
Shap-E: Generating Conditional 3D Implicit Functions

Heewoo Jun *
heewoo@openai.com

Alex Nichol *
alex@openai.com

Abstract

We present Shap-E, a conditional generative model for 3D assets. Unlike recent work on 3D generative models which produce a single output representation, Shap-E directly generates the parameters of implicit functions that can be rendered as both textured meshes and neural radiance fields. We train Shap-E in two stages: first, we train an encoder that deterministically maps 3D assets into the parameters of an implicit function; second, we train a conditional diffusion model on outputs of the encoder. When trained on a large dataset of paired 3D and text data, our resulting models are capable of generating complex and diverse 3D assets in a matter of seconds. When compared to Point-E, an explicit generative model over point clouds, Shap-E converges faster and reaches comparable or better sample quality despite modeling a higher-dimensional, multi-representation output space. We release model weights, inference code, and samples at <https://github.com/openai/shap-e>.

1 Introduction

With the recent explosion of generative image models [47, 40, 48, 17, 52, 53, 70, 15], there has been increasing interest in training similar generative models for other modalities such as audio [44, 10, 5, 31, 2, 25], video [24, 57, 23], and 3D assets [26, 28, 45, 41]. Most of these modalities lend themselves to natural, fixed-size tensor representations that can be directly generated, such as grids of pixels for images or arrays of samples for audio. However, it is less clear how to represent 3D assets in a way that is efficient to generate and easy to use in downstream applications.

Recently, implicit neural representations (INRs) have become popular for encoding 3D assets. To represent a 3D asset, INRs typically map 3D coordinates to location-specific information such as density and color. In general, INRs can be thought of as resolution independent, since they can be queried at arbitrary input points rather than encoding information in a fixed grid or sequence. Since they are end-to-end differentiable, INRs also enable various downstream applications such as style transfer [72] and differentiable shape editing [3]. In this work, we focus on two types of INRs for 3D representation:

- **A Neural Radiance Field (NeRF)** [38] is an INR which represents a 3D scene as a function mapping coordinates and viewing directions to densities and RGB colors. A NeRF can be rendered from arbitrary views by querying densities and colors along camera rays, and trained to match ground-truth renderings of a 3D scene.
- **DMTet** [56] and its extension **GET3D** [18] represent a textured 3D mesh as a function mapping coordinates to colors, signed distances, and vertex offsets. This INR can be used to construct 3D triangle meshes in a differentiable manner, and the resulting meshes can be rendered efficiently using differentiable rasterization libraries [32].

*Equal contribution



Figure 1: Selected text-conditional meshes generated by Shap-E. Each sample takes roughly 13 seconds to generate on a single NVIDIA V100 GPU, and does not require a separate text-to-image model.

Although INRs are flexible and expressive, the process of acquiring them for each sample in a dataset can be costly. Additionally, each INR may have many numerical parameters, potentially posing challenges when training downstream generative models. Some works approach these issues by using auto-encoders with an implicit decoder to obtain smaller latent representations that can be directly modeled with existing generative techniques [43, 34, 30]. Dupont et al. [12] present an alternative approach, where they use meta-learning to create a dataset of INRs that share most of their parameters, and then train diffusion models [58, 60, 22] or normalizing flows [51, 13] on the free parameters of these INRs. Chen and Wang [6] further suggest that gradient-based meta-learning might not be necessary at all, instead directly training a Transformer [64] encoder to produce NeRF parameters conditioned on multiple views of a 3D object.

We combine and scale up several of the above approaches to arrive at Shap-E, a conditional generative model for diverse and complex 3D implicit representations. First, we scale up the approach of Chen and Wang [6] by training a Transformer-based encoder to produce INR parameters for 3D assets. Next, similar to Dupont et al. [12], we train a diffusion model on outputs from the encoder. Unlike previous approaches, we produce INRs which represent both NeRFs and meshes simultaneously, allowing them to be rendered in multiple ways or imported into downstream 3D applications.

When trained on a dataset of several million 3D assets, our models are capable of producing diverse, recognizable samples conditioned on text prompts (Figure 1). Compared to Point-E [41], a recently proposed explicit 3D generative model, our models converge faster and obtain comparable or superior results while sharing the same model architecture, datasets, and conditioning mechanisms.

Surprisingly, we find that Shap-E and Point-E tend to share success and failure cases when conditioned on images, suggesting that very different choices of output representation can still lead to similar model behavior. However, we also observe some qualitative differences between the two models, especially when directly conditioning on text captions. Like Point-E, the sample quality of our models still falls short of optimization-based approaches for text-conditional 3D generation. However, it is orders of magnitude faster at inference time than these approaches, allowing for a potentially favorable trade-off.

We release our models, inference code, and samples at <https://github.com/openai/shap-e>.

2 Background

2.1 Neural Radiance Fields (NeRF)

Mildenhall et al. [38] introduce NeRF, a method for representing a 3D scene as an implicit function defined as

$$F_{\Theta} : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma)$$

where \mathbf{x} is a 3D spatial coordinate, \mathbf{d} is a 3D viewing direction, \mathbf{c} is an RGB color, and σ is a non-negative density value. For convenience, we split F_{Θ} into separate functions, $\sigma(\mathbf{x})$ and $\mathbf{c}(\mathbf{x}, \mathbf{d})$.

To render a novel view of a scene, we treat the viewport as a grid of rays and render each ray by querying F_{Θ} at points along the ray. More precisely, each pixel of the viewport is assigned a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ which extends from the camera origin \mathbf{o} along a direction \mathbf{d} . The ray can then be rendered to an RGB color by approximating the integral

$$\hat{\mathbf{C}}(\mathbf{r}) = \int_0^{\infty} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt, \text{ where } T(t) = \exp\left(-\int_0^t \sigma(\mathbf{r}(s))ds\right)$$

Mildenhall et al. [38] use quadrature to approximate this integral. In particular, they define a sequence of increasing values $t_i, i \in [1, N]$ and corresponding $\delta_i = t_{i+1} - t_i$. The integral is then approximated via a discrete sum

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^N T_i(1 - \exp(-\sigma(\mathbf{r}(t_i))\delta_i))\mathbf{c}(\mathbf{r}(t_i), \mathbf{d}), \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma(\mathbf{r}(t_j))\delta_j\right)$$

One remaining question is how to select the sequence of t_0, \dots, t_N to achieve an accurate estimate. This can be especially important for thin features, where a coarse sampling of points along the ray

may completely miss a detail of the object. To address this problem, Mildenhall et al. [38] suggest a two-stage rendering procedure. In the first stage, timesteps t_i are sampled along uniform intervals of a ray, giving a coarse estimate of the predicted color \hat{C}_c . In computing this integral, they also compute weights proportional to the influence of each point along the ray:

$$w_i \sim T_i(1 - \exp(-\sigma(\mathbf{r}(t_i))\delta_i))$$

To sample timesteps for the fine rendering stage, Mildenhall et al. [38] use w_i to define a piecewise-constant PDF along the ray. This allows a new set of t_i to be sampled around points of high density in the scene. While Mildenhall et al. [38] use two separate NeRF models for the coarse and fine rendering stages, we instead share the parameters between the two stages but use separate output heads for the coarse and fine densities and colors.

