Grounded Compositional and Diverse Text-to-3D with Pretrained Multi-View Diffusion Model

Xiaolong Li Jiawei Mo Ying Wang Chethan Parameshwara Xiaohan Fei
Ashwin Swaminathan CJ Taylor Zhuowen Tu
Paolo Favaro Stefano Soatto
AWS AI Labs
lxiaolx@amazon.com

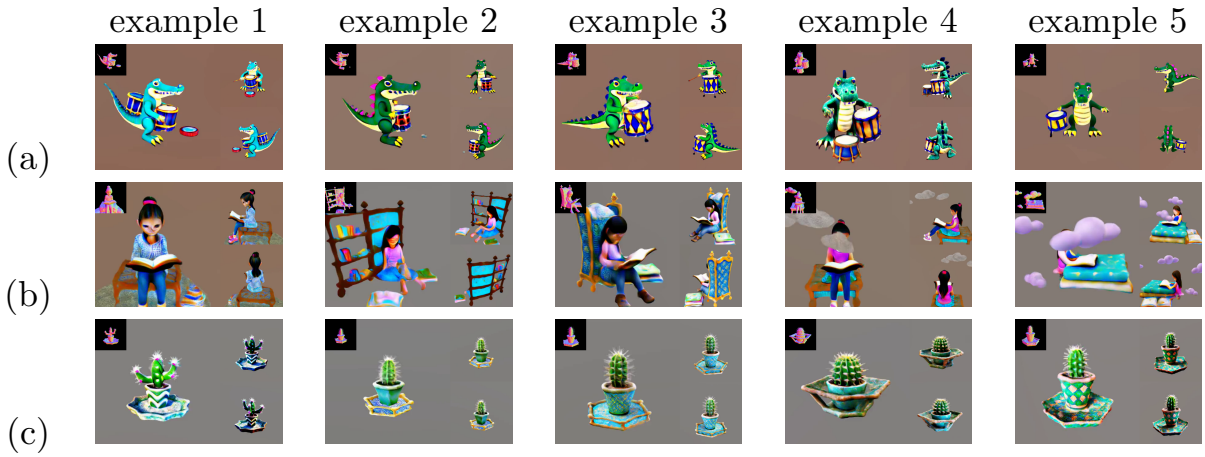


Figure 1. *Diverse 3D assets generated with Grounded-Dreamer* . Text prompts: (a) "a crocodile playing a drum set", (b) "A girl is reading a hardcover book in her room", (c) "a green cactus in a hexagonal cup on a star-shaped tray".

Abstract

In this paper, we propose an effective two-stage approach named Grounded-Dreamer to generate 3D assets that can accurately follow complex, compositional text prompts while achieving high fidelity by using a pre-trained multi-view diffusion model. Multi-view diffusion models, such as MVDream[35], have been shown to generate high-fidelity 3D assets using score distillation sampling (SDS). However, applied naively, these methods often fail to comprehend compositional text prompts, and may often entirely omit certain subjects or parts. To address this issue, we first advocate leveraging text-guided 4-view images as the bottleneck in the text-to-3D pipeline. We then introduce an attention refocusing mechanism to encourage text-aligned 4-view image generation, without the necessity to re-train the multi-view diffusion model or craft a high-quality compositional 3D dataset. We further propose a hybrid optimization strategy to encourage synergy between the SDS loss and the sparse RGB reference images. Our method consistently outperforms previous state-of-the-art (SOTA) methods in generating compositional 3D assets, excelling in both quality and accuracy, and enabling diverse 3D from the same text prompt.

1. Introduction

The quest to transform textual descriptions into vivid 3D models has seen remarkable advancements with methods like Score Jacobian Chaining [40], DreamFusion [29], and subsequent developments like [6, 19, 35]. However, the field still faces significant challenges in accurately rendering compositional prompts and ensuring diversity in the synthesized objects. Our research introduces a novel approach that not only addresses these challenges but also represents a paradigm shift in text-to-3D synthesis.

We draw inspiration from the 2D domain, where the same pre-trained diffusion models can generate compositionally correct images under multiple attempts. These im-

ages can serve as robust references for 3D synthesis, leading us to the key question: can we leverage these text-conditioned, diverse, compositionally correct views to enhance 3D asset creation? A naive solution is to combine text-to-image (T2I) and single-image-to-3D pipelines, such as [21, 22, 30, 34]. However, they often result in inconsistent geometry or semantics mainly due to inherent ambiguities and the domain gap when conditioning on a single image. This inconsistency is particularly evident in 3D assets where different views (front, side, rear) appear incongruent.

In our approach, we advocate for establishing a more robust foundation for Text-to-3D synthesis by utilizing multi-view images. Instead of relying on a single view, which often leads to ambiguities and inconsistencies, we generate four spatially distinct views, each separated by 90 degrees. This multi-view approach effectively constrains and defines an object's shape and appearance, bridging the gap between 2D imagery and 3D modeling. By employing a pre-trained multi-view diffusion model [35], we can generate these four views from a text prompt in a multi-view consistent manner. This process of generating and utilizing multiple views provides a more reliable and "grounded" basis for 3D reconstruction, as it reduces the uncertainty often associated with interpreting and extrapolating from a single image.

However, akin to the limitations of Stable Diffusion [31] in generating compositional single-view images [12], advanced models like MVDream can also struggle to consistently produce four-view images that accurately capture the correct compositional subjects, attributes, and their spatial relationships. To counter this, our first stage employs an attention refocusing mechanism during the inference phase, as inspired by [5]. This strategy ensures that each subject token from the text is precisely represented across all views, effectively addressing the ambiguities common in single-view reconstructions. By enhancing compositional accuracy without the need for re-training or fine-tuning the existing multi-view diffusion model, our method not only conserves resources but also leverages the rich knowledge embedded in the pre-trained text-guided diffusion model. This approach promotes greater adaptability and creativity in a wide range of scenarios.

In the second stage of our method, we implement a nuanced, coarse-to-fine reconstruction process. This stage is characterized by an integration of sparse-view Neural Radiance Fields (NeRF) with text-guided diffusion priors. The process begins by establishing a coarse 3D structure using sparse-view NeRF, grounded in the compositional accuracy achieved in the first stage. We then refine the details of this structure by introducing text-guided diffusion priors. A critical component of this stage is the implementation of a delayed Score Distillation Sampling (SDS) loss, coupled with an aggressively annealed timestep schedule. This combination is designed to refine textures and geometries in a scene-agnostic manner, ensuring that the enhancements do not distort the compositional accuracy established earlier.

It is important to note that a straightforward combination of sparse-view image supervision with existing Text-to-3D pipelines can lead to significant geometric distortions, such as the duplication of body parts (commonly referred to as the 'Janus' issue) or a complete disregard for compositional priors, resulting in a regression to the original MVDream Text-to-3D outputs. These common failure patterns, as illustrated in Fig. 3, underscore the need for a more sophisticated approach to integrating these elements. Our method's staged process, with its careful balance of NeRF and diffusion priors, is designed to avoid these pitfalls, ensuring a coherent and accurate 3D representation.

