SPAD: Spatially Aware Multi-View Diffuser

Yash Kant[1,2], Ziyi Wu[1], Michael Vasilkovsky[2], Guocheng Qian[2,3], Jian Ren[2],
Riza Alp Guler[2], Bernard Ghanem[3], Sergey Tulyakov[2,*], Igor Gilitschenski[1,*], Aliaksandr Siarohin[2,*]
[1]University of Toronto, [2]Snap Research, [3] KAUST
https://yashkant.github.io/spad/

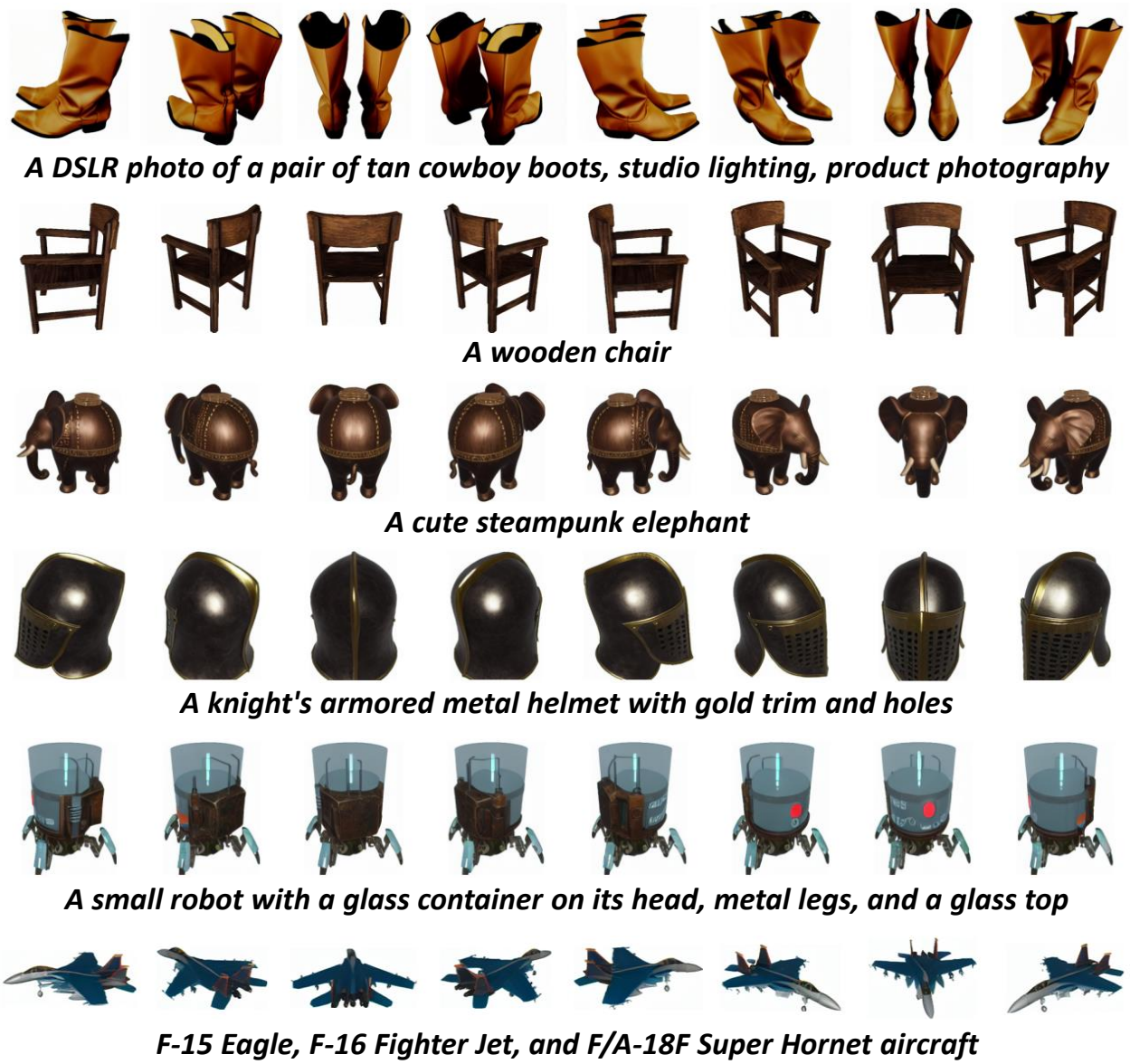


Figure 1. **Consistent multi-view generation from text with SPAD.** Given a text prompt, SPAD is capable of synthesizing many 3D consistent images of the same object, ranging from daily objects to highly complex machines. SPAD can generate many images from arbitrary camera viewpoints, while being trained only on four views. Here, we generate eight views sampled uniformly at a fixed elevation.

Abstract

We present SPAD, a novel approach for creating consistent multi-view images from text prompts or single images. To enable multi-view generation, we repurpose a pretrained 2D diffusion model by extending its self-attention layers with cross-view interactions, and fine-tune it on a high quality subset of Objaverse. We find that a naive extension of the self-attention proposed in prior work (e.g., MVDream [72]) leads to content copying between views. Therefore, we explicitly constrain the cross-view attention based on epipolar geometry. To further enhance 3D consistency, we utilize Plücker coordinates derived from camera rays and inject them as positional encoding. This enables SPAD to reason over spatial proximity in 3D well. Compared to concurrent works that can only generate views at fixed azimuth and elevation (e.g., MVDream [72], SyncDreamer [46]), SPAD offers full camera control and achieves state-of-the-art results in novel view synthesis on unseen objects from the Objaverse and Google Scanned Objects datasets. Finally, we demonstrate that text-to-3D generation using SPAD prevents the multi-face Janus issue.

1. Introduction

3D content generation holds great importance in a wide range of applications, including gaming, virtual reality, manufacturing, *etc*. Yet, the creation of high-quality 3D assets remains a time-consuming endeavor even for seasoned 3D artists. In recent years we have witnessed the emergence of generative models capable of creating 3D objects from a single or several 2D images, or just text inputs. Early methods in this field directly train models such as Variational Auto Encoders (VAEs) [41] and Generative Adversarial Networks (GANs) [27] on 3D shapes [1, 48, 95, 97]. These methods produce results with lower resolution to manage computational demands and have limited diversity due to the small scale of training dataset. Later approaches explored differentiable rendering to learn 3D GANs from monocular images [7, 8, 14, 26, 50, 54, 75, 76]. These methods improved resolution, but only show impressive results on relatively few categories (e.g., ShapeNet [10] furniture).

Recent advances in Diffusion Models (DMs) have revolutionized the field of 2D image generation [21, 34, 53]. Trained on billions of image-text data, state-of-the-art models [63, 65] learn generic object priors that enable high-quality and generalizable text-guided image generation. Recent works thus seek to leverage such 2D priors to generate 3D objects. One line of research proposes to optimize a NeRF [49] model by distilling a pre-trained text-to-image DM via Score Distillation Sampling (SDS) [58, 89], which enables single-view 3D reconstruction [19, 47, 59] and direct text-to-3D synthesis [36, 44, 92]. However, these methods lack understanding of the underlying object structures. The 2D prior provided by pre-trained DMs only considers one view at each optimization step, ignoring the geometric relationship across views. Even with hand-crafted prompts specifying explicit viewpoints [58],these methods continue to exhibit 3D inconsistencies, exemplified by issues such as the multi-faced Janus problem.

One natural solution is to equip 2D diffusion models with some form of 3D understanding. Recent work Zero-1-to-3 [45] proposes to condition Stable Diffusion [63] with one view and generate another one given the relative camera pose. However, conditioning in Zero-1-to-3 is performed by simply concatenating the input view, while disregarding any geometric priors. An alternate approach based on depth warping was proposed in iNVS [39]. It shows that, provided with an accurate depth map, one can establish dense correspondences between two views. This allows DMs to reconstruct high-quality novel views. Unfortunately, the generation quality of iNVS heavily relies on the precision of depth maps, while monocular depth estimation in itself is an unsolved problem.

Recent works [82, 101] have observed that Stable Diffusion can be utilized to obtain accurate image correspondences. Self-attention layers of text-to-image DMs can be directly used for establishing correspondences *within* the same image [86]. An interesting question to consider is whether the same layers can also find correspondences *between* different views, which can enable 3D geometric understanding. For this, we can modify the original self-attention into multi-view self-attention by running it over the concatenated feature maps across views. This approach trained on orthogonal multi-view images with known camera parameters can generate multiple novel views of the same object simultaneously, as shown in previous works such as MVDream [72]. However, we find such a model lacks precise camera control across views, and cannot generate arbitrarily free novel views. When tasked to generate two views that are close to each other (with significant overlap), such a model suffers from content copying problem – where the content of one view is just copied from another view without modification (see Fig. 5).

