\mathrm{SD}}(x_t|y)$. Our idea is to **diversify the condition y to y'** so that optimization with the prior $p_{\mathrm{SD}}(x_t|y')$ potentially yields a more substantial impact due to the alteration in the optimization landscape.

4.2 Method Overview

To implement the idea of diversifying y, inspired by textual inversion methods [11, 14] for 2D object personalization, we aim to condition text-to-3D synthesis such that the per-particle difference in $\epsilon_{\mathrm{SD}}(x^i_t, t, c, y'_i)$ is boosted by using different and distinct prompt y'_i for each particle θ_i. To this end, we devise a new approach that leverages HiPer textual inversion [14] to enhance the resulting diversity. Here we base our discussion on VSD, but the idea generalizes to SDS as well.

Our approach consists of two stages: HiPer tokens inversion and textual score distillation. Firstly, we select a reference image x^r_i for each particle and determine an optimized HiPer token h^*_i. This token is chosen such that the reference image can be faithfully reconstructed by the pretrained text-to-image diffusion model with the prompt $[y; h^*_i]$, where $[;]$ denotes concatenation. In the second stage, multiple particles θ_i are collectively optimized alongside a new shared domain adapter ϕ, which is also encoded as a learnable token. This process forms the augmented text prompt $[y; h^*_i; \phi]$ for each particle. The algorithm is depicted in Alg. 1 and illustrated in Fig. 3.

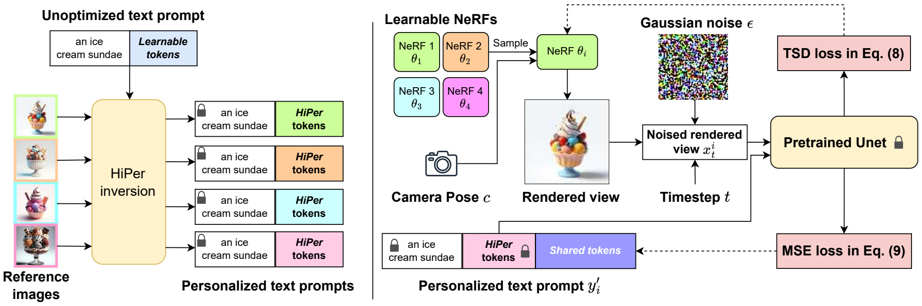


Fig. 3. We translate the diversity of augmented text prompts to the resulting 3D models via a two-stage method. **Stage 1: HiPer tokens inversion** (left): for each reference image, we seek to learn a HiPer token h_i so that the prompt $[y; h_i]$ reconstructs the reference image. **Stage 2: Textual score distillation** (right): we run a multi-particle variational inference for optimizing the 3D models from text prompt y. For each iteration in the optimization, we randomly sample a particle θ_i with its rendered image x_i. We use the augmented text prompt $y_i' = [y; h_i^*; \phi]$, with ϕ as shared embedding to condition the optimization of θ_i (Eq. (8) and Eq. (9)).

4.3 HiPer tokens inversion

We first sample K reference images $\{x_i^r\}_{i=1}^K$ corresponding to K particles from any text-to-image model given text prompt y. We empirically find that using Stable Diffusion (SD) [44] with additional guidance like “X with white background” gives the most suitable images for HiPer textual inversion. This is because we only have one image for inversion and we want to exclude noisy factors like background, facilitating faster and better textual inversion.