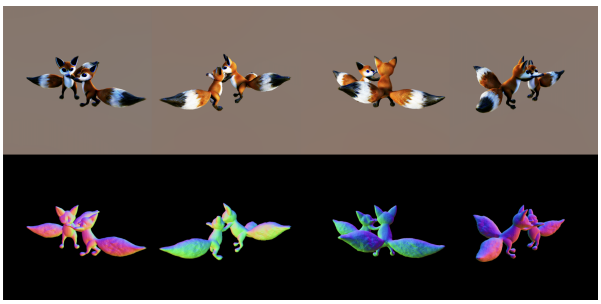
By integrating our novel two-stage framework with a pre-trained multi-view diffusion model, we develop an effective pipeline for compositional Text-to-3D synthesis that accurately adheres to complex text prompts. Our method not only generates diverse 3D assets for the same text prompts by varying the sets of four-view images but also marks significant advancements in the field:

- Innovative Two-Stage Framework: We introduce a new paradigm in Text-to-3D synthesis, where sparse-view images generated from a Text-to-Image (T2I) model serve as an intermediary, ensuring the preservation of compositional priors and facilitating diverse 3D generation.
- Compositional Alignment via Test-Time Optimization: Our method includes a novel test-time optimization technique for multi-view generation, significantly improving text-image alignment, particularly in terms of compositional accuracy.
- Hybrid Training Strategy for High-Fidelity 3D Assets: We propose a synergistic training approach that combines few-shot NeRF with Score Distillation Sampling (SDS)-based optimization. This strategy not only achieves high-fidelity, text-guided 3D asset generation but also maintains precise compositional relationships.

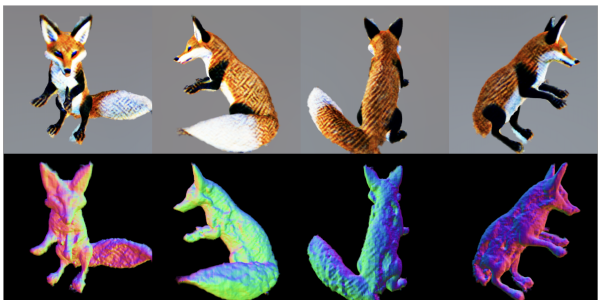
2. Related Works

2.1. Text-to-3D

Early works like CLIP-Mesh [27] and DreamField first show the possibility of generating 3D assets with text prompts using 2D priors. Later Score Distillation Sampling (SDS) is proposed by Boole *et al*. in DreamFusion [29] and followed by many [6, 7, 17, 19, 25, 35, 36, 38, 39, 41, 44, 47]. The key idea is to supervise NeRF [3, 26] training using the supervision signals from a pre-trained and frozen large text-to-image model [31, 32], which can be considered as distilling deterministic generators (i.e. neural radius field) as student models for a pre-trained large-scale diffusion models conditioned on specific text [23]. Several techniques have been proposed to improve the SDS frame-



(a) Running into the Janus issue



(b) Ignoring compositional priors

Figure 3. *Illustration of common failure patterns when naively combining sparse reference images to the SDS loss*. Text prompt: "Two foxes fighting". When combining 4 reference images, (a) ends up generating a fox with two tails, while (b) misses the 'two' information.

work including better viewpoint conditioning [1], better timestep scheduling [13], variational score distillation [44], accelerated NeRF representation [28], surface representation [33, 42, 46], improved efficiency using Gaussian Splatting [14], and improved fidelity [48]. However, these methods only demonstrate limited capability of generating compositional or diverse 3D assets.

2.2. Sparse Image-to-3D with Diffusion Models

A number of image-to-3D methods in existing literature attempt to synthesize 3D models from a single image, including RealFusion [24], Zero-1-to-3 [20], Magic123 [30], and Wonder3D [22]. Such methods are often used as part of a two-stage text-to-3D pipeline in which an image is first generated using a high quality text-to-image model, and subsequently used as a reference for 3D object synthesis. Although such approaches bear the advantage of the user being able to "select" the desired aesthetics before 3D synthesis commences, a single image is unable to fully capture multi-object compositions in 3D space in which objects may occlude each other, or when the prompt describes details that are impossible to capture from a single camera view, and easily becomes a bottleneck in preserving 3D compositional accuracy in the full text-to-3D pipeline.

2.3. 3D Compositional Generation

Previous works on compositional generation can be divided into two major tracks, layout or depth-conditioned 3D scene generation, like [2, 4, 37, 43, 45], or perform iterative 3D editing to add new compositional attributes [8, 18, 49]. The first track usually relies on a domain-specific dataset to learn scene priors, and the focus is different from ours on generating compositional 3D assets. While the second track can only add compistional attributes to the single target object through iterative editing, our methods can generate 3D assets with multiple compositional subjects or attributes in single round of training.

3. Preliminaries

3.1. Attend-and-Excite Revisited

Attend-and-Excite [5] was originally proposed to ease the catastrophic neglect issue in Text-to-Image generation domain, in which the text-guided image diffusion model can fail to generate one or more subjects specified in the target text prompt. Attend-and-Excite [5] comes up a test-time optimization framework over the noisy latents, and encourages the cross-attention layers to attend to all subjects in the text during the iterative denoising process.

The intuition lies in the adopted cross-attention mechanism to bring in text condition into image generation. At each timestep t, the text embeddings will be fed into the cross-attention layer of the U-Net, and each latent feature over the feature grid will perform attention operation with all the text embeddings, resulting an attention activation matrix per text token. The attention matrix can be reshaped to obtain a spatial map $A_t^s \in \mathcal{R}^{H \times W}$ per text token s. Intuitively, for a token to be manifested in the generated image, there should be at least one patch in its map with a high activation value. To guide such desired behavior, Attend-and-Excite introduces $f(z_t) = \mathcal{L}_{att} = \max_{s \in \mathcal{S}} \mathcal{L}_s, \mathcal{L}_s = 1 - \max(Gaussian(A_t^s))$, where z_t is the noisy latent. $Gaussian$ denotes applying Gaussian smoothing to the 2D activation map in order to cover a larger patch that later can emerge to the target objects. Such a loss will strengthen the activations of the most neglected subject token at the current timestep t.

3.2. Multi-View Diffusion Models

MVDream is a recent effort that adapts the common Text-to-Image diffusion model to have multi-view consistency, and enables Janus-free and high-fidelity Text-to-3D. Given a set of noisy images $\mathbf{x}_t \in \mathrm{R}^{F \times H \times W \times C}$, a text prompt as condition $\mathbf{y}$, and a set of extrinsic camera parameters $\mathbf{c} \in \mathrm{R}^{F \times 16}$, MVDream is trained to simultaneously denoise

and generate multiple images $\mathbf{x}_0 \in \mathrm{R}^{F\times H\times W\times C}$ that correspond to F different views of the same scene. At each step t, we have the predicted noise as $\epsilon_\theta(\mathbf{x}_t; \mathbf{y}, \mathbf{c}, t)$, where θ denotes the parameters of the latent U-Net. To inherit the generalizability of the 2D diffusion models, while also obtaining the capability of multi-view consistency, MVDream fine-tunes on Stable Diffusion v2.1. However, MVDream also inherits the same issue as Stable Diffusion and can fail in generating compositionally correct 4-view images, and the lack of large-scale compositional scene-level 3D data for fine-tuning makes the issue more significant when applying to Text-to-3D.

4. Method

To tackle the specific challenges when generating compositionally correct 3D assets while achieving diversity, we propose a novel 2-stage approach that well incorporates attention refocusing mechanism and sparse-view guidance in a unified framework, as drawed in Fig. 4. In Sec. 4.1, we detail our method for generating compositionally accurate four-view images. This stage focuses on ensuring that the subjects within these images are not only compositionally correct but also maintain the correct spatial relationships. Following this, in Sec. 4.2, we explore the integration of these consistent reference images with a pre-trained multi-view diffusion model. Here, we examine the optimal combination of sparse-view reference images and Score Distillation Sampling (SDS) loss to achieve high-fidelity 3D asset generation.