Inspired by [78], we design an Epipolar Attention layer, where feature map positions in one view can only attend to positions along the epipolar lines in other views. By restricting the cross-view attention maps, these layers enable better camera control and produce different views at viewpoints close to each other. While Epipolar Attention alone significantly improves 3D consistency, since epipolar lines do not have a direction, it still remains difficult for this model to disambiguate the direction of the camera ray. This ambiguity leads to flipping in predicted views, as observed in

* Equal supervision

iNVS [39] which also used epipolar lines. Motivated by recent works in the Light Field Networks [6, 88], we propose to represent rays passing through each pixel in Plücker coordinates, which assigns unique homogeneous coordinates for each ray. We use these coordinates as positional embeddings inside Epipolar Attention layers. These embeddings for rays hitting opposite sides of the object provide a high negative bias for self-attention, essentially preventing it from utilizing information from the wrong side. Additionally, Plücker embeddings also encourages pixels whose rays are close to each other to have similar representations, thus promoting self-attention to pick features from nearby positions.

Our method can operate in two modes: text-conditioned and image-conditioned. In text-conditioned mode, SPAD can simultaneously denoise several views given a text prompt. While in image-conditioned mode, given an image, SPAD denoises several other views. In both cases, the architecture of our method stays the same, and only the input and output changes. We evaluate SPAD in the task of text-conditioned multi-view generation and image-conditioned novel view synthesis on Google Scanned Objects (GSO) [22] and an unseen subset of Objaverse [62] datasets. The results show that SPAD is able to synthesize high-quality and 3D consistent images of objects. Finally, we enable high-quality text-to-3D generation using SPAD via a) feed-forward multi-view to 3D triplane generator, and b) multi-view Score Distillation Sampling similar to [72].

2. Related Works

3D Generative Models. 3D generative modeling is a long-standing problem in computer vision and graphics. Earlier works directly train generative models such as Variational Auto Encoders (VAEs) [41] on ground-truth 3D shapes [1, 25, 48, 80, 95, 97]. However, due to the small scale of 3D shape datasets, these methods produce less realistic and diverse results compared to their 2D counterparts. With the rapid development of Generative Adversarial Networks (GANs) [27] and differentiable rendering, later works focus on learning 3D GANs from monocular images, showing impressive generation of multi-view images [50, 51], radiance fields [7, 8, 20, 28, 54, 99, 104], and meshes [14, 15, 26, 55, 56]. Nevertheless, GANs still suffer from poor generalizability and training stability, preventing them from scaling to unconstrained objects and scenes. Recently, Diffusion Models (DMs) [34, 77] have achieved great success in general 2D image synthesis, and are also applied to 3D [23, 24, 38, 40, 42, 52, 73, 90]. Yet, these methods train 3D DMs from scratch on specific objects such as human faces or vehicles, limiting their generalization. Closer to ours are methods that adapt large-scale pre-trained 2D DMs [63] for 3D generation, which we will detail next.

Novel View Synthesis (NVS) with 2D Diffusion Models. Instead of reconstructing the entire 3D shape, NVS aims to generate 3D consistent images conditioned on a few input views [66, 94]. Early methods leverage the knowledge of epipolar geometry to perform interpolation between different input views [13, 17, 68, 107]. Since NVS is a 2D image-to-image translation task, recent works have re-purposed 2D DMs for it [9, 29, 79, 85, 93, 100]. To achieve 3D consistency, SparseFusion [105] builds a view-conditioned DM on the latent space of Stable Diffusion [63], and utilizes Epipolar Feature Transformer (EFT) [78] to fuse features from input views. Zero-1-to-3 [45] directly fine-tunes Stable Diffusion on multi-view images rendered from Objaverse [18].

The concurrent work MVDream [72] proposes to denoise four views jointly with multi-view self-attention layers. However, camera pose information is fed in as 1D features to these models, discarding the explicit constraint of 3D geometry. Thus this method does not allow accurate camera control. To cope with this issue MVDream [72] generates views at fixed camera positions spread apart 90 degrees from each other. However, this approach limits the maximum number of views that can be generated to only 4. Moreover, it limits the training data to only synthetic 3D model datasets, such as Objaverse [18], since it requires rendering with the same fixed camera view for each object.

Other works thus study more explicit pose conditioning. MVDiffusion [83] derives inter-view dense correspondence from homography transformation, which is used to guide the attention module in Stable Diffusion. iNVS [39] applies image wrapping based on depth maps to re-use pixels from the source view, and thus only needs to inpaint occluded regions in novel view images. While it can produce precise reconstructions when good depth maps are available, the quality of this method degrades drastically when depth maps are noisy or inaccurate. In addition, the depth ambiguity caused by epipolar lines used in iNVS results in the flipped prediction issue, where the model cannot differentiate two views from opposite directions. SyncDreamer [46] instead builds a 3D feature volume by up-projecting features from each view, and then re-projects it to ensure 3D consistency among views. However up-projection operation requires the network to explicitly understand the depth of each pixel, sharing the same issue with iNVS [39].

Different from prior works, we exploit the internal properties of large-scale pre-trained text-to-image diffusion models and enrich self-attention maps with the cross-view interactions derived from epipolar geometry. In addition, we use Plücker coordinates [37] as positional encodings to inject 3D priors of the scene into the diffusion model, further improving camera conditioning and disambiguating different sides of the object.

Lifting 2D Diffusion Models to 3D Generation. Instead of training a model on 3D data, several works adopt a per-instance optimization paradigm where pre-trained 2D DMs

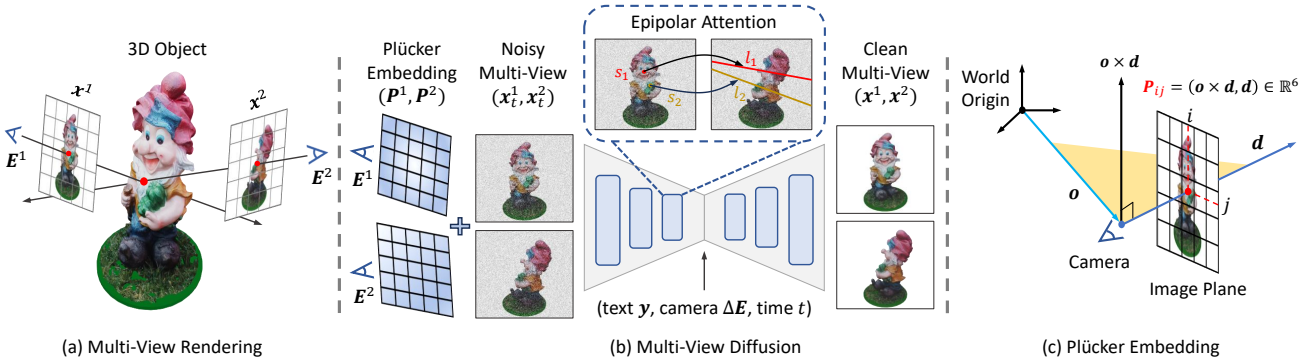


Figure 2. **Model pipeline.** (a) We initialize our multi-view diffusion model from pre-trained text-to-image model, and fine-tune it on multi-view renders of 3D objects. (b) Our model performs joint denoising on noisy multi-view images $\{\boldsymbol{x}_t^i\}_{i=1}^N$ conditioned on text $\boldsymbol{y}$ and relative camera poses $\Delta\boldsymbol{E}$. Here, we illustrate the pipeline using $N = 2$, which can be easily extended to more views. To enable cross-view interaction, we apply 3D self-attention by concatenating all views, and enforce epipolar constraints on the attention map. We further add (c) Plücker Embedding $\{\boldsymbol{P}^i\}_{i=1}^N$ to the attention layers as positional encodings, to enable precise camera control and prevent object flipping artefacts (as shown in Fig. 5).

provide image priors [58, 89]. Some of them apply it for single-view 3D reconstruction [19, 47, 59, 61, 71, 81, 98]. More relevant to ours are text-to-3D methods that optimize a NeRF model [49] by distilling the pre-trained text-to-image DM. Follow-up works have improved this text-to-3D distillation process in many directions, including more efficient 3D representations [12, 44, 70, 84, 96], better diffusion process [16, 36, 69], new loss functions [92, 106], and prompt design [2]. However, these methods still suffer from low visual quality and view consistency issues such as multi-face Janus and content drifting. SPAD generates multi-view images from a text prompt or a single input view with better 3D consistency and visual quality, which can mitigate these issues with multi-view distillation [72].