For notational convenience in later sections, we additionally define the transmittance of a ray as follows. Intuitively, this is the complement of the opacity or alpha value of a ray:

$$\hat{T}(\mathbf{r}) = 1 - \exp\left(-\sum_{i=1}^N \sigma(\mathbf{r}(t_i))\delta_i\right)$$

2.2 Signed Distance Functions and Texture Fields (STF)

Throughout this paper, we use the abbreviation *STF* to refer to an implicit function which produces both signed distances and texture colors. This section gives some background on how these implicit functions can be used to construct meshes and produce renderings.

Signed distance functions (SDFs) are a classic way to represent a 3D shape as a scalar field. In particular, an SDF f maps a coordinate \mathbf{x} to a scalar $f(\mathbf{x}) = d$, such that $|d|$ is the distance of \mathbf{x} to the nearest point on the surface of the shape, and $d < 0$ if the point is outside of the shape. As a result of this definition, the level set $f(\mathbf{x}) = 0$ defines the boundary of the shape, and $\text{sign}(d)$ determines normal orientation along the boundary. Methods such as marching cubes [35] or marching tetrahedra [11] can be used to construct meshes from this level set.

Shen et al. [56] present DMTet, a generative model over 3D shapes that leverages SDFs. DMTet produces SDF values s_i and displacements Δv_i for each vertex v_i in a dense spatial grid. The SDF values are fed through a differentiable marching tetrahedra implementation to produce an initial mesh, and then the resulting vertices are offset using the additional vector Δv_i . They also employ a subdivision procedure to efficiently obtain more detailed meshes, but we do not consider this in our work for the sake of simplicity.

Gao et al. [18] propose GET3D, which augments DMTet with additional texture information. In particular, they train a separate model to predict RGB colors \mathbf{c} for each surface point \mathbf{p} . This implicit model can be queried at surface points during rendering, or offline to construct an explicit texture. GET3D uses a differentiable rasterization library [32] to produce rendered images for generated meshes. This provides an avenue to train the implicit function end-to-end with only image-space gradients.

2.3 Diffusion Models

Our work leverages denoising diffusion [58, 60, 22] to model a high-dimensional continuous distribution. We employ the Gaussian diffusion setup of Ho et al. [22], which defines a diffusion process that begins at a data sample x_0 and gradually applies Gaussian noise to arrive at increasingly noisy samples x_1, x_2, \dots, x_T . Typically, the noising process is set up such that x_T is almost indistinguishable from Gaussian noise. In practice, we never run the noising process sequentially, but instead “jump” directly to a noised version of a sample according to

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is random noise, and $\bar{\alpha}_t$ is a monotonically decreasing noise schedule such that $\bar{\alpha}_0 = 1$. Ho et al. [22] train a model $\epsilon_\theta(x_t, t)$ on a data distribution $q(x_0)$ by minimizing the objective:

$$L_{\text{simple}} = E_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0, \mathbf{I}), t \sim U[1, T]} \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2$$

However, Ho et al. [22] also note an alternative but equivalent parameterization of the diffusion model, which we use in our work. In particular, we parameterize our model as $x_\theta(x_t, t)$ and train it to directly predict the denoised sample x_0 by minimizing

$$L_{x_0} = E_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0, \mathbf{I}), t \sim U[1, T]} \|x_\theta(x_t, t) - x_0\|_2^2$$

To sample from a diffusion model, one starts at a random noise sample x_T and gradually denoises it into samples x_{T-1}, \dots, x_0 to obtain a sample x_0 from the approximated data distribution. While early work on these models focused on stochastic sampling processes [58, 60, 22], other works propose alternative sampling methods which often draw on the relationship between diffusion models and ordinary differential equations [59, 61]. In our work, we employ the Heun sampler proposed by Karras et al. [27], as we found it to produce high-quality samples with reasonable latency.

For conditional diffusion models, it is possible to improve sample quality at the cost of diversity using a guidance technique. Dhariwal and Nichol [9] first showed this effect using image-space gradients from a noise-aware classifier, and Ho and Salimans [21] later proposed classifier-free guidance to remove the need for a separate classifier. To utilize classifier-free guidance, we train our diffusion model to condition on some information y (e.g. a conditioning image or textual description), but randomly drop this signal during training to enable the model to make unconditional predictions. During sampling, we then adjust our model prediction as follows:

$$\hat{x}_\theta(x_t, t|y) = x_\theta(x_t, t) + s \cdot (x_\theta(x_t, t|y) - x_\theta(x_t, t))$$

where s is a guidance scale. When $s = 0$ or $s = 1$, this is equivalent to regular unconditional or conditional sampling, respectively. Setting $s > 1$ typically produces more coherent but less diverse samples. We employ this technique for all of our models, finding (as expected) that guidance is necessary to obtain the best results.

2.4 Latent Diffusion

While diffusion can be applied to any distribution of vectors, it is often applied directly to signals such as pixels of images. However, it is also possible to use diffusion to generate samples in a continuous latent space.

Rombach et al. [52] propose Latent Diffusion Models (LDMs) as a two-stage generation technique for images. Under the LDM framework, they first train an encoder to produce latents $z = E(x)$ and a decoder to produce reconstructions $\tilde{x} = D(z)$. The encoder and decoder are trained in tandem to minimize a perceptual loss between \tilde{x} and x , as well as a patchwise discriminator loss on \tilde{x} . After these models are trained, a second diffusion model is trained directly on encoded dataset samples. In particular, each dataset example x_i is encoded into a latent z_i , and then z_i is used as a training example for the diffusion model. To generate new samples, the diffusion model first generates a latent sample z , and then $D(z)$ yields an image. In the original LDM setup, the latents z are lower-dimensional than the original images, and Rombach et al. [52] propose to either regularize z towards a normal distribution using a KL penalty, or to apply a vector quantization layer [63] to prevent z from being difficult to model.

Our work leverages the above approach, but makes several simplifications. First, we do not use a perceptual loss or a GAN-based objective for our reconstructions, but rather a simple L_1 or L_2 reconstruction loss. Additionally, instead of using KL regularization or vector quantization to bottleneck our latents, we clamp them to a fixed numerical range and add diffusion-style noise.