Subsequently, we want to optimize HiPer tokens for each reference image using the technique in [14]. Specifically, given a reference image x_i^r and a text prompt $y \in \mathbb{R}^{L_1 \times D}$ with L_1 as the number of text tokens and D as feature dimensions, we seek to find HiPer tokens $h \in \mathbb{R}^{L_2 \times D}$ with L_2 as the number of HiPer tokens to reflects the personalized identity of the object in x_i^r. To this end, we append the learnable tokens h_i to the original text prompt y to form new text personalized text prompt $y_i = [y; h_i] \in \mathbb{R}^{(L_1+L_2) \times D}$, and use HiPer [14] to optimize h_i with the objective:

$$\min_{h_i} \mathbb{E}_{t,\epsilon} \left[\|\omega(t) \epsilon_{\text{SD}}(x_{t,i}^r, t, [y; h_i]) - \epsilon\|_2^2 \right] . \tag{6}$$

Note that HiPer use Stable Diffusion (version 1.4) for textual inversion. This stage is visualized in Fig. 3 (Left). The optimized h_i^* is leveraged as the key component to diversify the results of text-to-3D synthesis in the next step.

4.4 Textual score distillation (TSD)

With the learned personalized text prompts $y_i = [y; h_i^*]$, we are ready to use them to replace the original text prompt y in any text-to-3D approaches such as

Algorithm 1 Algorithm of DiverseDream.

```
Input: $K$ particles, $K$ reference images $\{x_i^r\}_{i=1}^K$ from prompt $y$, pretrained text-to-image model $\epsilon_{\text{SD}}$.
Stage 1: HiPer tokens inversion
initialize $K$ HiPer tokens $\{h_i\}_{i=1}^K$.
for i=1 to K do
    Optimize $h_i$ given $x_i^r$ following Eq. (6) to obtain $h_i^*$.
end for
return optimized $\{h_i^*\}_{i=1}^K$
Stage 2: Textual score distillation
initialize $K$ NeRFs $\{\theta_i\}_{i=1}^K$, shared learnable tokens $\phi$.
while not converged do
    Sample noise $\epsilon$, camera pose $c$, timestep $t$, and index $i$, obtain $\theta_i$, and form text prompt $y_i' = [y; h_i^*; \phi]$.
    Render image $x_i = g(\theta_i, c)$ from NeRF $\theta_i$ at pose $c$, and compute $x_t^i$.
    Update $\theta_i$ following Eq. (8).
    Update $\phi$ following Eq. (9).
end while
return optimized $\{\theta_i^*\}_{i=1}^K$
```

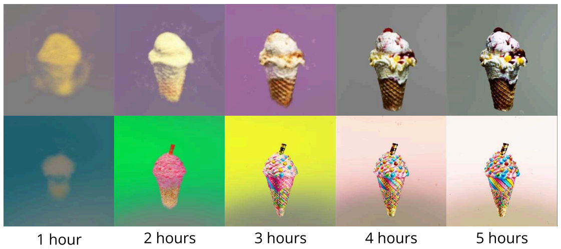


Fig. 4. Optimization progress of VSD (upper) vs ours (lower). TSD with less #learnable parameters converges faster than VSD. Prompt: "A high-quality ice cream sundae".

ProlificDreamer [57] to enhance the diversity of these approaches. However, we discover that the Domain Adaptor ϕ in ProlificDreamer, which is implemented using LoRA [18], can be further replaced by the textual inversion technique like HiPer [14]. This is similar to the problem of 2D object personalization where LoRA Dreambooth [45] can be replaced by Textual Inversion [11] or HiPer [14] with similar performance. The observation motivates us to devise a new Domain Adaptor $\phi \in \mathbb{R}^{L_3 \times D}$ in the form of shared learnable tokens in the text prompt among particles. That is, the new personalized text prompt:

$$y_i' = [y; h_i^*; \phi] \in \mathbb{R}^{(L_1+L_2+L_3)\times D}, \tag{7}$$

where L_3 is the number of shared learnable tokens. The new text prompt y_i' can replace the LoRA implementation ϕ of ProlificDreamer, resulting in the following Textual Score Distillation (TSD):

$$\nabla_{\theta_i} \mathcal{L}_{\text{TSD}} \triangleq \mathbb{E}_{t,\epsilon,c}\left[\omega(t)(\epsilon_{\text{SD}}(x_t^i, t, y_i) - \epsilon_{\text{SD}}(x_t^i, t, y_i'))\frac{\partial g(\theta_i, c)}{\partial \theta_i}\right]. \tag{8}$$