3. Method

Task Formulation. Our goal is to generate consistent and many novel views of the same object given a text prompt or an image, along with relative camera poses as input. Towards this goal, we train a multi-view diffusion model that is made spatially aware by explicitly encoding the 3D knowledge of the scene.

We build upon a state-of-the-art 2D text-to-image diffusion model (Sec. 3.1). Our specific adaptations enable 3D-aware interactions between views (Sec. 3.2), which include 3D self-attention (Sec. 3.2.1), Epipolar Attention (Sec. 3.2.2), and Plücker Embeddings (Sec. 3.2.3).

3.1. Preliminary: Text-to-Image Diffusion Models

Diffusion models (DMs) [34, 77] are generative models that learn a target data distribution $p_\theta(\boldsymbol{x}_0)$ by gradually denoising a standard Gaussian distribution, denoted as $p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T})\, d\boldsymbol{x}_{1:T}$, where $\boldsymbol{x}_{1:T}$ are intermediate noisy samples. DMs leverage a forward process that iteratively adds Gaussian noise $\boldsymbol{\epsilon}$ to the clean data $\boldsymbol{x}_0$, which is controlled by a pre-defined variance schedule $\{\bar{\alpha}_t\}_{t=1}^T$. During training, we manually construct noisy samples $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, and train a denoiser model $\epsilon_\theta(\boldsymbol{x}_t, t)$ to predict the added noise conditioned on the denoising time step t:

$$\mathcal{L}_{\text{DM}} = ||\boldsymbol{\epsilon}_t - \epsilon_\theta(\boldsymbol{x}_t, t)||^2, \text{ where } \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \quad (1)$$

Generally, the denoiser ϵ_θ is parameterized as a U-Net [64], which comprises of interleaved residual blocks [31] and self-attention layers [87]. Within this U-Net, we are interested primarily in self-attention layers [87], and we refer the reader to the original paper [63] for the overview of other blocks. The self-attention layer takes as input a feature map $\boldsymbol{F}$ and compute attention of feature in location s with entire feature map:

$$\tilde{\boldsymbol{F}}_s = \text{SoftMax}\left(\frac{Q(\boldsymbol{F}_s)\cdot K(\boldsymbol{F})^\top}{\sqrt{d}}\right)\cdot V(\boldsymbol{F}), \quad (2)$$

where Q, K, V are linear projection layers, $\boldsymbol{F} \in \mathbb{R}^{(hw)\times d}$ is a flattened feature map obtained from the 2D denoiser ϵ_θ, where d is the feature dimension, and h, w are intermediate spatial dimensions. $\boldsymbol{F}_s$, $\tilde{\boldsymbol{F}}_s$ is the input and output feature for location s respectively. In practice, the self-attention operation occurs at multiple resolutions in ϵ_θ.

3.2. Multi-View Diffusion Models

Inspired by the success of text-to-image DMs, we propose to generate multi-view images by fine-tuning a pre-trained 2D DM on multi-view rendered images of 3D assets. Fig. 2 shows the overall pipeline of our framework, SPAD. In this section, we use $N = 2$ views to explain our method for brevity. However, note that SPAD is easily extensible to generate an arbitrary number of views.

3.2.1 Multi-View Self-attention

The goal of our multi-view DM $\epsilon_\theta(\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, t, \boldsymbol{y}, \Delta\boldsymbol{E})$ is to generate 3D consistent images $(\boldsymbol{x}^1, \boldsymbol{x}^2) \in \mathbb{R}^{H\times W\times 3}$ of an object guided by a text input $\boldsymbol{y}$ and their relative camera

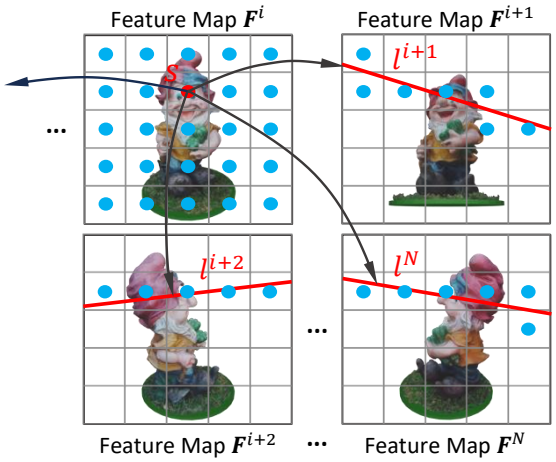


Figure 3. **Epipolar Attention.** For each point s (red point) on a feature map $\boldsymbol{F}^i$, we compute its epipolar lines $\{l^j\}_{j\neq i}$ on all other views $\{\boldsymbol{F}^j\}_{j\neq i}$. Point s will only attend to features along these lines plus all the points on itself (blue points).

pose $\Delta\boldsymbol{E} \in \mathbb{R}^{3\times4}$. To enable cross-view interaction, we concatenate the feature maps of two views side-by-side as input to the self-attention layers, denoted as $[\boldsymbol{F}^1|\boldsymbol{F}^2]$. This allows each location s on $\boldsymbol{F}^1$ to attend to all locations on itself and $\boldsymbol{F}^2$, calculated as:

$$\tilde{\boldsymbol{F}}_s^1 = \text{SoftMax}\left(\frac{Q(\boldsymbol{F}_s^1)\cdot K([\boldsymbol{F}^1|\boldsymbol{F}^2])^\top}{\sqrt{d}}\right)\cdot V([\boldsymbol{F}^1|\boldsymbol{F}^2]). \tag{3}$$

Camera conditioning. We embed the relative camera pose $\Delta\boldsymbol{E}$ with an MLP and fuse it with timestep embedding of DM to condition the residual blocks as shown in Fig. 4, similar to [72].

Issues with vanilla self-attention. We find empirically that such an unconstrained multi-view self-attention leads to content-copying between views (shown in Figure 5), i.e., the model generates similar images when the camera pose difference $\Delta\boldsymbol{E}$ is small, ignoring the underlying 3D geometry. We hypothesize that this could be the reason concurrent works such as MVDream [72] opt to generate images with 90 degree view change (along azimuths and fixed elevation) – as it diminishes the overlap between different views.

3.2.2 Multi-View Epipolar Attention

To enable SPAD to synthesize views at arbitrary relative poses, and address the above content-copying challenge, we propose to replace the vanilla self-attention operation with Epipolar Attention [78]. Epipolar Attention works by restricting the positions a point in feature map can attend to in other views – by exploiting epipolar geometry. Fig. 3 presents this mechanism. Specifically, given a source point s on a feature map $\boldsymbol{F}^i$, we compute its epipolar lines (implemented as a set of points) $\{l^j\}_{j\neq i}$ on all the other views $\{\boldsymbol{F}^j\}_{j\neq i}$. When computing the attention map between views, we ignore points that do not lie on these epipolar lines, so that the source point s only has access to features that lie along the camera ray (in other views) as well as all points in its own view for denoising:

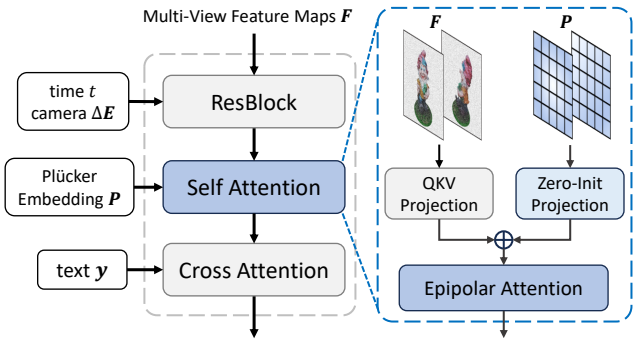


Figure 4. **Illustration of one block in our multi-view diffusion model**, which consists of a residual block, a self-attention layer, and a cross-attention layer. The residual block guides the model on the denoising timestep t and the relative camera pose $\Delta\boldsymbol{E}$, while the cross-attention layer conditions on text $\boldsymbol{y}$. We add Plücker Embedding $\boldsymbol{P}$ to feature maps $\boldsymbol{F}$ in the self-attention layer by inflating the original QKV projection layers with zero projections.

$$\tilde{\boldsymbol{F}}_s^i = \text{SoftMax}\left(\frac{Q(\boldsymbol{F}_s^i)\cdot K([\boldsymbol{F}^i|\boldsymbol{F}_{lj}^j])^\top}{\sqrt{d}}\right)\cdot V([\boldsymbol{F}^i|\boldsymbol{F}_{lj}^j]). \tag{4}$$

In practice, we dilate the epipolar lines with a 3×3 filter to consider neighboring target points for better robustness. Overall, Epipolar Attention enhances our generalizability to unseen viewpoint differences and objects.