4.1. Attention Refocusing for Accurate Compositional 4-View Generation

Despite the success of Attend-and-Excite in Text-to-Image generation with multiple subjects, it is non-trivial extending the attention refocusing to Text-to-3D generation to optimize the target NeRF. Instead of optimizing a single-view latent, now we need to jointly optimize the 4-view latents without breaking the multi-view consistency, and we don't want to result in latent updates that lead to the latent becoming out-of-distribution. While it looks appealing to directly train a NeRF with combined attention refocusing loss and SDS loss, it renders a more challenging optimization problem since the attention refocusing loss is not optimizing the NeRF directly but on the rendered noisy latents from NeRF. The asynchronous NeRF updates can easily violate the assumption of in-distribution noisy latents on the attention refocusing loss. As we will show in the ablation study Sec. 6, such an attempt can easily lead to sub-optimal solutions and significantly enlarge the convergence time.

To design an more effective paradigm for composition control in Text-to-3D, we thus first adapt attention refocusing mechanism into compositionally correct 4-view generation, then we use the sparse-view images as additional compositional constraints in the 2nd-stage SDS-based NeRF training. When using the pre-trained multi-view diffusion model to generate 4-view images following a text prompt, we will have attention activation map $A_t^s \in \mathcal{R}^{F\times H\times W}$ per text token s, F is the number of frames. Instead of naively updating the per-view latent using a per-view $\mathcal{L}_{att}$, we found that if first aggregating the attention maps across the 4 views using average operation, we tend to get more reasonable 4-view images, in which the final loss is

$$\mathcal{L}_{att} = \max_{s\in\mathcal{S}}(1 - \max(\text{mean}(Gaussian(A_t^s[v, :, :])))). \tag{1}$$

Our final algorithm for applying attention refocusing in multi-view generation is drawn in Algorithm 1. Compared to Attend-and-Exicte, we also perform such optimization at more timesteps especially on the early stage instead of a few selected steps, which can still be done in minutes.

Algorithm 1 A Single Denoising Step on Compositional 4-View Generation

Input: A text prompt $\mathbf{y}$, 4-view camera poses $\mathbf{c}$, a set of subject token indices $\int$, a timestep t, a set of iterations for refinement $\{t_1, \cdots, t_k\}$, a set of thresholds $\{T_1, \cdots, T_k\}$, and a trained multi-view diffusion model SD_{mv}.

Output: A noised latent z_{t-1} for the next timestep

$\_, A_t \leftarrow SD_{mv}(z_t, \mathbf{y}, \mathbf{c}, t)$
$A_t \leftarrow Softmax(A_t - \langle sot \rangle)$
for $s \in \mathcal{S}$ **do**
  $A_t^s \leftarrow Mean_v(A_t[v, :, :, s])$
  $A_t^s \leftarrow Gaussian(A_t^s)$
  $\mathcal{L}_s \leftarrow 1 - max(A_t^s)$
end for
$\mathcal{L} \leftarrow max_s(\mathcal{L}_s)$
$z_t' \leftarrow z_t - \alpha_t \cdot \Delta_{z_t}\mathcal{L}$
if $t \in \{t_1, \cdots, t_k\}$ **then**
  if $\mathcal{L} > 1 - T_t$ **then**
    $z_t \leftarrow z_t'$
    Go to Step 1
  end if
end if
$z_{t-1}, \_ \leftarrow SD_{mv}(z_t', \mathbf{y}, \mathbf{c}, t)$
return z_{t-1}

4.2. Coarse-to-Fine Synergistic Reconstruction With Diffusion Priors

We adopt an optimization-based reconstruction framework that leverages both the 4-view reference images, and a pre-trained multi-view diffusion model for priors-augmented reconstruction. To avoid running into the failure patterns as mentioned above, we hypothesize that the key lies in designing an effective training strategy that can create synergy between the two different supervision signals. Our key insight is that, the rough 4 reference views give coarse but nearly complete information about geometry of the target

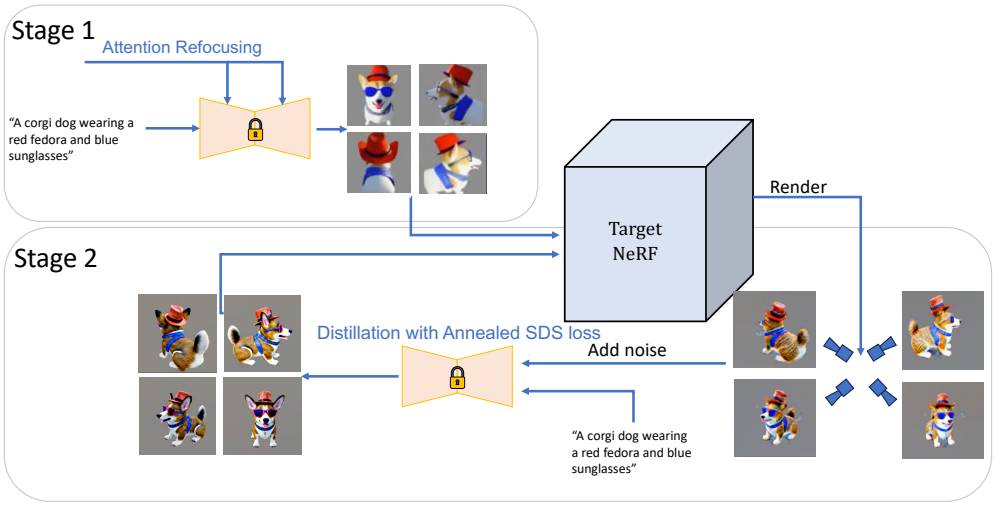


Figure 4. *Illustration of the two-stage pipeline with our Grounded-Dreamer.* Given a text prompt, we first generate compositionally correct 4-view images using iterative latent optimization at selected DDIM sampling steps. The 4-view reference images together with the masks are combined with score distillation sampling (SDS) loss in our hybrid training strategy, which will create high fidelity 3D assets while preserving the compositional priors accurately.

scene, especially on the compositional subjects, their interaction poses and spatial arrangement. These can provide a coarse initialization to later diffusion-based 3D distillation process.

Early few-shot NeRF training At early stages, we hypothesize that GT reference images can be more informative and stable compared to SDS loss with the pre-trained MVDream diffusion model given a large timestep. Thus we introduce sparse-view NeRF to establish the coarse geometry and texture. We adopt the hierarchical hash-grid MLP introduced in Instant-NGP [28] as our learnable NeRF representation, and we can simply rely on RGB and mask reconstruction loss which are sufficient to obtain a coarse NeRF representation. The reconstruction loss is defined as $\mathcal{L}_{img} = \mathcal{L}_{RGB} + \mathcal{L}_{mask}$.

3D distillation with warm-start SDS loss When the rough compositional geometry, and the associated texture emerge from the early few-shot NeRF training, we marry sparse-view NeRF with SDS-based 3D distillation to enable high-fidelity 3D generation, while preserving the compositional priors. The key idea is to bring in additional supervision from multiple unobserved viewpoints. In addition to the 4 fixed observed viewpoints that already have the roughly correct ground-truth images, we will randomly sample unobserved views and render multi-view images, the gradients come from SDS loss with a pre-trained multi-view diffusion model under text guidance. The total loss for NeRF training is defined as:

$$\mathcal{L}_{total} = \lambda_{img}(i) * \mathcal{L}_{img}^{\text{fixed views}} + \lambda_{SDS}(i) * \mathcal{L}_{SDS}^{\text{random views}} \tag{2}$$

$$\mathcal{L}_{SDS}^{\text{random views}} = \mathcal{E}_{t,\mathbf{c},\epsilon}[\omega(t)\|\epsilon_\theta(\mathbf{x}_t; y, \mathbf{c}, t(i)) - \epsilon)\|_2^2] \tag{3}$$

When the NeRF representation of the target scene is optimized to a certain level, further relying on poor-quality 4-view reference images may hinder the creation of high-fidelity 3D assets. Hence, we choose to gradually reduce the weight of the image reconstruction loss to 0, while increasing the weight of the SDS loss.