3 Related Work

An existing body of work aims to generate 3D models by training auto-encoders on explicit 3D representations and then training generative models in the resulting latent space. Achlioptas et al. [1] train an auto-encoder on point clouds, and experiment with both GANs [19] and GMMs [8] to model the resulting latent space. Yang et al. [69] likewise train a point cloud auto-encoder, but their decoder is itself a conditional generative model (i.e. a normalizing flow [51]) over individual points in the point cloud; they also employ normalizing flows to model the latent space. Luo and Hu [36] explore a similar technique, but use a diffusion model for the decoder instead of a normalizing flow. Zeng

et al. [71] train a hierarchical auto-encoder, where the second stage encodes a point cloud of latent vectors instead of a single latent code; they employ diffusion models at both stages of the hierarchy. Sanghi et al. [55] train a two-stage vector quantized auto-encoder [63, 50] on voxel occupancy grids, and model the resulting latent sequences autoregressively. Unlike our work, these approaches all rely on explicit output representations which are often bound to a fixed resolution or lack the ability to fully express a 3D asset.

More similar to our own method, some prior works have explored 3D auto-encoders with implicit decoders. Fu et al. [16] encode grids of SDF samples into latents which are used to condition an implicit SDF model. Sanghi et al. [54] encode voxel grids into latents which are used to condition an implicit occupancy network. Liu et al. [34] train a voxel-based encoder and separate implicit occupancy and color decoders. Kosiorek et al. [30] encode rendered views of a scene into latent vectors of a VAE, and this latent vector is used to condition a NeRF. Most similar to our encoder setup, Chen and Wang [6] use a transformer-based architecture to directly produce the parameters of an MLP conditioned on rendered views. We extend this prior body of work with Shap-E, which produces more expressive implicit representations and is trained at a larger scale than most prior work.

While the above methods all train both encoders and decoders, other works aim to produce latent-conditional implicit 3D representations without a learned encoder. Park et al. [43] train what they call an “auto-decoder”, which uses a learned table of embedding vectors for each example in the dataset. In their case, they train an implicit SDF decoder that conditions on these per-sample latent vectors. Bautista et al. [4] uses a similar strategy to learn per-scene latent codes to condition a NeRF decoder. Dupont et al. [12] employ meta-learning to encode dataset examples as implicit functions. In their setup, they “encode” an example into (a subset of) the parameters of an implicit function by taking gradient steps on a reconstruction objective. Concurrently to our work, Erkoç et al. [14] utilize diffusion to directly generate the implicit MLP weights; however, akin to [12], their method requires fitting NeRF parameters for each scene through gradient-based optimization. Wang et al. [66] pursue a related approach, jointly training separate NeRFs for every sample in a dataset, but share a subset of the parameters to ensure that all resulting models use an aligned representation space. These approaches have the advantage that they do not require an explicit input representation. However, they can be expensive to scale with increasing dataset size, as each new sample requires multiple gradient steps. Moreover, this scalability issue is likely more pronounced for methods that do not incorporate meta-learning.

Several methods for 3D generation use gradient-based optimization to produce individual samples, often in the form of an implicit function. DreamFields [26] optimizes the parameters of a NeRF to match a text prompt according to a CLIP-based [46] objective. DreamFusion [45] is a similar method with a different objective based on the output of a text-conditional image diffusion model. Lin et al. [33] extend DreamFusion by optimizing a mesh representation in a second stage, leveraging the fact that meshes can be rendered more efficiently at higher resolution. Wang et al. [65] propose a different approach for leveraging text-to-image diffusion models, using them to optimize a differentiable 3D voxel grid rather than an MLP-based NeRF. While most of these approaches optimize implicit functions, Khalid et al. [28] optimize the numerical parameters of a mesh itself, starting from a spherical mesh and gradually deforming it to match a text prompt. One common shortcoming of all of these approaches is that they require expensive optimization procedures, and a lot of work must be repeated for every sample that is generated. This is in contrast to direct generative models, which can potentially amortize this work by pre-training on a large dataset.

4 Method

In our method, we first train an encoder to produce implicit representations, and then train diffusion models on the latent representations produced by the encoder. Our method proceeds in two steps:

1. We train an encoder to produce the parameters of an implicit function given a dense explicit representation of a known 3D asset (Section 4.2). In particular, the encoder produces a latent representation of a 3D asset which is then linearly projected to obtain weights of a multi-layer perceptron (MLP).
2. We train a diffusion prior on a dataset of latents obtained by applying the encoder to our dataset (Section 4.3). This model is conditioned on either images or text descriptions.

We train all of our models on a large dataset of 3D assets with corresponding renderings, point clouds, and text captions (Section 4.1).

4.1 Dataset

For most of our experiments, we employ the same dataset of underlying 3D assets as Nichol et al. [41], allowing for fairer comparisons with their method. However, we slightly extend the original post-processing as follows:

- For computing point clouds, we render 60 views of each object instead of 20. We found that using only 20 views sometimes resulted in small cracks (due to blind spots) in the inferred point clouds.
- We produce point clouds of 16K points instead of 4K.
- When rendering views for training our encoder, we simplify the lighting and materials. In particular, all models are rendered with a fixed lighting configuration that only supports diffuse and ambient shading. This makes it easier to match the lighting setup with a differentiable renderer.

For our text-conditional model and the corresponding Point-E baseline, we employ an expanded dataset of underlying 3D assets and text captions. For this dataset, we collected roughly 1 million more 3D assets from high-quality data sources. Additionally, we gathered 120K captions from human labelers for high-quality subsets of our dataset. During training of our text-to-3D models, we randomly choose between human-provided labels and the original text captions when both are available.

4.2 3D Encoder

Our encoder architecture is visualized in Figure 2. We feed the encoder both point clouds and rendered views of a 3D asset, and it outputs the parameters of a multi-layer perceptron (MLP) that represents the asset as an implicit function. Both the point cloud and input views are processed via cross-attention, which is followed by a transformer backbone that produces latent representations as a sequence of vectors. Each vector in this sequence is then passed through a latent bottleneck and projection layer whose output is treated as a single row of the resulting MLP weight matrices. During training, the MLP is queried and the outputs are used in either an image reconstruction loss or a distillation loss. For more details, see Appendix A.1.

We pre-train our encoder using only a NeRF rendering objective (Section 4.2.1), as we found this to be more stable to optimize than mesh-based objectives. After NeRF pre-training, we add additional output heads for SDF and texture color predictions, and train these heads using a two-stage process (Section 4.2.2). We show reconstructions of 3D assets for various checkpoints of our encoder with both rendering methods in Figure 3.