Issues with Epipolar Attention.

However, solely using epipolar lines to constrain attention masks models can cause flipped predictions especially under large viewpoint changes. This happens because in the absence of precise depth of the object surface, the model can leverage information from any point along these lines. Consider the feature map $\boldsymbol{F}^N$ in Fig. 3, the source point s may attend to the front face of the figure or the back of it. The latter will cause a flipped prediction. iNVS [39] solves this problem with a monocular depth estimator. Yet, their imprecise depth leads to deformed object surfaces and distorted textures. We instead address this issue with a Plücker Ray Embedding which is detailed next.

3.2.3 Plücker Ray Embedding

Given a camera with its center placed at $\boldsymbol{o} \in \mathbb{R}^3$ in the world coordinate system, we represent a ray passing through a point on the feature map $\boldsymbol{F}_{ij}$ along a normalized direction $\boldsymbol{d} \in \mathbb{R}^3$ as $\boldsymbol{r}_{ij}$. We embed this ray as the positional encoding to help our model distinguish between different views. We find the simple ray parametrization $\boldsymbol{r}_{ij} = (\boldsymbol{o}, \boldsymbol{d})$ to be insufficient here. As an example, consider two rays in the same direction but with different camera origins which lie along this ray, $\boldsymbol{r}_{ij}^1 = (\boldsymbol{o}, \boldsymbol{d})$ and $\boldsymbol{r}_{ij}^2 = (\boldsymbol{o} + t\boldsymbol{d}, \boldsymbol{d})$, their embeddings turn out to be notably different, despite them representing essentially the same ray.

Inspired by recent works in Neural Light Fields [11, 74], we adopt the Plücker Ray Embedding $\boldsymbol{P}_{ij} = (\boldsymbol{o} \times \boldsymbol{d}, \boldsymbol{d})$ where $\times$ is the cross product. See Fig. 2 (c) for illustration. This parametrization is able to map $\boldsymbol{r}^1$ and $\boldsymbol{r}^2$ to the same embedding as we have:

$$(\boldsymbol{o} + t\boldsymbol{d}) \times \boldsymbol{d} = \boldsymbol{o} \times \boldsymbol{d} + t\boldsymbol{d} \times \boldsymbol{d} = \boldsymbol{o} \times \boldsymbol{d}. \quad (5)$$

We simply pass the Plücker Embedding $\boldsymbol{P} \in \mathbb{R}^{(hw)\times 6}$ through a linear projection layer to project it into d dimension and add it to the multi-view feature maps $\boldsymbol{F}$, which serves as the input to the Epipolar Attention layer, shown in Fig. 4. To avoid disturbing the pre-trained model, the weights of the projection layer are initialized as zeros and learned during fine-tuning similar to ControlNet [102].

3D geometric priors in Plücker Embedding. With the Plücker coordinates, rays that are close in the 3D space share similar embeddings, which leads to higher values in the pre-softmax self-attention map $Q(\boldsymbol{F}) \cdot K(\boldsymbol{F})^\top$. This encourages feature points to look at spatially nearby locations in other views, enhancing the 3D consistency across views. On the other hand, two rays passing through the same 3D location from opposite cameras will have Plücker coordinates with flipped (positive/negative) sign. Their embeddings will have a smaller dot product and result in a smaller attention value. The two pixels will thus attend less to each other, effectively addressing the flipped prediction problem.

4. Experiments

We conduct extensive experiments to answer the following questions: **(i)** Can SPAD generate high-quality multi-view images from diverse (non-orthogonal and overlapping) viewpoints that are aligned with input text or image? (Sec. 4.2) **(ii)** Are the synthesized views 3D consistent? (Sec. 4.3) **(iii)** To what extent do Plücker Positional Embeddings and Epipolar Attention contribute to the overall performance? (Sec. 4.3) **(iv)** Lastly, can SPAD enable high-quality text-guided 3D asset generation? (Sec. 4.5)

4.1. Experimental Setup

Training Data Curation. Instead of using the entire Objaverse [18] which consists of many flat and primitive shapes that can drift the diffusion model away from high-quality generation, we filter Objaverse using a few simple but useful heuristics based on its metadata. We select 150K objects with the most like, view, and comment count available in metadata, as well as the top 50K objects that contain the highest number of mesh polygons and vertex counts. We use Blender [4] to render 12 multi-view images for each object at a resolution of 256×256. All objects are centered and re-scaled to a unit cube. We randomly sample camera positions with elevations in $[-90°, 90°]$, azimuths in $[0°, 360°]$, and fix the distance to origin as 3.5 and FOV as 40.26°.

Training Details. We initialize SPAD from the pre-trained weights of Stable Diffusion v1.5 [63]. We train two versions of our model: one with text conditioning and another one with image conditioning for the novel view synthesis task. In both the variants, we set the number of views N to 2 while training. For the text-conditioned model, we jointly denoise both the views. For the image-conditioned model, we feed in one clean source view image and denoise the target view. All our baselines and reported numbers follow this setup. First, we train our models two-view models for 40K iterations on Objaverse, with an effective batch size of 1728 samples per iteration, on eight H100 GPUs. Later we train a larger text-conditioned model with $N = 4$ views on sixteen H100 GPUs for 100K steps, and use it to generate all visuals (except ablation study).

Evaluation Datasets and Metrics. For text-conditioned multi-view image generation, we follow MVDream [72] and randomly sample 1,000 Objaverse captions as text prompts to generate images. We use CLIP-score [60] to measure the image-text alignment. We also report the Inception Score (IS) [67] and Fréchet Inception Distance (FID) [32] of the generated images to evaluate image generation quality. It is important to highlight that these metrics only measure generation quality of individual images, and do not provide any information about their multi-view 3D consistency.

For image-conditioned novel view synthesis, we select 1,000 unseen Objaverse objects which are not contained in our training set for testing. Following [39, 45], we also adopt real-world scanned objects from the *Google Scanned Objects* (GSO) dataset [22] to evaluate the generality of our method. We render each object from two views following the same setup in training data generation, where one view serves as the model input and another view is the target novel view image. We report PSNR, SSIM [91], and LPIPS [103] metrics to measure the accuracy of the synthesized novel views.

Baselines. Since it is difficult to replicate and control for training and rendering setups used in prior works, we choose the following variants of our model as primary baselines: *Vanilla MV-DM* that only adds 3D self-attention on concatenated multi-view feature maps without Epipolar Attention and Plücker Embedding; *MV-DM (Epipolar)* and *MV-DM (Plücker)* which incorporate the two components, respectively. We also compare SPAD with two *concurrent* works: *MVDream* [72] and *SyncDreamer* [46]. Different from SPAD, both methods can only generate views at a fixed elevation and azimuth ranges. In the image-conditioned novel view synthesis task, we compare with additional baselines *Zero-1-to-3* [45] and *iNVS* [39]. We used the official codebase and pre-trained weights of these methods on our testing data to report their results.

Method	IS ↑	CLIP-score ↑
MVDream (v2.1) † [72]	13.36±0.87	30.22±3.83
MVDream (v1.5) † [72]	9.72±0.43	28.55±4.05
SyncDreamer ‡ [46]	11.69±0.24	27.76±4.84
Vanilla MV-DM	11.04±0.81	28.52±3.69
SPAD (**Ours**)	11.18±0.97	29.87±3.33

Table 1. **Quantitative results on text-conditioned multi-view image generation.** We randomly sample 1,000 captions from Objaverse, and evaluate the FID, Inception Score (IS), and CLIP-score. † We ran MVDream's code on the same captions we used. ‡ We first generated single-view images using Stable Diffusion [63] on the same captions we used and removed their backgrounds. Then, we ran SyncDreamer's code to generate multi-view images.

4.2. Text-conditioned Multi-View Generation

We use single view quality metrics to compare methods, similar to MVDream. We evaluate two MVDream variants, which are fine-tuned from Stable Diffusion v1.5 (same as ours) and v2.1, respectively. For SyncDreamer, we follow the text-to-image-to-3D pipeline described in their paper to first generate a single-view image from a text prompt using Stable Diffusion, and then generate multiple views from it. We make sure that the single-view image is aligned with the text, and pre-process it using the script provided in their official codebase.