We have $t(i) \sim \mathcal{U}(T_{min}(i), T_{max}(i))$. In prior works [13, 35, 48] a time-annealing approach is implemented, where $T_{max}(i_{start})$ is set close to the total number of timesteps. We found such implementation can lead to large variation in SDS loss and drastic content change in NeRF output towards entirely different directions. It can raise the resurfacing alignment issues with the compositional priors drawn from the reference images. To solve this and preserve composition, the key modification we made is to set the initial $T_{max}(i_{start})$ as small as 680, thus the SDS loss can be leveraged to add more details, and refine NeRF to high fidelity. Specifically,

$$T_{max}(i) = c_1 + (c_2 - c_1) * \frac{i - a}{b - a} \tag{4}$$

The above hierarchical training design and modification to timestep annealing are simple but critical. It works effectively to have SDS loss and sparse image supervisions to make synergy between each other, an example is shown in

Fig. 5. We will next demonstrate through experiments and ablations.

5. Experiments

5.1. Experimental Setup

Implementation details We use the same settings for our method on all the text prompts. The early few-shot NeRF is trained for 200 steps, and then we add SDS loss by setting the $T_{max}(i)$ to linearly reduce from 680 to 500, while $T_{min}(i)$ is linearly reduced from 380 to 20. During the first 5000 steps we train the NeRF at 64x64 resolution, and switch to 256x256 for the latter 5000 steps for refinement. We gradually reduce the weighting on image reconstruction loss from 1000 to 100, while the weighting on SDS loss is increased from 0.025 to 0.25. For most of the baselines, we adopt default implementations within [9], and we run Wonder3D [22] using their released code, and use the scripts provided by Magic123 [30] for background removal. All the experiments are conducted on Nvidia A100 GPUs.

Prompts set To cover various scenarios of compositional Text-to-3D, we select and categorize our text prompts into (1) compositional-objects, which involve multiple subjects arranged in specific spatial relationships, e.g. "a green spoon on a red cake in a yellow tray"; (2) compositional-animals, which contain scenarios of animal-object interaction, or specific activities, e.g. "An artist is painting on a blank canvas" or "two foxes fighting". We select 50 for each subgroup based on existing text prompts from [29] and [8], with 100 text prompts in total.

Baselines We consider MVDream [35] as our baseline for multi-view generation. For text to 3D generation, we adopt recent SOTA Text-to-3D methods such as Magic3D [19], ProlificDreamer [44], and the MVDream-ThreeStudio [35] as the baselines. To further illustrate the unique advantages of 4-view input, we compare with recent single-view-to-3D methods like Magic-123 [30] by first using the text to generate a proper single-view image. To demonstrate the unique benefits of our hierarchical NeRF optimization with text guided diffusion priors, we also compare the results with a recent SOTA image-conditioned multi-view diffusion model named Wonder3D [22], which is essentially a single-view-to-4-view-to-3D method. Wonder3D learns 3D priors that are capable of generating normal maps and RGB images on 3 additional viewpoints, however, they adopt a normal-assisted few-shot NeRF to get the final 3D assets.

Metrics CLIP-based metrics might fail to measure the fine-grained correspondences between described objects and binding attributes, thus we only use CLIP R-Precision following [29], which measures the relative closeness between all generated images and their corresponding text prompts. Following a recent effort T^3Bench [10], we adopt a VQA-like pipeline by first using image-to-text models like BLIP-2 [16] to generate captions for rendered multi-view images, and then evaluate the alignment score between the compressed predicted captions and the given text prompt, named as T3 Score II. The purpose is to evaluate the capability of handling general text prompts with complex semantics in a fine-grained manner. We also conduct user study and add human preferences score on text-image alignment as additional metric. For measuring images realism, we compute the FID score following [11].

5.2. Enhanced Compositional 4-view Synthesis

We show both quantitative and qualitative results in Fig. 6 and Tab. 1. Our inference-stage editing method improves on the text-image alignment, while achieving comparable image realism.

Method	FID Score ↓	T3 Score II (%) ↑
MVDream	**71.43**	2.72
Ours	74.16	**2.85**

Table 1. *Quantitative evaluation on 4-view generation*. We run each method on 100 prompts with different random seeds, and compare the CLIP score of the generated images. Our inference-stage editing can generate more accurate images regarding the composition of different target subjects, while not breaking the multi-view consistency.

5.3. Compositional and Diverse 3D Generation

We show quantitative and qualitative comparisons in Tab. 2 and Fig. 7 for text-guided 3D generation with multiple subjects present. As detailed in Tab. 2, our method consistently outperforms existing SOTA baselines considering the overall performance of text-image alignment, view quality and consistency.

Specifically, our approach excels in the CLIP-R-Precision and T3 Score II metrics, indicative of superior performance in consistent text-guided 3D generation. In terms of view consistency, our method performs on par with MVDream while significantly exceeding all other models. Compared with MVDream, our method largely outperforms it for compositional generation. As illustrated in Fig. 7, MVDream frequently overlooks certain compositional elements, leading to incomplete or imprecise representations. In contrast, our method generates more compositionally complete views. Also thanks to the hierarchical optimization strategy, our method achieves high-fidelity 3D generation as evidenced by the FID score. ProlificDreamer, on the other hand, also achieves remarkable visual quality, producing sharp and highly detailed results, however it suffers from slow training speeds (refer to our supplementary material for an efficiency report) and is prone to issues such as

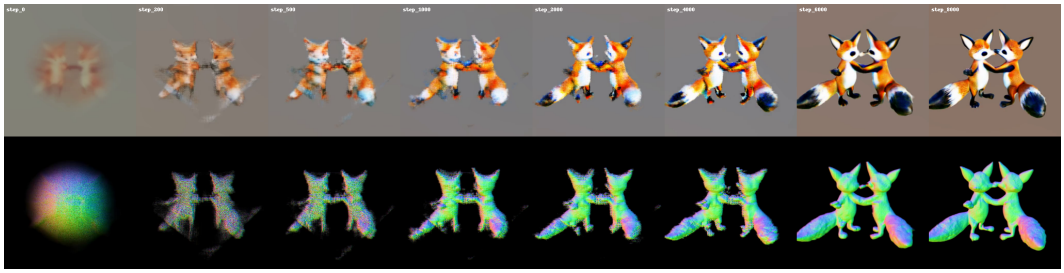


Figure 5. *Illustration on the 2nd-stage training progress with Grounded-Dreamer* . Here we are showing a fixed front-view rendering of the target NeRF at different optimization steps. Our method can gradually create high fidelity 3D assets while preserving the compositional priors accurately.