4.2.1 Decoding with NeRF Rendering

We mostly follow the original NeRF formulation [38], except that we share the parameters between the coarse and fine models.² We randomly sample 4096 rays for each training example, and minimize an L_1 loss³ between the true color $C(\mathbf{r})$ and the predicted color from the NeRF:

$$L_{\text{RGB}} = E_{\mathbf{r} \in R} [||\hat{C}_c(\mathbf{r}) - C(\mathbf{r})||_1 + ||\hat{C}_f(\mathbf{r}) - C(\mathbf{r})||_1]$$

We also add an additional loss on the transmittance of each ray. In particular, the integrated density of a ray gives transmittance estimates $\hat{T}_c(r)$ and $\hat{T}_f(r)$ for coarse and fine rendering, respectively. We use the alpha channel from the ground-truth renderings to obtain transmittance targets $T(r)$, giving a second loss:

$$L_T = E_{\mathbf{r} \in R} [||\hat{T}_c(\mathbf{r}) - T(\mathbf{r})||_1 + ||\hat{T}_f(\mathbf{r}) - T(\mathbf{r})||_1]$$

²We use different linear output heads to produce coarse and fine predictions.

³In preliminary scans, we found that L_1 loss outperformed L_2 loss on PSNR after an initial warmup period where L_1 was worse.

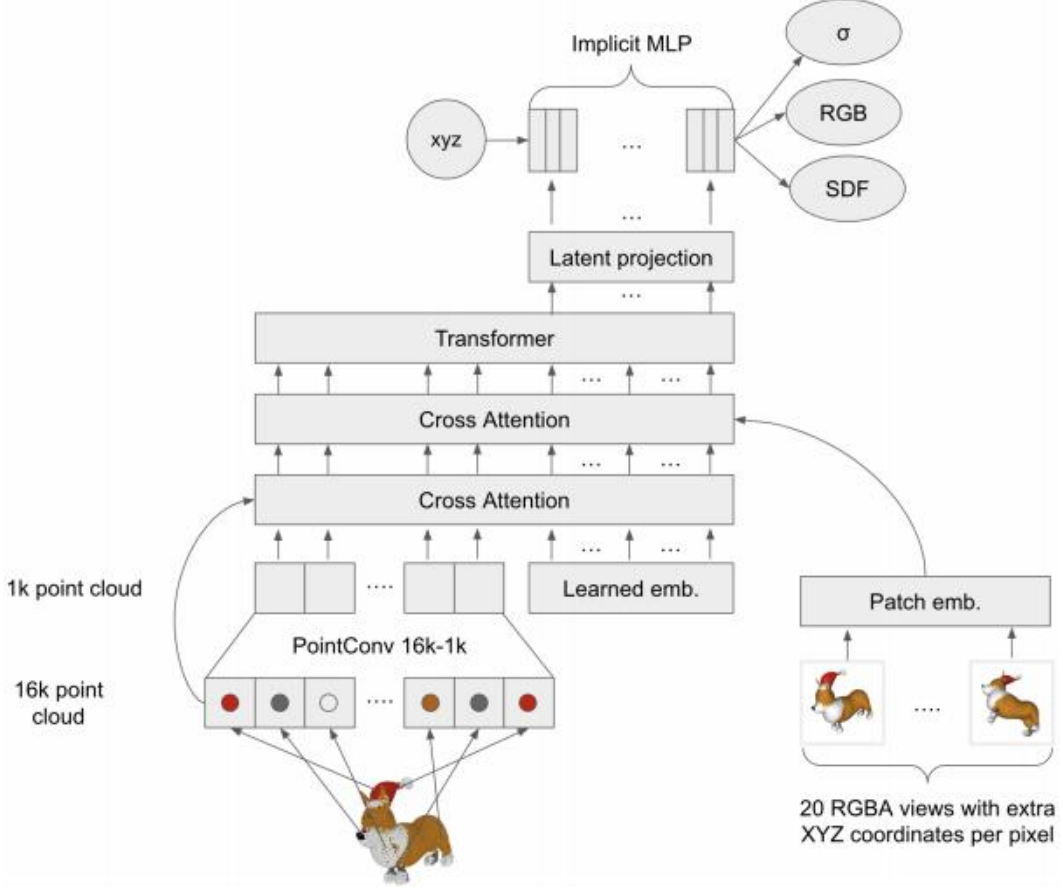


Figure 2: An overview of our encoder architecture. The encoder ingests both 16k resolution RGB point clouds and rendered RGBA images with augmented spatial coordinates for each foreground pixel. It outputs parameters of an MLP, which then acts as both a NeRF and a signed texture field (STF).

We then optimize the joint objective:

$$L_{\text{NeRF}} = L_{\text{RGB}} + L_T$$

4.2.2 Decoding with STF Rendering

After NeRF-only pre-training, we add additional STF output heads to our MLPs which predict SDF values and texture colors. To construct a triangle mesh, we query the SDF at vertices along a regular 128^3 grid and apply a differentiable implementation of Marching Cubes 33 [62]. We then query the texture color head at each vertex of the resulting mesh. We differentiably render the resulting textured mesh using PyTorch3D [49]. We always render with the same (diffuse) lighting configuration which is identical to the lighting configuration used to preprocess our dataset.

In preliminary experiments, we found that randomly-initialized STF output heads were unstable and difficult to train with a rendering-based objective. To alleviate this issue, we first distill approximations of the SDF and texture color into these output heads before directly training with differentiable rendering. In particular, we randomly sample input coordinates and obtain SDF distillation targets using the Point-E SDF regression model, and RGB targets using the color of the nearest neighbor in the asset’s RGB point cloud. During distillation training, we use a sum of distillation losses and the pre-training NeRF loss:

$$L_{\text{distill}} = L_{\text{NeRF}} + E_{\mathbf{x} \sim U[-1,1]^3} [||\text{SDF}_{\theta}(\mathbf{x}) - \text{SDF}_{\text{regression}}(\mathbf{x})||_1 + ||\text{RGB}_{\theta}(\mathbf{x}) - \text{RGB}_{\text{NN}}(\mathbf{x})||_1]$$

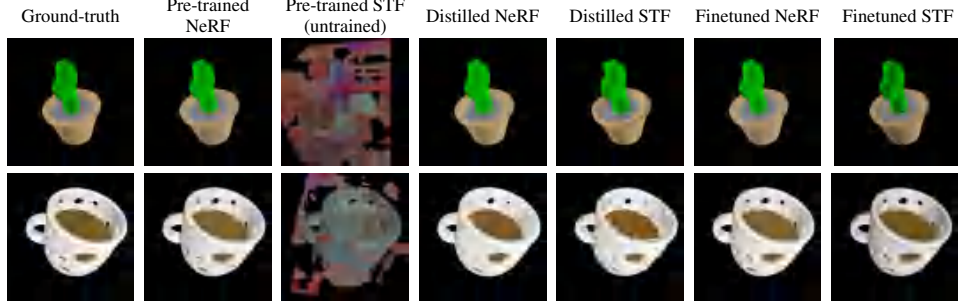


Figure 3: 3D asset reconstructions from different rendering modes and checkpoints. Surprisingly, we find that randomly initialized STF heads still produce some elements of the original shape, likely because the previous layer activations are used for NeRF outputs. While distillation improves STF rendering results, it produces rough looking objects. Fine-tuning on both rendering methods yields the best reconstructions.