Compared to the LoRA implementation, the term $\epsilon_{\text{SD}}(x_t^i, t, y_i')$ with shared learnable tokens has the advantage of faster training speed since the number of our learnable parameters ϕ (about 30K parameters) is much smaller than those of LoRA (about 1.3M parameters). The training of the shared learnable tokens is similar to the LoRA implementation, i.e., via a separate updating step from the updating step of each particle as:

$$\min_{\phi} \mathbb{E}_{t,\epsilon,c} \left[\|\omega(t)(\epsilon_{\text{SD}}(x_t, t, y_i') - \epsilon)\|_2^2 \right]. \tag{9}$$

In Fig. 3 (Right), we show how we train the sampled NeRF model θ_i and shared token ϕ using the proposed TSD (Eq. (8)) and MSE (Eq. (9)) losses respectively. As can be seen in the Fig. 4, our method can produce higher quality samples than those produced by VSD [57] given the same amount of optimization time.

5 Experiments

Metrics. The common metrics for generative models such as FID [15] do not separately measure fidelity and diversity.

Inspired by [56], we proposed to use a modified version of *Inception Quality (IQ)* and *Inception Variance (IV)* to measure the quality and diversity of our models. Our IQ and IV are formulated as follows:

$$\text{IQ}(\theta) = \mathbb{E}_{i,c} \left[\mathcal{H}[p(y \mid x_i = g(\theta_i, c))] \right], \tag{10}$$

$$\text{IV}(\theta) = \mathcal{H} \left[\mathbb{E}_{i,c}[p(y \mid x_i = g(\theta_i, c))] \right], \tag{11}$$

where $p(y \mid x_i = g(\theta_i, c))$ is the pretrained classifier given the rendered images x_i from particles i. The entropy $\mathcal{H}$ serves as an indicator of the classifier's confidence when presented with an input-rendered image. The IQ metric captures the expected entropy, reflecting the classifier's certainty across all views, which indicates the image quality to some degree *(the lower the better)*. Conversely, the IV metric quantifies the entropy of the expected classifier outputs *(the higher the better)*. A higher IV score is achieved when the classifier outputs are uniformly distributed, indicating greater diversity in the input images. We rendered 120 views for each particle to compute IQ and IV.

We also propose the *Cosine Sim* metric to quantify the diversity of our particles. Specifically, for a given set of K particles, we render the same view for each particle. The rendered images are then fed through a feature extractor (e.g., as DINO [63]) to obtain feature vectors. We then calculate the cosine similarity for these feature vectors across $\binom{K}{2}$ pairs and take the average. To ensure robustness, we also average the results over 120 rendered views.

Implementation details. Our method is implemented with threestudio, an open-source framework for text-to-3D synthesis [13]. We use $K = 6$ particles for both VSD and our framework, TSD. The training of all experiments is conducted for 50K iterations. We use $L_2 = 5$ as HiPer tokens optimized in 1.4K iterations. Also, we use $L_3 = 8$ as shared learnable tokens ϕ and trained in 50K iterations.

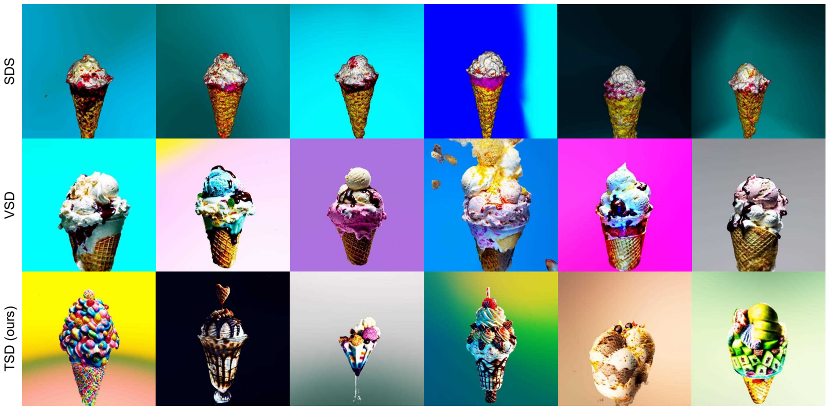


Fig. 5. Diversity comparison between SOTAs and our method.