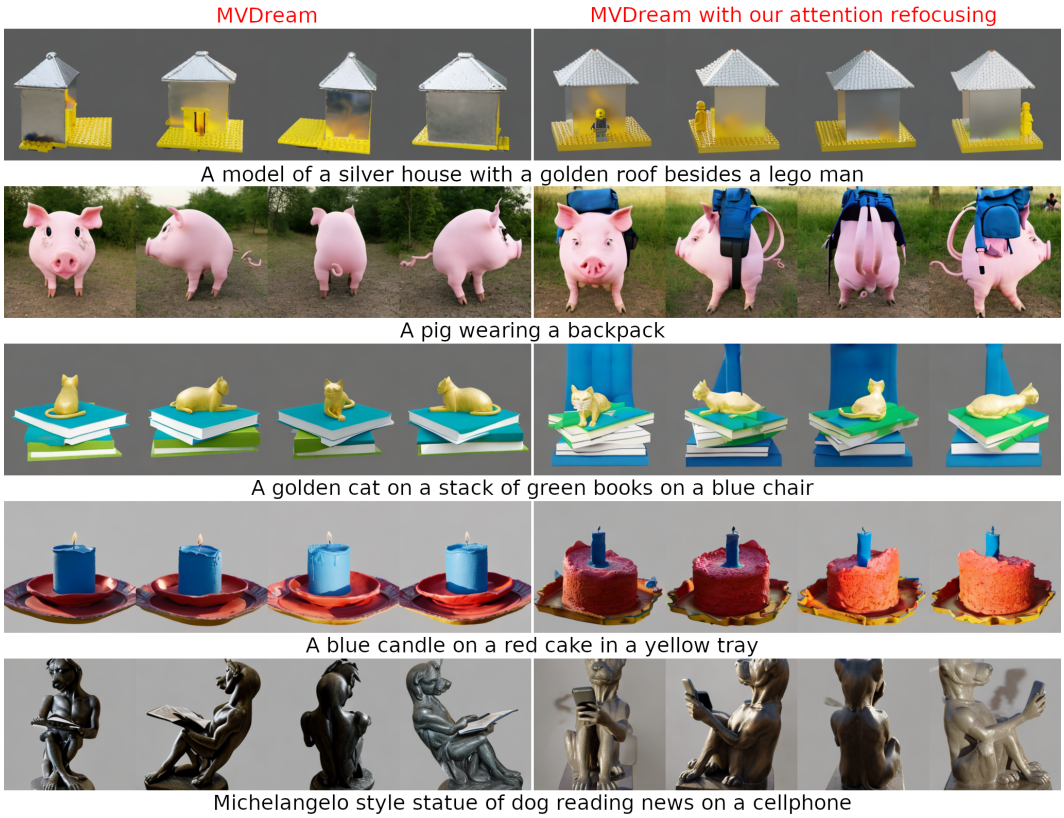


Figure 6. *4-view generation, each pair uses the same random seed.* Our inference-stage optimization encourages compositionally correct 4-view generation compared to the original MVDream.

corrupted flat geometries or 'Janus' problems, as shown in the case of row 4. From Tab. 2 we can also see its inferior view consistency performance.

Our method also largely outperforms single-image-to-3D approaches like Magic-123 [30], Wonder3D [22] on the benchmark text prompts. While Magic-123 can effectively reconstruct the front view of objects, it struggles with side or back perspectives, leading to incorrect anatomy, like in the case "a zoomed out DSLR photo of a chimpanzee holding a cup of hot coffee". Other recent works, such as Wonder3D, typically face challenges in reconstruction quality, particularly when relying solely on sparse-view images. In contrast, our method innovatively integrates text-guided natural image priors with sparse-view supervision. Our generated 3D assets can harvest the natural images priors from a pre-trained diffusion model, which not only enhances the quality of reconstruction but also ensures generalizability across various text prompts describing a wide range of 3D compositional subjects. The results demonstrate the capability of our method in achieving high-quality, composition-

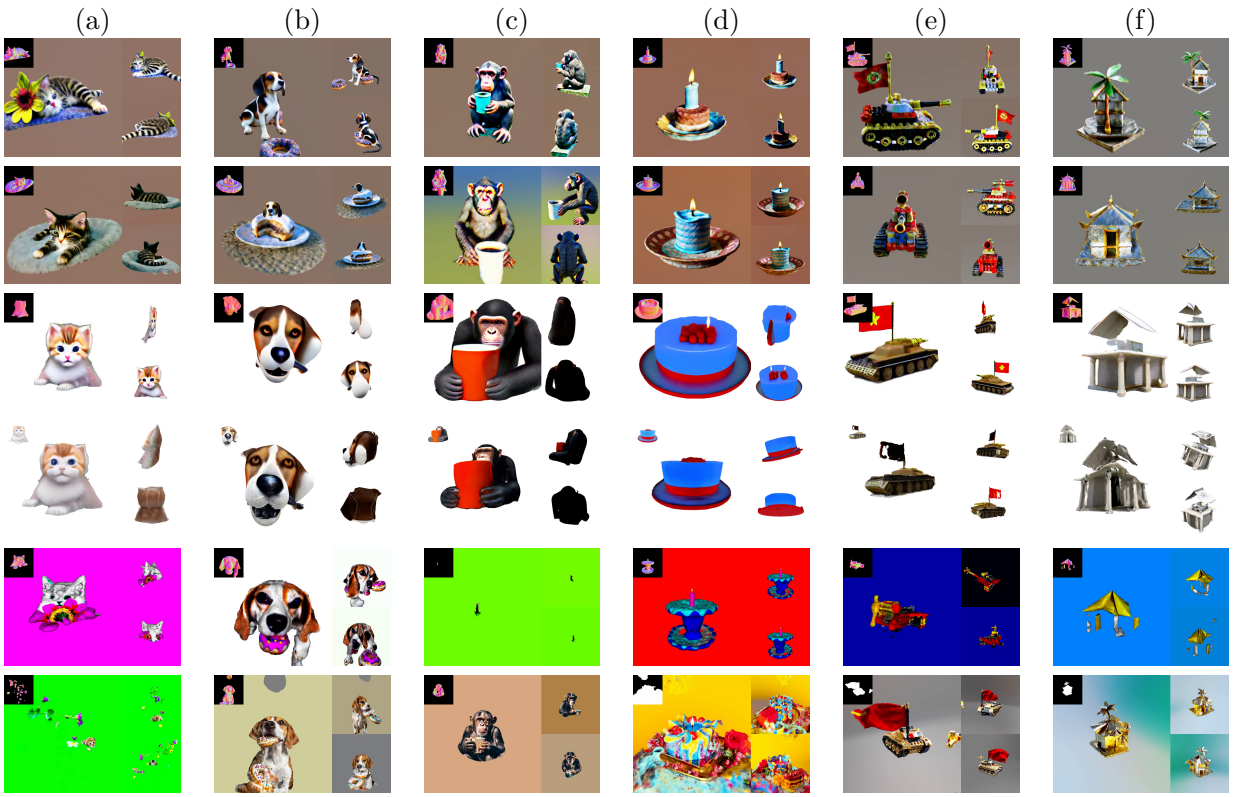


Figure 7. *Qualitative results comparison for compositional Text-to-3D*. From top to down, the methods are: our Grounded-Dreamer, MVDream [35], Magic123 [30], Wonder3D [22], Magic3D [19], ProlificDreamer [44]. Our method generates more compositionally complete views with high quality. Text prompts: (a) "a zoomed out DSLR photo of an adorable kitten lying next to a flower", (b) "a zoomed out DSLR photo of a beagle eating a donut", (c) "a zoomed out DSLR photo of a chimpanzee holding a cup of hot coffee", (d) "a blue candle on a red cake in a yellow tray", (e) "a lego tank with a golden gun and a red flying flag", (f) "a model of a silver house with a golden roof beside an origami coconut tree".

Method	T3 Score II ↑	CLIP R-P.(%) ↑	Good alignment ↑	Freq. of Janus↓	FID Score ↓
Magic3D	2.29/5.0	27.10	27.10	58.88	137.45
ProlificDreamer	**2.68/5.0**	**48.91**	**60.87**	77.17	129.11
Magic-123	2.32/5.0	24.74	28.87	64.95	121.16
Wonder3D	1.96/5.0	20.22	35.60	21.25	129.93
MVDream	2.33/5.0	44.95	44.44	**5.88**	**109.78**
Ours	**2.53/5.0**	**62.73**	**56.71**	**17.15**	**115.94**

Table 2. *Quantitative evaluation on 3D composition*. We run each method on 100 prompts with same random seeds. Under "Good alignment", for each method, we show the percentage of the generated 3D outputs that human reviewers annotate as aligning well with the text prompts. Under "Freq. of Janus", we show the preference ratio of generated outputs for each model in terms of view consistency.

ally accurate 3D reconstructions without compromising details and structural integrity.