Once the STF output heads have been initialized to reasonable values via distillation, we fine-tune the encoder for both NeRF and STF rendering end-to-end. We found it unstable to use L_1 loss for STF rendering, so we instead use L_2 loss only for this rendering method. In particular, we optimize the following loss for STF rendering:

$$L_{\text{STF}} = \frac{1}{N \cdot s^2} \sum_{i=1}^N ||\text{Render}(\text{Mesh}_i) - \text{Image}_i||_2^2$$

where N is the number of images, s is the image resolution, Mesh_i is the constructed mesh for sample i , Image_i is a target RGBA rendering for image i , and $\text{Render}(x)$ renders a mesh using a differentiable renderer. We do not include a separate transmittance loss, since this is already captured by the alpha channel of the image.

For this final fine-tuning step, we optimize the summed objective:

$$L_{\text{FT}} = L_{\text{NeRF}} + L_{\text{STF}}$$

4.3 Latent Diffusion

For our generative models, we adopt the transformer-based diffusion architecture of Point-E, but replace point clouds with sequences of latent vectors. Our latents are sequences of shape 1024×1024 , and we feed this into the transformer as a sequence of 1024 tokens where each token corresponds to a different row of the MLP weight matrices. As a result, our models are roughly compute equivalent to the base Point-E models (i.e. have the same context length and width) while generating samples in a much higher-dimensional space due to the increase of input and output channels.

We follow the same conditioning strategies as Point-E. For image-conditional generation, we prepend a 256-token CLIP embedding sequence to the Transformer context. For text-conditional generation, we prepend a single token containing the CLIP text embedding. To support classifier-free guidance, we randomly set the conditioning information to zero during training with probability 0.1.

Unlike Point-E, we do not parameterize our diffusion model outputs as ϵ predictions. Instead, we directly predict x_0 , which is algebraically equivalent to predicting ϵ , but produced more coherent samples in early experiments. The same observation was made by Ramesh et al. [48], who opted to use x_0 prediction when generating CLIP latent vectors with diffusion models.

5 Results

5.1 Encoder Evaluation

We track two render-based metrics throughout the encoder training process. First, we evaluate the peak signal-to-noise ratio (PSNR) between reconstructions and ground-truth rendered images. Additionally, to measure our encoder’s ability to capture semantically relevant details of 3D assets,

Table 1: Evaluating the encoder after each stage of training. We evaluate PSNR between reconstructions and ground-truth renders, as well as CLIP R-Precision on reconstructions of samples from Point-E 1B (where the peak performance is roughly 46.8%).

Stage	NeRF PSNR (dB)	STF PSNR (dB)	NeRF Point-E CLIP R-Precision	STF Point-E CLIP R-Precision
Pre-training (300K)	33.2	-	44.3%	-
Pre-training (600K)	34.5	-	45.2%	-
Distillation	32.9	23.9	42.6%	41.1%
Fine-tuning	35.4	31.3	45.3%	44.0%

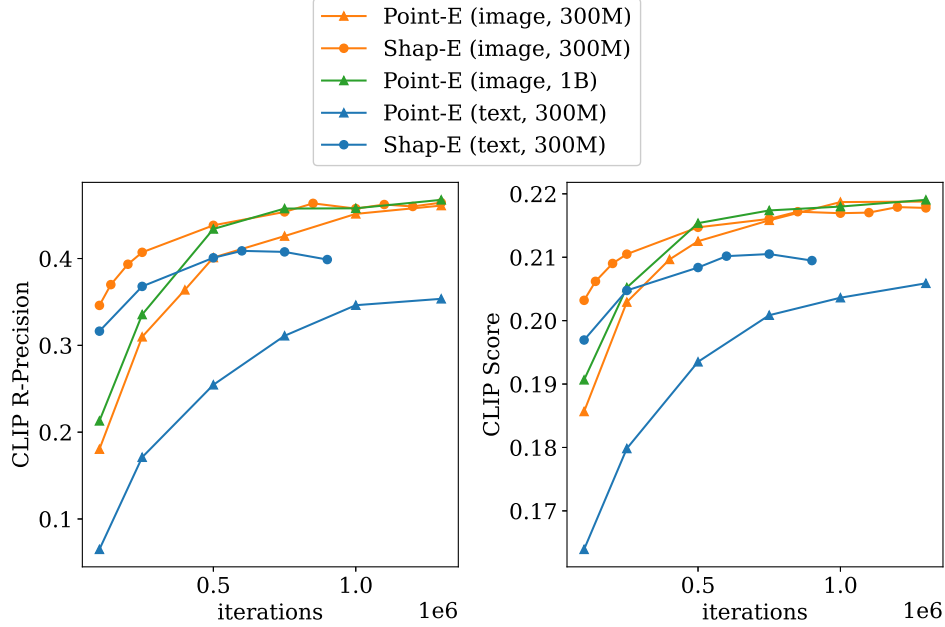


Figure 4: Evaluations throughout training for both Shap-E and Point-E. For each checkpoint for both models, we take the maximum value when sweeping over guidance scales $\{2.0, 3.0, 4.0, 5.0, 8.0, 10.0, 15.0\}$.

we encode meshes produced by the largest Point-E model and re-evaluate the CLIP R-Precision of the reconstructed NeRF and STF renders. Table 1 tracks these two metrics over the different stages of training. We find that distillation hurts NeRF reconstruction quality, but fine-tuning recovers (and slightly boosts) NeRF quality while drastically increasing the quality of STF renders.

5.2 Comparison to Point-E

Our latent diffusion model shares the same architecture, training dataset, and conditioning modes as Point-E.⁴ As a result, comparing to Point-E helps us isolate the effects of generating implicit neural representations rather than an explicit representation. We compare these methods throughout training on sample-based evaluation metrics in Figure 4. As done by Jain et al. [26] and various follow-up literature, we compute CLIP R-Precision [42] on a set of COCO validation prompts. We also evaluate CLIP score on these same prompts, as this metric is often used for measuring image generation quality [40]. We only train comparable 300 million parameter models, but we also plot evaluations for the largest (1 billion parameter) Point-E model for completeness.