For the camera embedding, we use the same implementation as threestudio [13]. Regarding the resolution, we train each particle at 256 × 256 for all methods.

Table 1. Comparison with SOTAs

	IQ ↓	**IV** ↑	**Cosine Sim** ↓
SDS [38]	3.695	4.577	0.720
VSD [57]	**3.345**	4.586	0.476
TSD (ours)	3.614	**5.075**	**0.380**

Table 2. Study on #HiPer tokens.

#tokens	**IQ** ↓	**IV** ↑	**Cosine Sim** ↓
1	**3.375**	4.908	**0.403**
5	3.790	4.886	0.415
10	4.428	**5.138**	0.409
15	4.721	4.862	0.445
20	5.193	4.978	0.425

5.1 Comparison with Prior Methods

Baselines. We compare our method with two prominent text-to-3D methods including DreamFusion [38] and ProlificDreamer [57]. Note that we do not compare to other variants such as Magic3D [28] or Fantasia3D [8] since these methods address different issues of SDS, which is orthogonal to our method which focuses on diversity. For validation, we select a set of 60 text prompts from DreamFusion [38] and 10 text prompts generated randomly from ChatGPT [37].

Quantitative comparison. We present our quantitative results in Tab. 1. It is evident that we outperform VSD and SDS in terms of IV Score and Cosine Sim, which measure the diversity between the particles. The IQ score of our method is slightly lower than VSD, which demonstrates there remains some fidelity-diversity trade-off, which is well known for existing text-to-image methods. Our method outperforms SDS in both diversity and quality.

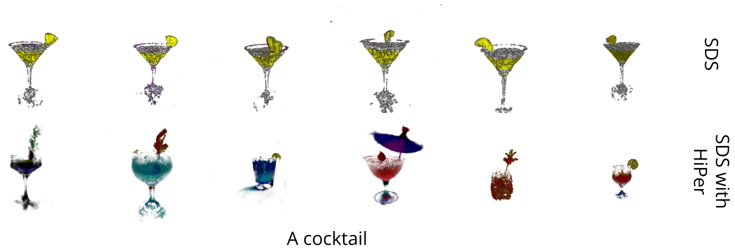


Fig. 6. We demonstrate that our method remains effective for SDS.

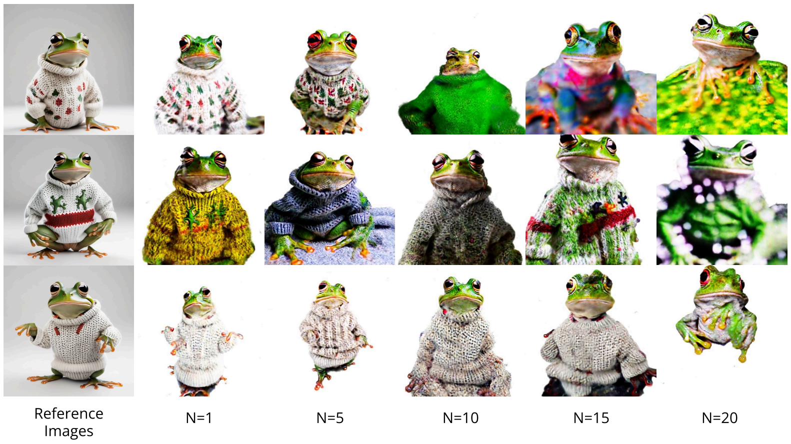


Fig. 7. Visual results with different numbers of HiPer tokens.