Compared to ProlificDreamer, our method can generate diverse 3D assets in a well time-bounded manner, as shown in Fig. 1. The diversity can be easily controlled with different pairs of edited 4-view images as additional guidance.

6. Ablation Study

Method	FID Score ↓	CLIP R-P. (%) ↑	GPU-hours ↓
Ours-2-stage	115.94	**62.73**	**1.91**
Ours-1-stage a	114.68	48.18	3.90
Ours-1-stage b	**112.45**	40.91	5.25

Table 3. *Ablation with 1-stage designs*.

1-stage design with attention refocusing loss We also implement and train two 1-stage variants without first generating 4-view images, each with balanced losses: a) we directly add the attention score loss to the SDS loss; b) we first update latents using the attention loss, then we use the updated noisy latents for the SDS loss. Tab. 3 show the ablation results. Compared to our 2-stage approach: 1) the 1-stage ones significantly increase the training time by two to three times; 2) stark performance drop on compositional accuracy; 3) corrupted shapes, like the tree mixes with the house in the 2nd row of Fig. 8.

Different strategies for multi-view latent updates For example, if naively updating each view sequentially using Attend-and-Excite, we can end up with corrupted views as shown in Fig. 9. Our later mean operation across the 4 views

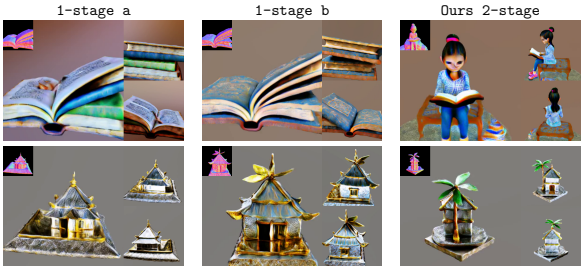


Figure 8. *Results visualization with different pipeline designs.*

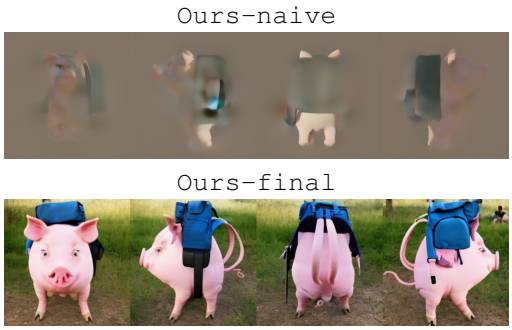


Figure 9. *Examples on "a pig wearing a backpack".*

latents is simple and already works sufficiently well on creating compositionally correct 4-view images.

Effects of warm-start timestep in SDS loss One of the keys to our method's success is to set the initial timestep of SDS loss to be relatively smaller, so the SDS loss can guide NeRF towards adding more geometric and texture details, while preserving the compositional priors. As shown in Fig. 10, if following the default setting like $[T_{min}(t_0), T_{max}(t_0)] = [0.98, 0.98]$ out of 1000, or$[T_{min}(t_0), T_{max}(t_0)] = [0.74, 0.86]$, it can miss the compositional information like "two" in the first example, or "beside an origami coconut tree" in the 5th example.

Figure 10. *3D generation under different initial timestep sampling range.* We pick 5 text prompts, and visualize the reference 4-view images, and two side-view of the generated 3D assets with associated normal images attached to the top-left corner.

Effects of backbone used in SDS loss To validate the effects of using pre-trained multi-view diffusion model, we can replace our Score Distillation Sampling (SDS) loss backbone with SD v2.1. The results, as depicted in the accompanying Tab. 4, demonstrate that using pre-trained multi-view diffusion model for SDS loss help consistently yield better textural quality.

Method	Freq. of Janus (%) ↓	FID Score ↓
ours with SD	40.90	135.60
Ours	**17.15**	**115.94**

Table 4. *Text-to-3D with different pretrained T2I models.*

7. Conclusion

In summary, our work introduces a novel two-stage framework for Text-to-3D synthesis, effectively overcoming challenges in compositional accuracy and diversity. The first stage leverages a multi-view diffusion model for generating spatially coherent views from text, while the second stage synergizes sparse-view NeRF with text-guided diffusion priors for refined 3D reconstruction. This approach not only enhances the fidelity and compositional integrity of 3D models from complex text prompts, but also paves the way for future explorations in seamless 2D-to-3D transitions and model versatility. Our method demonstrates a significant leap in Text-to-3D synthesis, offering a robust foundation for further advancements in this evolving field.

References

[1] Mohammadreza Armandpour, Huangjie Zheng, Ali Sadeghian, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023. 3

[2] Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. *arXiv preprint arXiv:2303.12074*, 2023. 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[4] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022. 3

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 3

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 1, 2

[7] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2

[8] Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts. *arXiv preprint arXiv:2310.11784*, 2023. 3, 6

[9] 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

[10] Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023. 6

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

[12] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 2

[13] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3, 5

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[17] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *arXiv preprint arXiv:2310.02596*, 2023. 2

[18] Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608*, 2023. 3

[19] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1, 2, 6, 8

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[21] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[22] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2, 3, 6, 7, 8

[23] Weijian Luo. A comprehensive survey on knowledge distillation of diffusion models. *arXiv preprint arXiv:2304.04262*, 2023. 2

[24] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023. 3

[25] 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*, pages 12663–12673, 2023. 2

[26] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[27] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3, 5

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

[30] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2, 3, 6, 7, 8

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

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[33] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[34] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2

[35] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 5, 6, 8

[36] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 2

[37] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Scene graph denoising diffusion probabilistic model for generative indoor scene synthesis. *arXiv preprint arXiv:2303.14207*, 2023. 3

[38] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[39] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*, 2023. 2

[40] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 1

[41] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[42] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3

[43] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, pages 106–115. IEEE, 2021. 3

[44] 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. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3, 6, 8

[45] Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. Lego-net: Learning regular rearrangements of objects in rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19037–19047, 2023. 3

[46] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3

[47] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 2

[48] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 3, 5

[49] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. *arXiv preprint arXiv:2306.13455*, 2023. 3

Grounded Compositional and Diverse Text-to-3D with Pretrained Multi-View Diffusion Model

Supplementary Material

In this supplementary material, we provide additional implementation details in Sec. 8, offering deeper insights into our methodology. For a comprehensive understanding of our method's efficiency, a comparative speed analysis against selected methods is detailed in Sec. 9. We also discuss the failure modes of our approach in Sec. 10, highlighting areas for potential improvement. Additionally, Sec. 12 includes an expanded set of qualitative visualizations, complementing those presented in Fig. 12, to better illustrate the capabilities and limitations of our method. Finally, you can find all text prompts we used for our experiments in Sec. 11.

8. Additional Implementation Details

During stage 1, we use the pre-trained multi-view text-to-image (T2I) diffusion model from [35] to perform inference-stage editing and generate 4 views, we denote this model as MVDream-T2I. Our offline editing approach treats the generation of 4-view, text-aligned images as a latent optimization problem. In this framework, the attention score map associated with each subject token forms the basis of our objective function. This function is then used to optimize the latent noise at selected steps. During the 50-step DDIM-based inference sampling, we specifically target iterations 0,1,2,3,4,5,10,20,30,40 for latent optimization. This process is capped at a maximum of 25 iterations, with a loss threshold set at 0.1. This additional optimization step does extend the inference time from an initial 10 seconds to approximately 45 seconds. In stage 2, we also use the pre-trained MVDream-T2I model for SDS loss, without doing any fine-tuning or re-training.

Metrics In our main paper Tab. 2, we adopt "Good alignment" and "Freq. of Janus", both are new metrics proposed and defined in another work that we are submitting. Basically "Good alignment" shows the percentage of the generated 3D models that that human reviewers annotate as aligning well with the text prompts, while "Freq. of Janus" measures how often the Janus problem appears in the generated 3D content.