In the text-conditional setting, we observe that Shap-E improves on both metrics over the comparable Point-E model. To rule out the possibility that this gap is due to perceptually small differences,

⁴However, note that Shap-E depends on a separate encoder model, while Point-E depends on separate upsampler and SDF models. Only the base diffusion model architecture is the same.

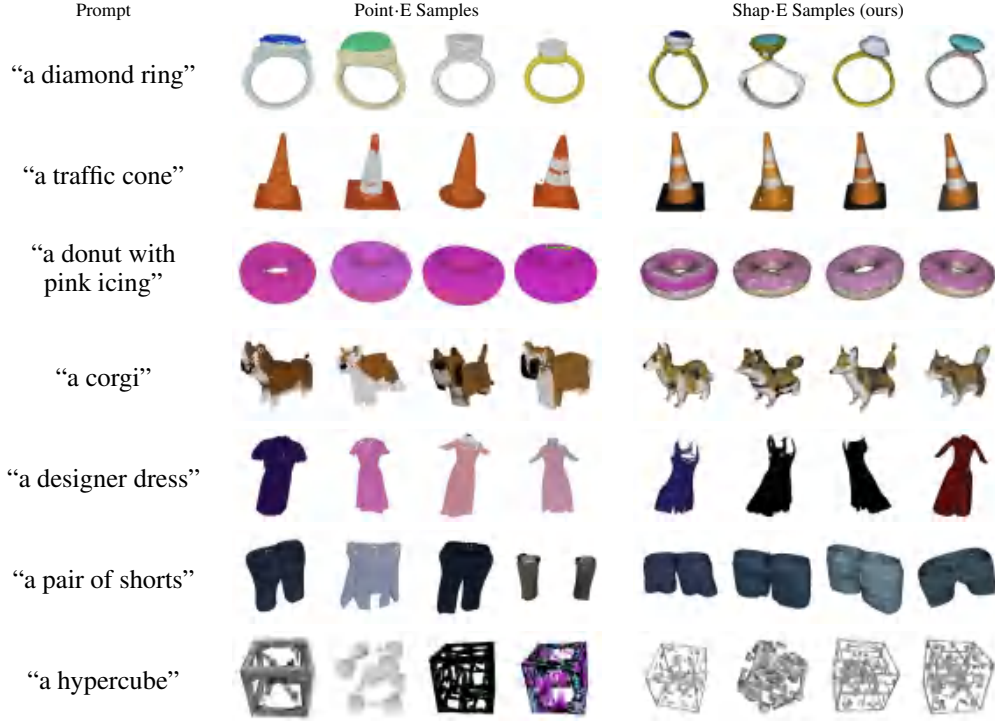


Figure 5: Examples of text prompts for which text-conditional Point-E and Shap-E consistently exhibit qualitatively different behavior. For each prompt, we show four random samples from both models, which were trained on the same dataset with the same base model size.

we also show qualitative samples in Figure 5, finding that these models often produce qualitatively different samples for the same text prompts. We also observe that our text-conditional Shap-E begins to get worse on evaluations before the end of training. In Appendix B, we argue that this is likely due to overfitting to the text captions, and we use an early-stopped checkpoint for all figures and tables.

Unlike the text-conditional case, our image-conditional Shap-E and Point-E models reach roughly the same final evaluation performance, with a slight advantage for Shap-E in CLIP R-Precision and a slight disadvantage in CLIP score. To investigate this phenomenon more deeply, we inspected samples from both models. We initially expected to see qualitatively different behavior from the two models, since they produce samples in different representation spaces. However, we discovered that both models tend to share similar failure cases, as shown in Figure 6a. This suggests that the training data, model architecture, and conditioning images affect the resulting samples more than the chosen representation space.

However, we do still observe some qualitative differences between the two image-conditional models. For example, in the first row of Figure 6b, we find that Point-E sometimes ignores the small slits in the bench, whereas Shap-E attempts to model them. We hypothesize that this particular difference could occur because point clouds are a poor representation for thin features or gaps. Also, we observe in Table 1 that the 3D encoder slightly reduces CLIP R-Precision when applied to Point-E samples. Since Shap-E achieves comparable CLIP R-Precision as Point-E, we hypothesize that Shap-E must generate qualitatively different samples for some prompts which are not bottlenecked by the encoder. This further suggests that explicit and implicit modeling can still learn distinct features from the same data and model architecture.

5.3 Comparison to Other Methods

We compare Shap-E to a broader class of 3D generative techniques on the CLIP R-Precision metric in Table 2. As done by Nichol et al. [41], we include sampling latency in this table to highlight that the superior sample quality of optimization-based methods comes at a significant inference cost. We also



(a) Shared failure cases between image-conditional Shap-E and Point-E. In the first example, both models counter-intuitively infer an occluded handle on the mug. In the second, both models incorrectly interpret the proportions of the depicted animal.



(b) Conditioning images for which both Shap-E and Point-E succeed.

Figure 6: Randomly selected image-conditional samples from both Point-E and Shap-E for the same conditioning images.

note that Shap-E enjoys faster inference than Point-E because Shap-E does not require an additional upsampling diffusion model.

6 Limitations and Future Work

While our text-conditional model can understand many single object prompts with simple attributes, it has a limited ability to compose concepts. In Figure 7, we find that this model struggles to bind multiple attributes to different objects, and fails to reliably produce the correct number of objects when asked for more than two. These failures are likely the result of limited paired training data, and could potentially be alleviated by gathering or generating larger annotated 3D datasets.

Table 2: Comparison of 3D generation techniques on the CLIP R-Precision metric on COCO evaluation prompts. Compute estimates and other methods’ values are taken from Nichol et al. [41]. *The best text-conditional results are obtained using our expanded dataset of 3D assets.