SPAD is a strong 2D text-to-image generator. The results on image generation quality are presented in Tab. 1. SPAD outperforms or matches both baselines on 2D Image Quality metrics when compared against the methods utilizing the same underlying Stable Diffusion v1.5 base model. This confirms that our method while being more 3D consistent, does not compromise either text-to-image alignment or overall image quality, but rather improves it compared to our baseline MV-DM.

We provide qualitative results in Fig. 1, Fig. 11, and Fig. 12. SPAD is able to generate consistent multi-view images of diverse 3D subjects, ranging from daily objects to highly complex machines. Additionally, we put preliminary investigations of training SPAD with v2.1 base model in Appendix B.3.

4.3. Image-conditioned Novel View Synthesis

Since image quality metrics do not provide any indication of multi-view consistency or the quality of camera control. For evaluation of multi-view consistency, we mostly rely on image-conditioned experiments. For this evaluation, given an input view and relative camera pose, we generate the target view and compare it against ground truth. Tab. 2 and Tab. 3 present the novel view synthesis results on Objaverse and GSO, respectively.

SPAD preserves structural and perceptual details faithfully. We find that SPAD outperforms all baselines on LPIPS metrics across both datasets, while matching its performance to SyncDreamer on SSIM. Moreover, we find that adding each component (Epipolar and Plücker) gradually improves scores across the board, and leads to state-of-the-art performance with our full model. This confirms our main hypothesis that imparting 3D understading to MV-DMs is helpful.

We also find that iNVS [39] is able to achieve the highest PSNR, since it directly copies pixels from the source view (via depth-based reprojection). However, it particularly performs worse on SSIM and LPIPS metrics, which measure the structural and semantic accuracy of the generated view. This is because of deformations introduced by reprojection when viewpoint changes are large and monocular depth from ZoeDepth [3] is inaccurate.

The official inference code of SyncDreamer always generates 16 views at fixed azimuth angles uniformly distributed in $[0°, 360°]$, which is incompatible with our random view generation setup. We modified their code to consider the exact target camera pose as model input, but found it performed worse than choosing the prediction at the azimuth that is closest to the target azimuth. Therefore, we report SyncDreamer results using the closest view, where the error is usually smaller than $10°$.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3 † [45]	18.16	0.81	0.201
iNVS [39]	**20.52**	0.81	0.178
SyncDreamer † [46]	19.51	**0.84**	0.174
Vanilla MV-DM	17.56	0.81	0.20
MV-DM (Epipolar)	18.90	0.82	0.19
MV-DM (Plücker)	17.98	0.81	0.20
SPAD (**Ours**)	20.29	**0.84**	**0.166**

Table 2. **Quantitative results on image-conditioned novel view synthesis on Objaverse.** We report PSNR, SSIM, and LPIPS on the generated novel view images of 1,000 unseen Objaverse objects. † We run the official codebases of SyncDreamer and Zero-1-to-3 to report results.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Zero-1-to-3 [45]	16.10	0.82	0.183
iNVS [39]	**18.53**	0.80	0.180
SyncDreamer † [46]	17.18	**0.83**	0.178
Vanilla MV-DM	15.98	0.81	0.20
MV-DM (Epipolar)	17.13	0.82	0.19
MV-DM (Plücker)	16.15	0.81	0.20
SPAD (**Ours**)	17.99	**0.83**	**0.169**

Table 3. **Quantitative results on image-conditioned novel view synthesis on GSO.** We report PSNR, SSIM, and LPIPS on the generated novel view images of GSO objects. † SyncDreamer only reports results on 30 selected objects from GSO in their paper [46]. We ran their code and test it on all the GSO objects here.

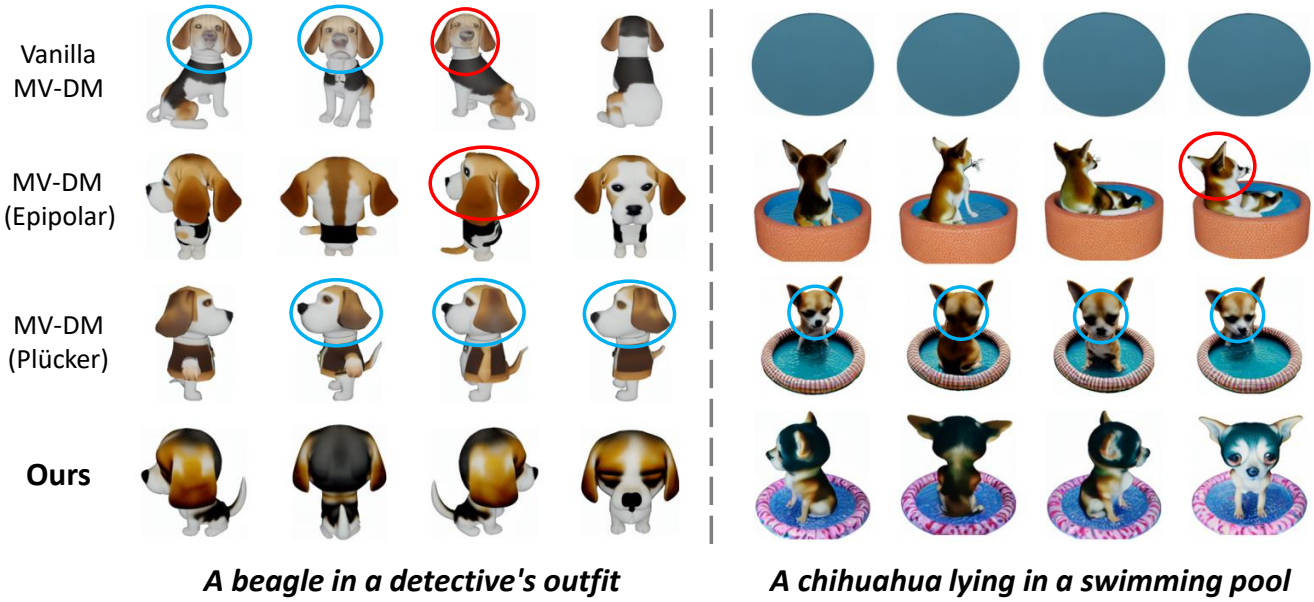


Figure 5. **Qualitative comparison between SPAD and its variants.** We prompt models trained on two views to generate four views at 90 degree intervals for clear visual distinctions. The flipped predicted views are highlighted with red circles, while the content-copying issues are indicated by blue circles.

4.4. Qualitative Analysis

We also conduct qualitative analysis to visually understand the usefulness of each component of our model. The text-conditioned multi-view generation results of baselines and SPAD are shown in Fig. 5, where all models are trained with two views while prompted to generate four views. The elevation is fixed for all the views, and the azimuth spans uniformly between $[0°, 360°]$.

Epipolar Attention promotes better camera control in SPAD. We find that the vanilla (full) 3D self-attention used in Vanilla MV-DM and MV-DM (Plücker) models often leads to content copying. This is highlighted in the figure using blue circles, where the generated dogs face in similar direction, ignoring the target camera poses. We hypothesize that the readaptation of the self-attention layer of SD originally trained to attend only to itself hinders with generalization and controllability of this model.

Additionally, since these models are trained only to generate two views, we hypothesize that they overfit to predicting only two novel views. In contrast, Epipolar Attention constrains cross-view interactions to only happen between spatially related pixels, reducing the search space in establishing correspondences across images. Despite not being trained on four views, the model is still able to generate 3D consistent images by attending to the correct regions.

Plücker Embeddings help prevent generation of flipped views. When the difference in camera positions between two views is large, the epipolar lines introduce ambiguities in the ray directions. Indeed, Vanilla MV-DM and MV-DM (Epipolar) sometimes predict image regions that are rotated by $180°$. For example, the dog's head highlighted by red circles looks in the opposite direction of the body, which is inconsistent with other views. Instead, Plücker Embeddings bias the model to pay less attention to camera views on opposite sides of the object, while leveraging more information from spatially closer views.

4.5. Text-to-3D Generation

Multi-view SDS. Inspired by [45, 72], we also adopt the multi-view Score Distillation Sampling (SDS) [58] to perform text-to-3D generation using the four-view SPAD variant. Concretely, we integrate our model into the state-of-the-art text-to-3D generation codebase threestudio [30], and follow the setup similar to MVDream [72] for stable NeRF [49] distillation. Fig. 6 shows the multi-view rendered images of the trained NeRF models. We find that SPAD is able to reconstruct consistent geometry without Janus problem, while maintaining good visual quality.