9. Speed Comparison

While our primary focus is not on efficiency, our method is capable of generating high-quality 3D assets within a reasonable time of approximately 1.9 hours. In contrast, while ProlificDreamer [44] also achieves diverse results, it does so at the cost of time efficiency, requiring upwards of 15 hours to produce final 3D assets. Our method strikes a more balanced approach, offering both diversity and a more manageable time investment. We anticipate that future developments in this field will further enhance the efficiency of our method.

Method	GPU-hours ↓
`Magic3D` [19]	2.40
`ProlificDreamer` [44]	15.50
`MVDream` [35]	1.83
`Ours -- 1st stage`	0.02
`Ours -- full model`	1.91

Table 5. *GPU-hours needed to generate one 3D model given a prompt.*

10. Failure Cases

Our analysis identifies two primary failure modes. The first involves incomplete foreground segmentation, leading to the omission of key compositional elements. For instance, as depicted in Fig. 11, when the roof of a house is not included in three views of the segmented final images, our Stage 2 process is unable to fully compensate for these missing elements. We expect more advanced segmentation tool like SAM [15] to help mitigate such issue. The second failure pattern relates to inaccuracies in additional attributes, particularly on color attribute in our case. An example of this, shown in Fig. 11, is the inability of both the initial four-view generation and the final 3D asset to accurately represent specified colors, such as yellow and blue. These limitations predominantly stem from the current capabilities and precision of the diffusion-based Text-to-Image (T2I) model backbone.

11. List of Prompts

In our actual implementation, we end up with having 111 text prompts, with the first 55 for capturing interactive compositional animals, and the last 56 involving static compositional objects with specific spatial arrangement.

1. "a pig wearing a backpack"
2. "a blue poison-dart frog sitting on a water lily"
3. "a bumblebee sitting on a pink flower"
4. "a crocodile playing a drum set"
5. "a corgi dog wearing a red fedora and blue sunglasses"
6. "two foxes fighting"
7. "Michelangelo style statue of dog reading news on a cellphone"
8. "a tiger playing the violin"

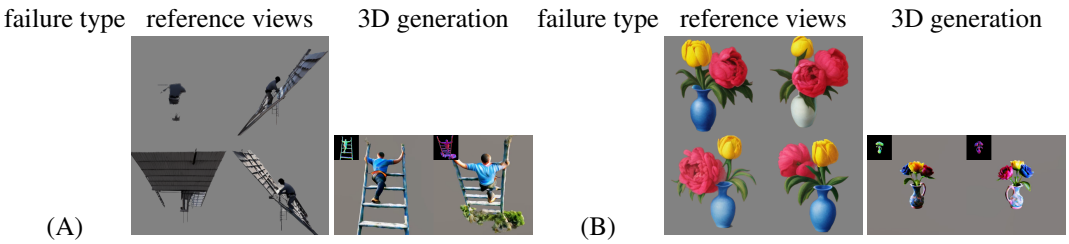


Figure 11. *Failure cases with Grounded-Dreamer* . We visualize the two side-view of the generated 3D assets with associated normal images attached to the top-left corner. Text prompts: (1) “A worker is climbing a ladder to repair a roof”; (2) “a blue peony and a red rose and a yellow tulip in a pink vase”.

9. “An artist is painting on a blank canvas”
10. “A student is typing on his laptop”
11. “A young gymnast trains with a balance beam”
12. “A chef is making pizza dough in the kitchen”
13. “A footballer is kicking a soccer ball”
14. “A man is holding an umbrella against rain”
15. “A girl is reading a hardcover book in her room”
16. “A woman putting lipstick on in front of a mirror”
17. “A worker is climbing a ladder to repair a roof”
18. “A florist is making a bouquet with fresh flowers”
19. “A boy is flying a colorful kite in the sky”
20. “A gardener is watering plants with a hose”
21. “A photographer is capturing a beautiful butterfly with his camera”
22. “A scientist is examining a specimen under a microscope”
23. “A drummer is beating the drumsticks on a drum”
24. “A fisherman is throwing the fishing rod in the sea”
25. “A baby is reaching for a teddy bear on the bed”
26. “a DSLR photo of a corgi wearing a beret and holding a baguette, standing up on two hind legs”
27. “a DSLR photo of a fox holding a videogame controller”
28. “a DSLR photo of a ghost eating a hamburger”
29. “a DSLR photo of a humanoid robot using a laptop”
30. “a DSLR photo of a koala wearing a party hat and blowing out birthday candles on a cake”
31. “a DSLR photo of Two locomotives playing tug of war”
32. “a lego man wearing a chefŽ2019s hat and riding a golden motorcycle”
33. “a lego man wearing a silver crown and riding a golden motorcycle”
34. “a lego man wearing a wooden top hat and riding a golden motorcycle”
35. “a metal monkey wearing a chefŽ2019s hat and driving an origami sport car”
36. “a metal monkey wearing a golden crown and driving an origami sport car”
37. “a metal monkey wearing a wooden top hat and driving an origami sport car”
38. “a zoomed out DSLR photo of a colorful camping tent in a patch of grass”
39. “a zoomed out DSLR photo of a dachsund riding a unicycle”
40. “a zoomed out DSLR photo of a hippo biting through a watermelon”
41. “a zoomed out DSLR photo of a monkey riding a bike”
42. “a zoomed out DSLR photo of a pig playing the saxophone”
43. “a zoomed out DSLR photo of a raccoon astronaut holding his helmet”
44. “a zoomed out DSLR photo of a tiger eating an ice cream cone”
45. “a zoomed out DSLR photo of an adorable kitten lying next to a flower”
46. “a wide angle zoomed out DSLR photo of A red dragon dressed in a tuxedo and playing chess. The chess pieces are fashioned after robots”
47. “a zoomed out DSLR photo of a beagle eating a donut”
48. “a zoomed out DSLR photo of a chimpanzee holding a cup of hot coffee”
49. “an orange cat wearing a yellow suit and cyan boots”
50. “an orange cat wearing a yellow suit and green sneakers”
51. “an orange cat wearing a yellow suit and green sneakers and pink cap”
52. “an orange cat wearing a yellow suit and red pumps”
53. “a golden cat on a stack of green books on a blue chair”
54. “A black cat sleeps peacefully beside a carved pumpkin”
55. “a monkey-rabbit hybrid”
56. “a blue candle on a red cake in a yellow tray”
57. “a blue peony and a red rose and a yellow tulip in a pink vase”
58. “a blue peony and a yellow tulip in a pink vase”
59. “a bunch of colorful marbles spilling out of a red velvet bag”
60. “a ceramic tea pot and a cardboard box on a golden table”
61. “a ceramic tea pot and a lego car on a golden table”
62. “a ceramic tea pot and a pair of wooden shoes on a golden table”
63. “a ceramic tea pot and an origami box and a green apple on a golden table”
64. “a ceramic tea pot and an origami box on a golden table”
65. “a ceramic upside down yellow octopus holding a blue green ceramic cup”
66. “a purple rose and a red rose and a yellow tulip in a pink vase”
67. “a red rose in a hexagonal cup on a star-shaped tray”
68. “a white tulip and a red rose and a yellow tulip in a pink vase”
69. “a zoomed out DSLR photo of a beautifully carved wooden knight chess piece”
70. “an orchid flower planted in a clay pot”
71. “a nest with a few white eggs and one golden egg”
72. “a pink peach in a hexagonal cup on a round cabinet”
73. “a plate of delicious tacos”
74. “a DSLR photo of a bagel filled with cream cheese and lox”
75. “a DSLR photo of a beautiful violin sitting flat on a table”