Method	ViT-B/32	ViT-L/14	Latency
DreamFields	78.6%	82.9%	~ 200 V100-hr
CLIP-Mesh	67.8%	74.5%	~ 17 V100-min
DreamFusion	75.1%	79.7%	~ 12 V100-hr
Point-E (300M, text-only)	33.6%*	35.5%*	25 V100-sec
Shap-E (300M, text-only)	37.8%*	40.9%*	13 V100-sec
Point-E (300M)	40.3%	45.6%	1.2 V100-min
Point-E (1B)	41.1%	46.8%	1.5 V100-min
Shap-E (300M)	41.1%	46.4%	1.0 V100-min
Conditioning images	69.6%	86.6%	-

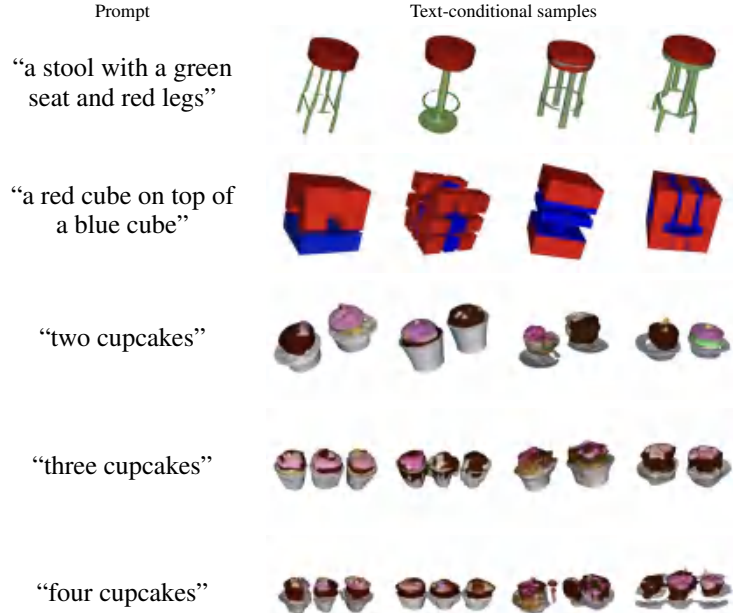


Figure 7: Examples of text-conditional Shap-E samples prompts which require counting and attribute binding.

Additionally, while Shap-E can often produce recognizable 3D assets, the resulting samples often look rough or lack fine details. Notably, Figure 3 shows that the encoder itself sometimes loses detailed textures (e.g. the stripes on the cactus), indicating that improved encoders could potentially recover some of the lost generation quality.

For the best results, Shap-E could potentially be combined with optimization-based 3D generative techniques. For example, a NeRF or mesh produced by Shap-E could be used to initialize an optimization-based approach such as DreamFusion, potentially leading to faster convergence. Alternatively, image-based objectives could be used to guide the Shap-E sampling process, as we briefly explore in Appendix D.

7 Conclusion

We present Shap-E, a latent diffusion model over a space of 3D implicit functions that can be rendered as both NeRFs and textured meshes. We find that Shap-E matches or outperforms a similar explicit generative model given the same dataset, model architecture, and training compute. We also find that our pure text-conditional models can generate diverse, interesting objects without relying on images as an intermediate representation. These results highlight the potential of generating implicit representations, especially in domains like 3D where they can offer more flexibility than explicit representations.

8 Acknowledgments

Our thanks go to Prafulla Dhariwal, Joyce Lee, Jack Rae, and Mark Chen for helpful discussions, and to all contributors of ChatGPT, which provided valuable writing feedback.

References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. *arXiv:1707.02392*, 2017.
- [2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023. URL <https://arxiv.org/abs/2301.11325>.
- [3] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. *arXiv:2303.13277*, 2023.
- [4] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Josh Susskind. Gaudi: A neural architect for immersive 3d scene generation. *arXiv:2207.13751*, 2022.
- [5] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audioldm: a language modeling approach to audio generation, 2022. URL <https://arxiv.org/abs/2209.03143>.
- [6] Yinbo Chen and Xiaolong Wang. Transformers as meta-learners for implicit neural representations, 2022. URL <https://arxiv.org/abs/2208.02801>.
- [7] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models, 2022.
- [8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977. ISSN 00359246. URL <http://www.jstor.org/stable/2984875>.
- [9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv:2105.05233*, 2021.
- [10] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.
- [11] Akio Doi and Akio Koide. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE Transactions on Information and Systems*, 74:214–224, 1991.
- [12] Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. *arXiv:2201.12204*, 2022.

- [13] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *arXiv:1906.04032*, 2019.
- [14] Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Hyperdiffusion: Generating implicit neural fields with weight-space diffusion, 2023.
- [15] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv:2210.15257*, 2022.
- [16] Rao Fu, Xiao Zhan, Yiwen Chen, Daniel Ritchie, and Srinath Sridhar. Shapecrafter: A recursive text-conditioned 3d shape generation model. *arXiv:2207.09446*, 2022.
- [17] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv:2203.13131*, 2022.
- [18] 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. *arXiv:2209.11163*, 2022.
- [19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv:1406.2661*, 2014.
- [20] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016.
- [21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL <https://openreview.net/forum?id=qw8AKxfYbI>.
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv:2006.11239*, 2020.
- [23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.
- [24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.
- [25] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. Noise2music: Text-conditioned music generation with diffusion models, 2023. URL <https://arxiv.org/abs/2302.03917>.
- [26] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. *arXiv:2112.01455*, 2021.
- [27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv:2206.00364*, 2022.
- [28] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. *arXiv:2203.13333*, 2022.
- [29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
- [30] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J Rezende. NeRF-VAE: A geometry aware 3D scene generative model. *arXiv:2104.00587*, April 2021.

- [31] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation, 2022. URL <https://arxiv.org/abs/2209.15352>.
- [32] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *arXiv:2011.03277*, 2020.
- [33] 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. *arXiv:2211.10440*, 2022.
- [34] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. *arXiv:2203.14622*, 2022.
- [35] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In Maureen C. Stone, editor, *SIGGRAPH*, pages 163–169. ACM, 1987. ISBN 0-89791-227-6. URL <http://dblp.uni-trier.de/db/conf/siggraph/siggraph1987.html#LorensenC87>.
- [36] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. *arXiv:2103.01458*, 2021.
- [37] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. *arXiv:1710.03740*, 2017.
- [38] 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. *arXiv:2003.08934*, 2020.
- [39] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv:2102.09672*, 2021.
- [40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021.
- [41] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv:2212.08751*, 2022.
- [42] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL <https://openreview.net/forum?id=bKBhQhPeKaF>.
- [43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *arXiv:1901.05103*, 2019.
- [44] Christine Payne. Musenet. *OpenAI blog*, 2019. URL <https://openai.com/blog/musenet>.
- [45] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv:2209.14988*, 2022.
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021.
- [47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv:2102.12092*, 2021.
- [48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

- [49] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.
- [50] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *arXiv:1906.00446*, 2019.
- [51] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv:1505.05770*, 2015.
- [52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv:2112.10752*, 2021.
- [53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*, 2022.
- [54] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. *arXiv:2110.02624*, 2021.
- [55] Aditya Sanghi, Rao Fu, Vivian Liu, Karl Willis, Hooman Shayani, Amir Hosein Khasahmadi, Srinath Sridhar, and Daniel Ritchie. Textcraft: Zero-shot generation of high-fidelity and diverse shapes from text. *arXiv:2211.01427*, 2022.
- [56] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *arXiv:2111.04276*, 2021.
- [57] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022.
- [58] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv:1503.03585*, 2015.
- [59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020.
- [60] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv:arXiv:1907.05600*, 2020.
- [61] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020.
- [62] Evgueni Tcherniaev. Marching cubes 33: Construction of topologically correct isosurfaces. 01 1996.
- [63] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv:1711.00937*, 2017.
- [64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762*, 2017.
- [65] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv:2212.00774*, 2022.
- [66] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion, 2022. URL <https://arxiv.org/abs/2212.06135>.