Multi-view Triplane Generator. Inspired by concurrent works [35, 43], we trained a multi-view to triplane generator that takes as input generated views from SPAD and generates a NeRF in a single feed-forward pass, which is nearly two orders of magnitude faster than SDS optimization. Fig. 7 shows the results from this experiment. We find that SPAD can be used as a faithful base model to facilitate such generations.

5. Conclusion

In this paper, we propose SPAD, a novel framework for generating multiple views from text or image input. We propose to transform the self-attention layers of the pre-trained text-to-image diffusion model into Epipolar Atten-

Figure 6. **Text-to-3D generation using multi-view SDS with SPAD.** We adopt the multi-view SDS proposed in MVDream [72] to train a NeRF model. Thanks to the 3D consistency of our model, we do not suffer from the multi-face Janus issue.

tion to promote multi-view interactions and improve camera control. Moreover, we augment self-attention layers with Plücker positional encodings to further improve camera control by preventing flipping view prediction of the object. We provide rigorous evaluations of these modifications and demonstrate state-of-the-art results in terms of image-conditioned novel view synthesis.

Limitations and Future Work. While our method improves the 3D consistency of multi-view diffusion models, there still remains lots of scope for improvements. For example, a larger Stable Diffusion such as SDXL [57] can further improve performance while preventing lossy compres-

Figure 7. **Text-to-3D generation using multi-view Triplane generator with SPAD.** Following [35, 43] we trained a multi-view conditioned triplane generator that outputs a NeRF using four outputs of SPAD in a single feed-forward pass. We show the rendered NeRF on the top row (zoomed) and corresponding multi-view outputs from SPAD in the bottom row. For entire 360-degree videos see our website.

sion of image conditioning. We can use a monocular depth estimator similar to iNVS [39] to further improve the correspondences established by epipolar self-attention. Finally, we plan to explore the usage of SPAD to generate dynamic 4D assets and multi-object scenes. Additionally, datasets of monocular videos and pre-trained text-to-video generators can be explored to improve the quality and consistency of the generated results.

Acknowledgments

We would like to thank Xuanchi Ren and Weize Chen, for valuable discussions and support.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *Proceedings of the International Conference on Machine Learning*, 2018. 2, 3

[2] 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. 4

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 7

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2022. 6

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 15

[6] Junli Cao, Huan Wang, Pavlo Chemerys, Vladislav Shakhrai, Ju Hu, Yun Fu, Denys Makoviichuk, Sergey Tulyakov, and Jian Ren. Real-time neural light field on mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[8] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[9] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion

models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[11] Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. *arXiv preprint arXiv:2304.13681*, 2023. 6

[12] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 4

[13] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Special Interest Group on Computer Graphics and Interactive Techniques*, 1993. 3

[14] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in Neural Information Processing Systems*, 2019. 2, 3

[15] Wenzheng Chen, Joey Litalien, Jun Gao, Zian Wang, Clement Fuji Tsang, Sameh Khamis, Or Litany, and Sanja Fidler. Dib-r++: learning to predict lighting and material with a hybrid differentiable renderer. *Advances in Neural Information Processing Systems*, 2021. 3

[16] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 4

[17] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Special Interest Group on Computer Graphics and Interactive Techniques*, 1996. 3

[18] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3, 6

[19] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[20] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021. 2

[22] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2022. 3, 6

[23] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[24] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[25] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision*, 2018. 3

[26] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in Neural Information Processing Systems*, 2022. 2, 3

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014. 2, 3

[28] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[29] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *Proceedings of the International Conference on Machine Learning*, 2023. 3

[30] 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. 8

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4

[32] 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*, 2017. 6

[33] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2022. 15

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2, 3, 4

[35] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d, 2023. 8, 10

[36] 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. 2, 4

[37] Yan-Bin Jia. Plücker coordinates for lines in the space. *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 2020. 3

[38] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3

[39] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs: Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 2, 3, 5, 6, 7, 10, 15

[40] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[42] Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. Salad: Part-level latent diffusion for 3d shape generation and manipulation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[43] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *https://arxiv.org/abs/2311.06214*, 2023. 8, 10

[44] 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 Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[45] 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. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2, 3, 6, 7, 8, 15, 18

[46] 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, 3, 6, 7, 15

[47] 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 Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 4, 8

[50] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 3

[51] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in Neural Information Processing Systems*, 2020. 3

[52] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[53] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, 2021. 2

[54] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[55] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. *Advances in Neural Information Processing Systems*, 2020. 3

[56] Dario Pavllo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[57] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 9

[58] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proceedings of the International Conference on Learning Representations*, 2023. 2, 4, 8

[59] 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, 4

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 6

[61] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 4

[62] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 6, 7

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015. 4

[65] 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*, 2022. 2

[66] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[67] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 2016. 6

[68] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 3

[69] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023. 4

[70] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 4

[71] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 4

[72] 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. 2, 3, 4, 5, 6, 7, 8, 9, 15, 17

[73] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[74] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 2021. 6

[75] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *Advances in Neural Information Processing Systems*, 2022. 2

[76] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *Proceedings of the International Conference on Learning Representations*, 2023. 2

[77] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2015. 3, 4

[78] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *ECCV*, 2022. 2, 3, 5

[79] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[80] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[81] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. 2023. 4

[82] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[83] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *Advances in Neural Information Processing Systems*, 2023. 3

[84] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *Proceedings of the International Conference on 3D Vision*, 2023. 4

[85] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[86] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[88] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[89] 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 Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[90] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[91] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 6

[92] 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. *Advances in Neural Information Processing Systems*, 2023. 2, 4

[93] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[94] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[95] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 2016. 2, 3

[96] Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, and Errui Ding. Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation. *arXiv preprint arXiv:2307.16183*, 2023. 4

[97] Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3

[98] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 4

[99] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[100] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[101] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. 2023. 2

[102] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 6

[103] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 6

[104] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[105] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[106] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 4

[107] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 2004. 3

A. More Implementation Details

DDIM Initialization. We use a black Gaussian-blob on a white background to initialize the first 20 steps (out of 200) of our DDIM sampler, which ensures that our model correctly generates a single object placed at the center on the white background (similar to images during training). This trick is similar to the one used in iNVS [39] which starts inpainting with the partial image warped with depth.

We list our hyper-parameter choices and miscellaneous training details below:

Hyper-parameters	Value
Base Learning rate	1e-4
Learning rate decay	None
Loss Type	L2
Classifier-free guidance	7.5 (text-only)
Effective batch size	1152
DDIM Steps	200
Gaussian Blob Initialization Steps	20
CLIP Frozen	True
Renders background color	White
Image Resolution	256
Learning rate linear warmup	100 steps

Table 4. Hyperparameter choices for SPAD.

B. Additional Experiments and Results

B.1. Qualitative Results and Baseline Comparisons

Text-conditioned multi-view generations and comparison with MVDream [72]. Fig. 8 presents the results. We find SPAD synthesizes images with higher quality details and better alignment with the text prompt.

Image-conditioned novel view synthesis and comparison with Zero123 [45]. Fig. 9 presents the results. We find SPAD preserves the structural and perceptual details of objects and exhibits better 3D consistency.

Close viewpoints generations from SPAD. In Fig. 10, we put text-conditioned multi-view generations from SPAD where we increment the azimuth angle by 10 degrees per view. We find that SPAD can synthesize continuous moving views well, without content copying issues.

B.2. User Study comparing SPAD with MVDream

We conducted a user study on the visual quality, 3D consistency, and text alignment of multi-view generations. We distributed our questions via Amazon Mechanical Turk, where participants were given 4-view generations of SPAD and MVDream [72], and asked to choose the better one satisfying the above properties. We found that SPAD is preferred over MVDream with 59% vs 41%.

B.3. Training with Stable Diffusion v2.1 Weight

The SPAD model we evaluated in the main paper and Appendix B.1 is initialized from the weight of Stable Diffusion (SD) v1.5. Here, we train another model initializing from the weight of the stronger SD v2.1 release. Fig. 12 presents the multi-view generation results of this model. Indeed, we observe better alignment with the text input, especially with longer and more complicated prompts.