Qualitative comparison. The comparative results in Fig. 5 demonstrate that our method offers more diversity among particles compared to VSD and SDS. For example, when given "A high-quality ice cream sundae" prompts, VSD tends to collapse into cone-shaped ice cream, while our method is capable of generating glass shapes and other variants. Our 3D models inherit texture and structure from reference images (see Fig. 1), showcasing the potential of transferring 2D diversity to 3D through text prompt augmentation. Perfect inversion by HiPer is not necessary; capturing the essence of reference images suffices for diversity among personalized text prompts. Fig. 6 demonstrates the diversity of our results when applying HiPer to the SDS loss.

5.2 Ablation Study

In this section, we undertake an ablation study to examine the factors that influence our methods.

Number of HiPer tokens L_2. We vary the number of the HiPer tokens from 1, 5, 10, 15, and 20 using the prompt "A DSLR photo of a frog wearing a sweater", Fig. 7 and Tab. 2 show that while altering token length has a subtle impact on

Table 3. Ablation study of our method.

ϵ_ϕ	**HiPer**	**IQ** ↓	**IV** ↑	**Cosine Sim** ↓	**Training time (hours)**
LoRA	No	**3.345**	4.586	0.476	10.23
LoRA	Yes	3.662	**5.109**	**0.355**	9.50
Shared tokens	Yes	3.614	5.075	0.380	**7.16**

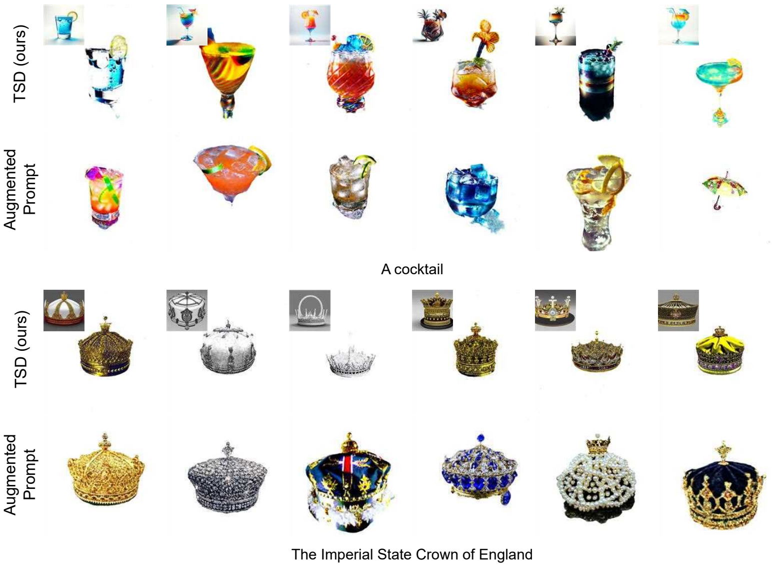


Fig. 8. Comparison between our HiPer inversion with LLM-generated augmentation for text prompt sampling. It can be seen that the use of reference images in our method leads to better fidelity and diversity.

the 3D model, the text-to-image model produces images more closely resembling the reference in 2D.

LoRA vs. shared learnable tokens. We further validate the effect of our shared learnable tokens. When replacing the LoRA layer with our learnable tokens, our method can still achieve diverse results compared to SDS and VSD, although the quality was not as good as our method with LoRA, as shown in Tab. 3. However, the advantage of using shared learnable tokens is that it achieves a better training speed compared to the use of LoRA.