76. “a DSLR photo of a candelabra with many candles on a red velvet tablecloth”
77. “a DSLR photo of a Christmas tree with donuts as decorations”
78. “a DSLR photo of a cup full of pens and pencils”
79. “a DSLR photo of a delicious chocolate brownie dessert with ice cream on the side”
80. “a DSLR photo of a pair of headphones sitting on a desk”
81. “a DSLR photo of a quill and ink sitting on a desk”
82. “a DSLR photo of A very beautiful tiny human heart organic sculpture made of copper wire and threaded pipes, very intricate, curved, Studio lighting, high resolution”
83. “a DSLR photo of a very cool and trendy pair of sneakers, studio lighting”
84. “a DSLR photo of a wooden desk and chair from an elementary school”
85. “a lego tank with a golden gun and a blue flying flag”
86. “a lego tank with a golden gun and a red flying flag”
87. “a lego tank with a golden gun and a yellow flying flag”
88. “a green apple in a hexagonal cup on a round cabinet”
89. “a green apple on a yellow tray on a red desk”
90. “a green cactus in a hexagonal cup on a star-shaped tray”
91. “a green spoon on a red cake in a yellow tray”
92. “a model of a round house with a spherical roof on a hexagonal park”
93. “a model of a round house with a spherical roof on a square park”
94. “a model of a silver house with a golden roof beside a lego man”
95. “a model of a silver house with a golden roof beside a wooden car”
96. “a model of a silver house with a golden roof beside an origami coconut tree”
97. “A candle burns beside an ancient, leather-bound book”
98. “An apple lays nestled next to a vintage, brass pocket watch”
99. “A quill pen lies across a stack of unmarked parchment paper”
100. “A wine bottle and two empty glasses glisten under a chandelier”
101. “A magnifying glass sits on top of a mysterious, printed map”
102. “A weathered straw hat hangs beside a freshly picked sunflower”
103. “An open diary lays flat, a single dried rose on its pages”
104. “A twinkling star ornament hangs closely with a snow globe”
105. “A sandy hourglass and a rugged compass lay side by side”
106. “A pair of spectacles lies open on a dog-eared paperback”
107. “Ripe apples cluster next to a gleaming knife”
108. “An abandoned teddy bear leans against a discarded toy car”
109. “A bottle of red wine stands alongside an empty wine glass”
110. “A colorful array of spices in tiny jars sits next to an unused cooking book”
111. “fries and a hamburger

12. More Qualitative Results

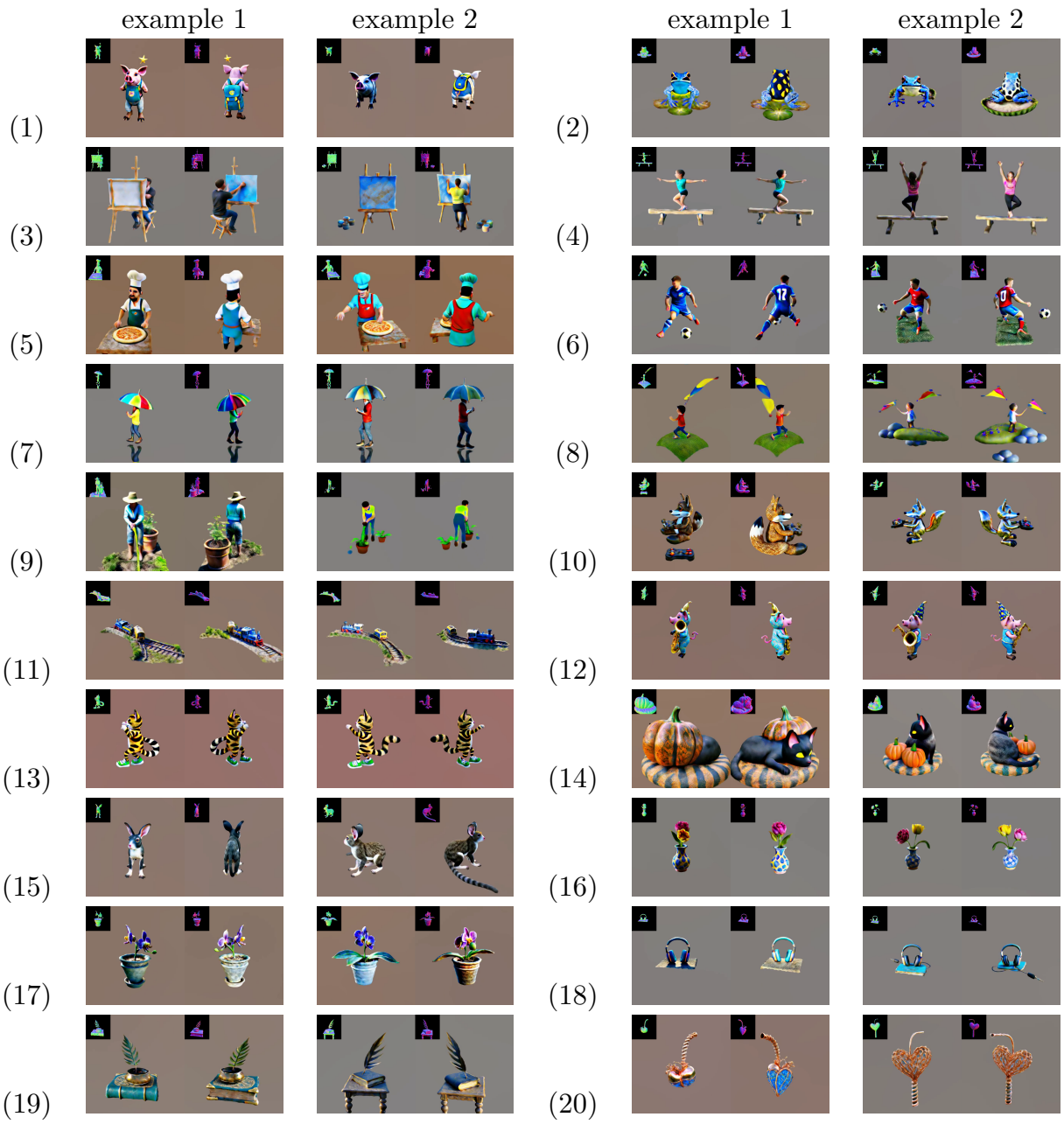


Figure 12. *Diverse 3D assets generated with Grounded-Dreamer* . We show results for 20 text prompts in total, and for each text prompt, we visualize two different generated samples, using their two side-view with associated normal images attached to the top-left corner. (1) "a pig wearing a backpack", (2) "a blue poison-dart frog sitting on a water lily", (3) "An artist is painting on a blank canvas", (4) "A young gymnast trains with a balance beam", (5) "A chef is making pizza dough in the kitchen", (6) "A footballer is kicking a soccer ball", (7) "A man is holding an umbrella against rain", (8) "A boy is flying a colorful kite in the sky", (9) "A gardener is watering plants with a hose", (10) "a DSLR photo of a fox holding a videogame controller", (11) "a DSLR photo of Two locomotives playing tug of war", (12) "a zoomed out DSLR photo of a pig playing the saxophone", (13) "an orange cat wearing a yellow suit and green sneakers and pink cap", (14) "A black cat sleeps peacefully beside a carved pumpkin", (15) "a monkey-rabbit hybrid", (16) "a blue peony and a yellow tulip in a pink vase", (17) "an orchid flower planted in a clay pot", (18) "a DSLR photo of a pair of headphones sitting on a desk", (19) "a DSLR photo of a quill and ink sitting on a desk", (20) "a DSLR photo of A very beautiful tiny human heart organic sculpture made of copper wire and threaded pipes, very intricate, curved, Studio lighting, high resolution".