This is also verified by the quantitative result. SPAD with SD v1.5 achieves a CLIP-score of 29.87±3.33. SPAD with SD v2.1 achieves a better CLIP-score of 30.39±3.30, which is also higher than MVDream [72] initialized from the same SD v2.1 weight (30.22±3.83).

B.4. Classifier-free Guidance

Classifier-free diffusion guidance [33] is a technique used to balance the quality and diversity of images produced by diffusion models. This method is particularly effective in class-conditional and text-conditional image generation, enhancing both the visual quality of images and their alignment with given conditions. Inspired by [5] we explore the integration of classifier-free guidance with Epipolar Attention and Plücker Embedding. Implementing classifier-free guidance involves simultaneous training of the diffusion model for both conditional and unconditional denoising tasks. During inference, these models' score estimates are merged. We have four different types of conditioning injected into our system:

- Text (c_T): Injected from CLIP text-encoder similar to Vanilla Stable Diffusion.
- Camera (c_C): Injected with timestep via Residual blocks.
- Epipolar Attention (c_E): Injected by applying mask during self-attention.
- Plücker Embedding (c_P): Injected by concatenation during self-attention.

During training, we extend classifier-free guidance over all these conditions. Therefore, our modified score estimate during inference is as follows:

$$\begin{aligned}\tilde{e_\theta}(z_t, c_T, c_C, c_E, c_P) &= e_\theta(z_t, \varnothing, \varnothing, \varnothing, \varnothing) \\ &+ s_T \cdot (e_\theta(z_t, c_T, \varnothing, \varnothing, \varnothing) - e_\theta(z_t, \varnothing, \varnothing, \varnothing, \varnothing)) \\ &+ s_C \cdot (e_\theta(z_t, c_T, c_C, \varnothing, \varnothing) - e_\theta(z_t, c_T, \varnothing, \varnothing, \varnothing)) \\ &+ s_E \cdot (e_\theta(z_t, c_T, c_C, c_E, \varnothing) - e_\theta(z_t, c_T, c_C, \varnothing, \varnothing)) \\ &+ s_P \cdot (e_\theta(z_t, c_T, c_C, c_E, c_P) - e_\theta(z_t, c_T, c_C, c_E, \varnothing))\end{aligned}$$

Outcome: As shown in Fig. 13, we find that classifier-free guidance beyond text conditioning does not provide additional benefits, and rather leads to over-saturated generations. This also aligns with our observations on MVDream.

B.5. Joint Multi-View Inference

Concurrent multi-view diffusion models [46, 72] are limited to generating the same number of views they were trained

on during testing. However, generating a high-quality 3D asset by e.g. training a NeRF model usually requires more than ten views of the asset. A naive solution is to use more views during training, which leads to quadratically increasing training costs due to the use of 3D self-attention. Instead, we propose a joint multi-view inference technique, which enables generating an infinite number of views using a model trained with fewer views.

Assume that we want to generate M views with a two-view model. We first initialize M noise maps $\{\boldsymbol{x}_T^i\}_{i=1}^M$, and then iteratively denoise all possible pairs of views:

$$(\boldsymbol{x}_{t-1}^i, \boldsymbol{x}_{t-1}^j) = \mathrm{Denoise}(\boldsymbol{x}_t^i, \boldsymbol{x}_t^j, \epsilon_\theta),\ \forall\, i, j \in [1, M], i \neq j. \tag{6}$$

Since the model is only trained on both views with the same noise level (i.e., timestep t), we sample (i, j) pairs without replacement and make sure to go over all possible combinations uniformly via simple heuristics.

Outcome: We find that this experiment trades off 3D consistency, as it only allows cross-view communication between two views at any given timestep of generation.

B.6. Fréchet Inception Distance (FID) Results

Compared to Vanilla MV-DM with an FID score of 55.25, our full model SPAD achieves a better FID score of 52.77 which shows further evidence of improvement in 2D generation quality.

FID Comparison with MVDream. Since our model generates images at random views, it has a much larger pose distribution mismatch compared to MVDream which uses orthogonal (90-degree varying) views in both ground-truth and generated images. Due to this reason, our FID cannot be compared directly with MVDream (trained with v2.1) which is reported to be 32.06 in the original work.

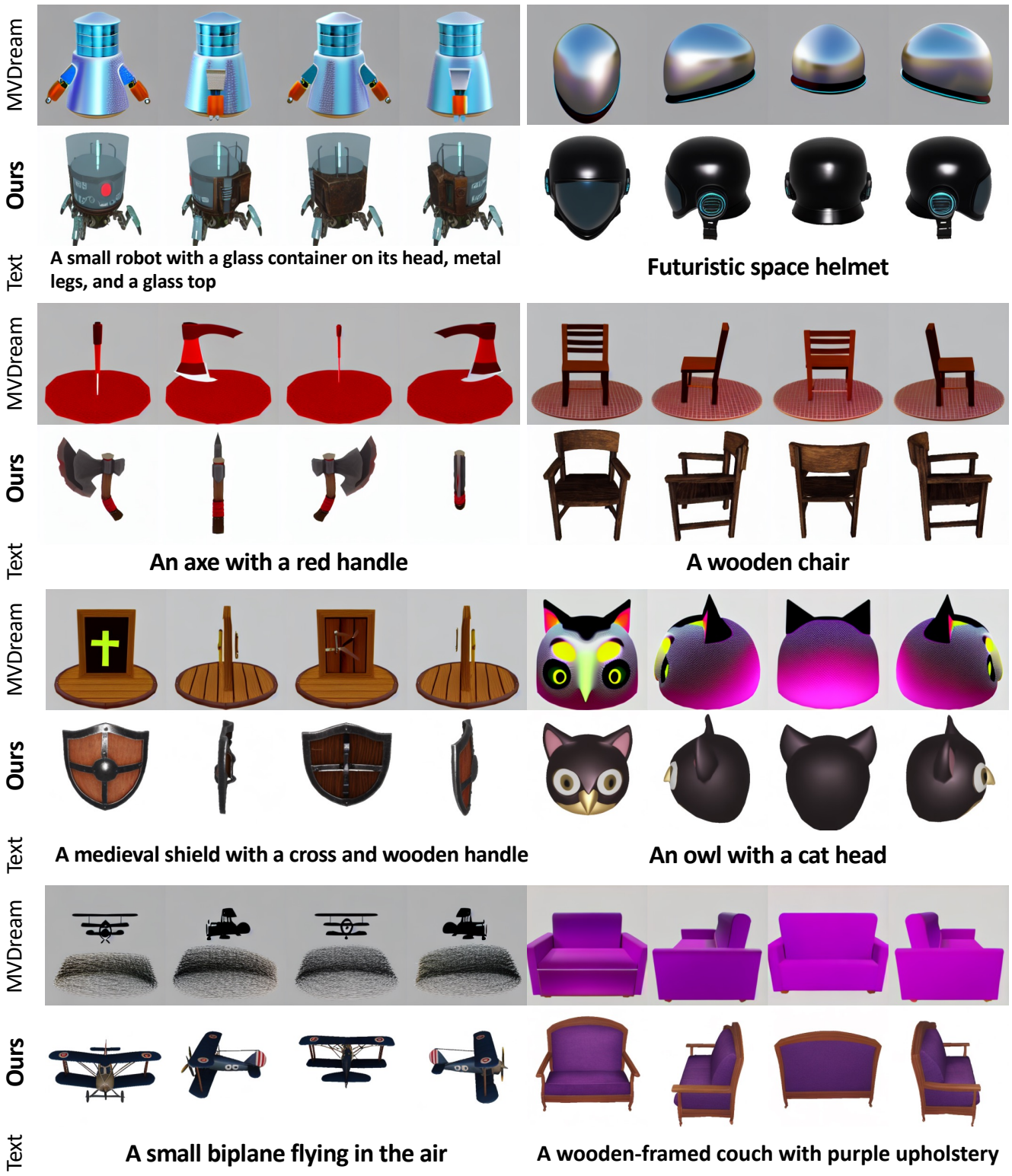


Figure 8. **Comparison of text-conditioned multi-view generation with MVDream [72].**