LLM-based text prompt augmentation. In addition to personalized text prompts obtained through image-to-text inversion using HiPer [14], we compare our approach to a different text prompt sampling technique using large language models (LLMs). Given an original text prompt y, we employ ChatGPT [37] to generate K prompts from y, enriching the description of the object. For instance, starting with "A high-quality photo of an ice cream sundae" as the original prompt, we obtain an augmented prompt like "A high-quality photo

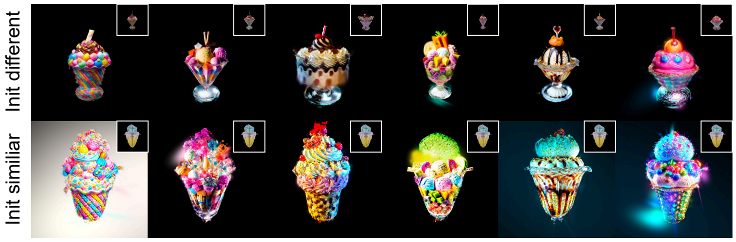


Fig. 9. Our method also achieve remarkable diversity on 3DGS. We also show the initialized point-cloud at the top right corner of each sample.

of an ice cream sundae with fresh berries and mint leaves". Subsequently, we utilize each of these prompts to condition the corresponding particle in the VSD loss. The results, shown in Fig. 8, indicate that while LLM-based augmented text prompts also lead to diverse 3D generations, their quality and diversity are inferior compared to our image-to-text inversion method.

5.3 Extension to 3DGS

To expedite training, our approach can be extended to 3D Gaussian Splatting (3DGS) [25]. Since the 3DGS method requires a point cloud to start the optimization, we utilize recent generative text-to-3D approaches such as 3DTopia [16] and Shape-E [22] to obtain the initial shapes for our training. Specifically, we conducted experiments on 3DGS using the same initial shape from 3DTopia and multiple initial shapes from Shape-E. As shown in Figure 9, our method can generate high-quality and diverse 3D renderings using the 3DGS backbone. The 3DGS representation, compared to Instant-NGP, significantly reduces the training time from approximately 7 hours to about 2 hours when using 6 particles.

6 Discussion and Conclusion

Limitations: Despite enhancing the diversity of the existing VSD framework, our model has limitations. It heavily relies on the HiPer inversion, which may struggle with outlier reference images, resulting in 3D models with unusual shapes and appearances. Also, our method shares VSD's limitations, such as the Janus problem, which can be addressed by orthogonal methods applicable to VSD.

Conclusion: In this paper, we have successfully introduced a new text-to-3D synthesis method that focuses on diversifying the 3D generation by using 2D reference images and textual inversion to build augmented text prompts for conditioning the optimization. In future work, we plan to experiment with our augmented text embedding technique for other text-to-3D methods [8, 29], and more 3D representations [25].

Acknowledgment. Binh-Son Hua is supported by Science Foundation Ireland under the SFI Frontiers for the Future Programme (22/FFP-P/11522).

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning, pp. 469–470. Springer (2006)
2. Blender Online Community: Blender - a 3D modelling and rendering package. Blender Foundation, Blender Institute, Amsterdam (2018)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
6. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
7. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (ECCV) (2022)
8. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023)
9. Chen, Y., Li, Z., Liu, P.: Et3d: Efficient text-to-3d generation via multi-view distillation (2023)
10. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2023)
12. Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: High-fidelity neural rendering at 200fps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14346–14355 (2021)
13. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation (2023)
14. Han, I., Yang, S., Kwon, T., Ye, J.C.: Highly personalized text embedding for image manipulation by stable diffusion. arXiv preprint arXiv:2303.08767 (2023)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
16. Hong, F., Tang, J., Cao, Z., Shi, M., Wu, T., Chen, Z., Wang, T., Pan, L., Lin, D., Liu, Z.: 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. arXiv preprint arXiv:2403.02234 (2024)

17. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: LRM: Large reconstruction model for single image to 3d. In: The Twelfth International Conference on Learning Representations (2024)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
19. Huang, T., Zeng, Y., Zhang, Z., Xu, W., Xu, H., Xu, S., Lau, R.W., Zuo, W.: Dreamcontrol: Control-based text-to-3d generation with 3d self-prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5364–5373 (June 2024)
20. Huang, Y., Wang, J., Shi, Y., Tang, B., Qi, X., Zhang, L.: Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In: The Twelfth International Conference on Learning Representations (2024)
21. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
22. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions (2023), https://arxiv.org/abs/2305.02463
23. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
24. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023)
26. Lee, K., Sohn, K., Shin, J.: Dreamflow: High-quality text-to-3d generation by approximating probability flow. In: The Twelfth International Conference on Learning Representations (2024)
27. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In: The Twelfth International Conference on Learning Representations (2024)
28. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
29. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
30. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
31. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9298–9309 (October 2023)
32. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Learning to generate multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)

33. Lorraine, J., Xie, K., Zeng, X., Lin, C.H., Takikawa, T., Sharp, N., Lin, T.Y., Liu, M.Y., Fidler, S., Lucas, J.: Att3d: Amortized text-to-3d object synthesis. arXiv preprint arXiv:2306.07349 (2023)
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
35. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
36. Nguyen-Ha, P., Sarafianos, N., Lassner, C., Heikkilä, J., Tung, T.: Free-viewpoint rgb-d human performance capture and rendering. In: European Conference on Computer Vision. pp. 473–491. Springer (2022)
37. OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023)
39. Qian, G., Cao, J., Siarohin, A., Kant, Y., Wang, C., Vasilkovsky, M., Lee, H.Y., Fang, Y., Skorokhodov, I., Zhuang, P., Gilitschenski, I., Ren, J., Ghanem, B., Aberman, K., Tulyakov, S.: Atom: Amortized text-to-mesh using 2d diffusion (2024)
40. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., Ghanem, B.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In: The Twelfth International Conference on Learning Representations (2024)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
42. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Mildenhall, B., Ruiz, N., Zada, S., Aberman, K., Rubenstein, M., Barron, J., Li, Y., Jampani, V.: Dreambooth3d: Subject-driven text-to-3d generation. ICCV (2023)
43. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (18–24 Jul 2021)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
45. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
47. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: StyleGAN-t: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 30105–30118. PMLR (23–29 Jul 2023)

48. Seo, J., Jang, W., Kwak, M.S., Kim, H., Ko, J., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. In: The Twelfth International Conference on Learning Representations (2024)
49. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3d generation. In: The Twelfth International Conference on Learning Representations (2024)
50. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of caption. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7454–7464 (2020)
51. Shi, Z., Peng, S., Xu, Y., Geiger, A., Liao, Y., Shen, Y.: Deep generative models on 3d representations: A survey. arXiv preprint arXiv:2210.15663 (2023)
52. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022)
53. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation (2024)
54. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
55. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. In: Computer Graphics Forum. vol. 41, pp. 703–735. Wiley Online Library (2022)
56. Wang, P., Xu, D., Fan, Z., Wang, D., Mohan, S., Iandola, F., Ranjan, R., Li, Y., Liu, Q., Wang, Z., Chandra, V.: 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). pp. 9037–9047 (June 2024)
57. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. NeurIPS (2023)
58. Weng, H., Yang, T., Wang, J., Li, Y., Zhang, T., Chen, C., Zhang, L.: Consistent123: Improve consistency for one image to 3d object synthesis. arXiv preprint arXiv:2310.08092 (2023)
59. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7467–7477 (2020)
60. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
61. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research (2022)
62. Yu, X., Guo, Y.C., Li, Y., Liang, D., Zhang, S.H., QI, X.: Text-to-3d with classifier score distillation. In: The Twelfth International Conference on Learning Representations (2024)
63. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2023)
64. Zhu, J., Zhuang, P., Koyejo, S.: HIFA: High-fidelity text-to-3d generation with advanced diffusion guidance. In: The Twelfth International Conference on Learning Representations (2024)