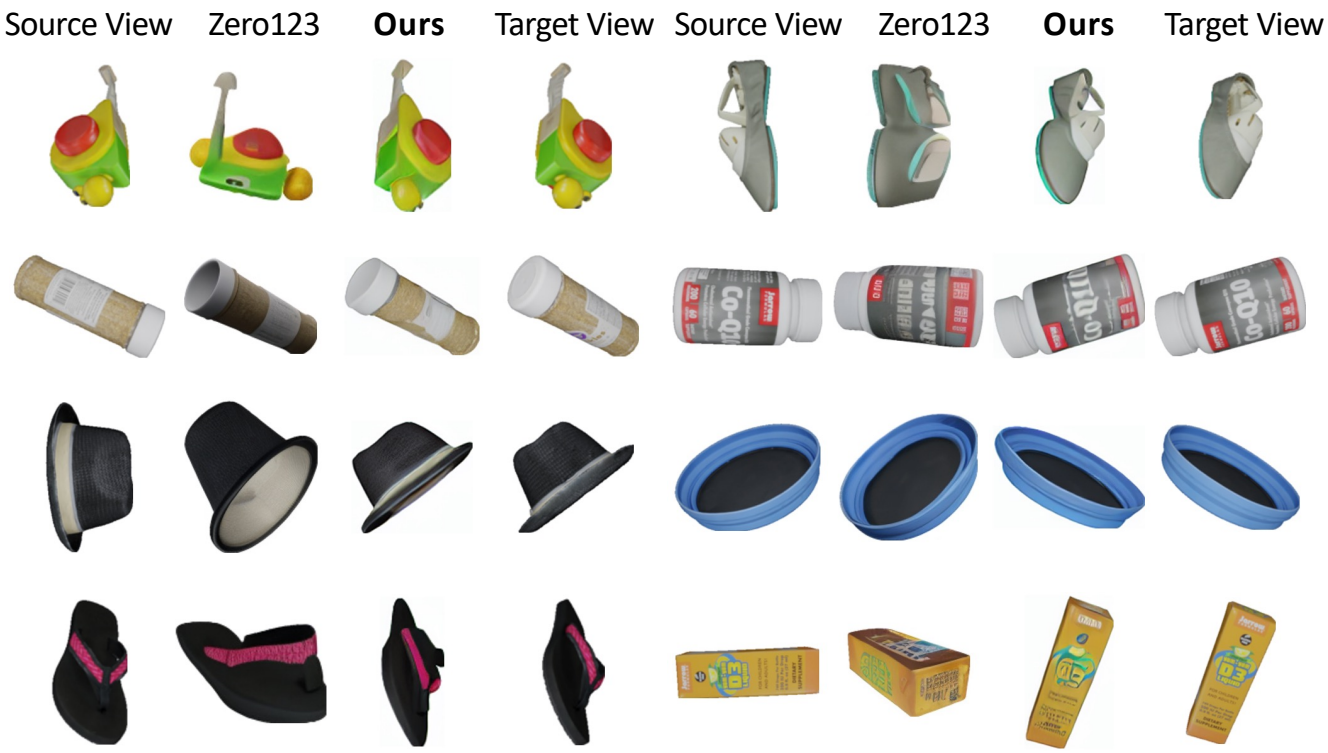


Figure 9. **Comparison of image-conditioned novel view synthesis with Zero123 [45].**

A white Ford F-150 King Ranch pickup truck

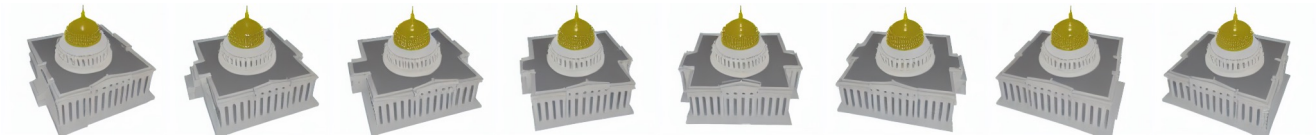
Red Fidget Spinner Model

A white marble Greek temple with columns and pillars

A blue muscle car

A dog skull

A loaf of bread or baguette



The US Capitol building with a white exterior and golden-yellow dome

A Porsche sports car collection, including 911, 991 Carrera, Cayman S, and Cayenne models

Figure 10. **Close-view generation results from SPAD**. We generate images at continuous viewpoints with an offset of 10 degrees.

Figure 11. **More multi-view generation results with SPAD.** The tested model is initialized with the weight of Stable Diffusion v1.5, and fine-tuned on Objaverse rendered images (same as Fig. 1 in the main paper).

Figure 12. **More multi-view generation results with SPAD.** The tested model is initialized with the weight of Stable Diffusion v2.1, and fine-tuned on Objaverse rendered images. Compared to results in Fig. 11 which adopts the weight of Stable Diffusion v1.5, this model is able to follow more complicated text prompts.

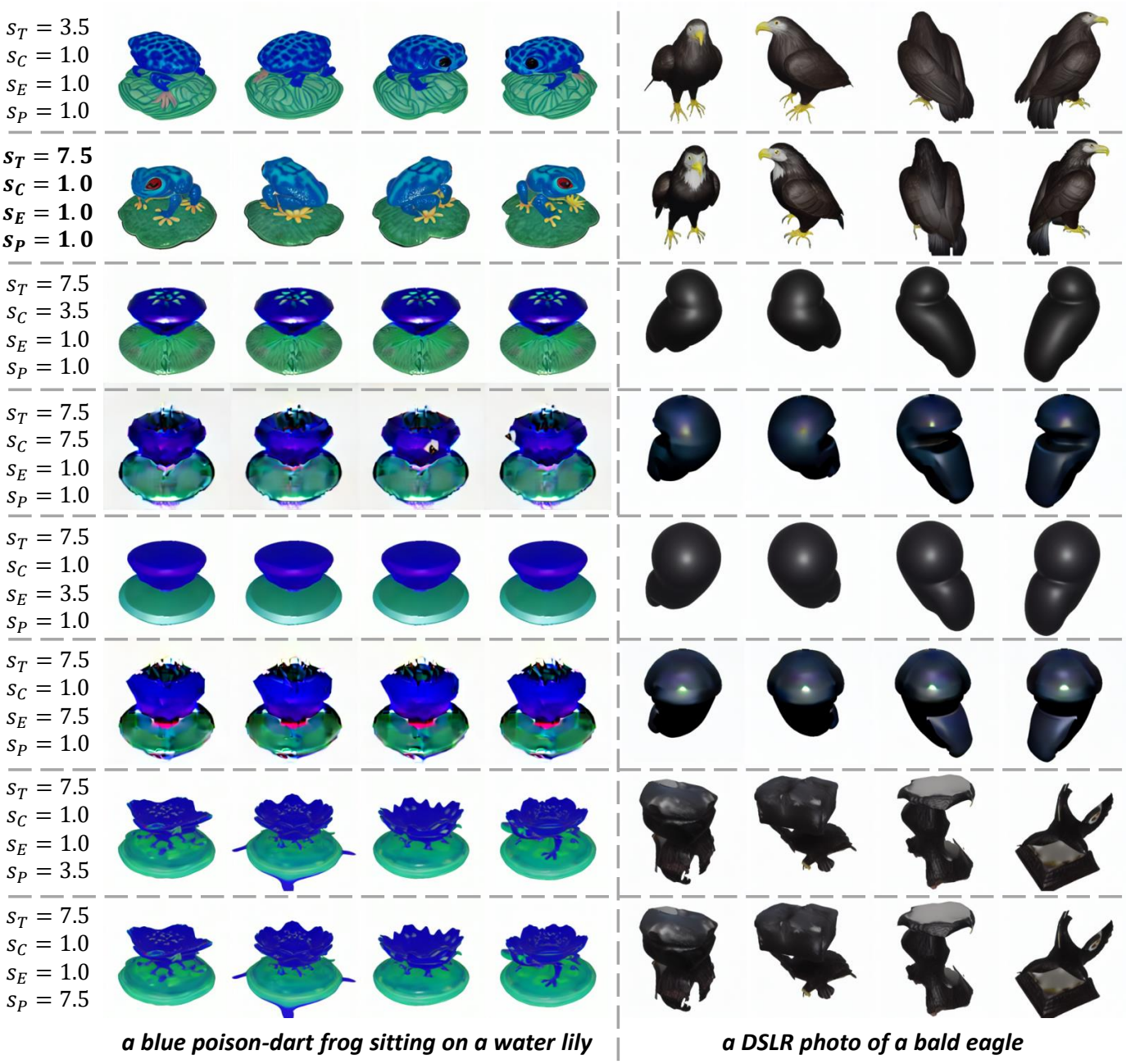


Figure 13. **Ablation study regarding the classifier-free guidance scales.** Using a large scale of $s_T = 7.5$ for text conditioning works the best (row 2), while increasing scales for camera embedding, Epipolar Attention, and Plücker Embedding all leads to over-saturated images.