- [67] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv:2210.04628*, 2022.
- [68] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9:40–53, 11/2019 2019. ISSN 1927-0321. doi: <http://doi.org/10.22215/timreview/1282>. URL timreview.ca/article/1282.
- [69] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. *arXiv:1906.12320*, 2019.
- [70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv:2206.10789*, 2022.
- [71] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv:2210.06978*, 2022.
- [72] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. *arXiv:2206.06360*, 2022.

Algorithm 1 High-level pseudocode of our encoder architecture.

Input point cloud p , multiview point cloud m , learned input embedding sequence h_l .**Outputs:** latent variable h and MLP parameters θ .

```
1:  $h \leftarrow \text{Cat}([\text{PointConv}(p), h_l])$ 
2:  $h \leftarrow \text{CrossAttend}(h, \text{Proj}(p))$ 
3:  $h \leftarrow \text{CrossAttend}(h, \text{PatchEmb}(m))$ 
4:  $h \leftarrow \text{Transformer}(h)$ 
5:  $h \leftarrow h[-\text{len}(h_l) :]$ 
6:  $h \leftarrow \tanh(h)$ 
7:  $h' \leftarrow \text{DiffusionNoise}(h)$ 
8:  $\theta \leftarrow \text{Proj}(h')$ 
9: return  $h, \theta$ 
```

A Hyperparameters

A.1 Encoder Architecture

To capture details of the input 3D asset, we feed our encoder two separate representations of a 3D model:

- **Point clouds:** For each 3D asset, we pre-compute an RGB point cloud with 16,384 points.
- **Multiview point clouds:** In addition to a point cloud, we render 20 views of each 3D asset from random camera angles at 256×256 resolution. We augment each foreground pixel with an (x, y, z) surface coordinate, giving an image of shape $256 \times 256 \times 7$. We apply an 8×8 patch embedding these augmented renderings, resulting in a sequence of 20,480 vectors representing a *multiview point cloud*.

Our encoder begins by using a point convolution layer to downsample the input point cloud into a set of 1K embeddings. This set of embeddings is concatenated with a learned input embedding h_l to obtain a query sequence h . We then update h with a single cross-attention layer that references the input point cloud. Next, we update h again by cross-attending to the patch embedded multiview point cloud m . Next, we apply a transformer to h and take the 1K suffix tokens as latent vectors. We then apply a $\tanh(x)$ activation to these latents to clamp them to the range $[-1, 1]$. At this stage, we have obtained the latent vector that we target with our diffusion models, but we do not yet have the parameters of an MLP.

After computing the sequence of latents, we apply Gaussian diffusion noise $q(h_t)$ to the latents with probability 0.1. For this diffusion noise, we use the schedule $\bar{\alpha}_t = 1 - t^5$ which typically produces very little noise. After the noise and bottleneck layers, we project each latent vector to 256 dimensions and stack the resulting latents into four MLP weight matrices of size 256×256 . Our full encoder architecture is described in Algorithm 1.

A.2 Encoder Training

We pre-train our encoders for 600K iterations using Adam [29] with a learning rate of 10^{-4} and a batch size of 64. We perform STF distillation for 50K iterations with a learning rate of 10^{-5} and keep the batch size at 64. We query 32K random points on each 3D asset for STF distillation. We fine-tune on STF renders for 65K iterations with the same hyperparameters as for distillation. For each stage of training, we re-initialize the optimizer state. For pre-training, we use 16-bit precision with loss scaling [37], but we found full 32-bit precision necessary to stabilize fine-tuning.

A.3 Implicit Representations

We represent our INRs as 6-layer MLPs where the first four layers are determined by the output of an encoder; the final two layers are shared across dataset examples. We do not use any biases in these models. Input coordinates are concatenated with sinusoidal embeddings following the work of Mildenhall et al. [38] and Watson et al. [67]. In particular, each coordinate dimension x is expanded as

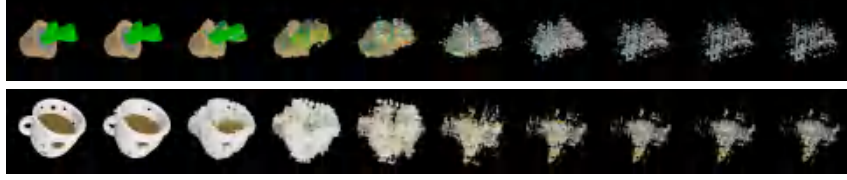


Figure 8: NeRF reconstructions of noised latents using our diffusion schedule $\bar{\alpha}_t = e^{-12t}$. The timestep t is linearly swept from 0 to 1 from left to right.

$$[x, \cos(2^0 x), \sin(2^0 x), \dots, \cos(2^{14} x), \sin(2^{14} x)]$$

Our MLPs use SiLU activations [20] between intermediate layers. The NeRF density and RGB heads are followed by sigmoid and ReLU activations, respectively. The SDF and texture color heads are followed by tanh and sigmoid activations, respectively.

Although we use direction independent lighting in all of our experiments, we found our encoders unstable to train unless we augmented their input coordinates with ray direction embeddings. Unlike typical NeRF models, our models’ density head can be influenced by the ray direction, potentially leading to view-inconsistent objects. To ensure view-consistency at test time, we always set the ray direction embeddings to zero. Despite the out-of-distribution inputs, this approach is effective, likely because the model learns to disregard the ray direction with sufficient training. It remains an open question why the ray direction is beneficial during initial pre-training, yet appears irrelevant in later stages of training.

A.4 Diffusion Models

When training our diffusion models, we employ the same hyperparameters as used for the 300M parameter Point-E models. The only difference is that we use larger input and output projections to accommodate 1024 feature channels (instead of 6).

For diffusion models, the choice of noise schedule $\bar{\alpha}_t$ can often have a big impact on sample quality [39, 27, 7]. Intuitively, the relative scale between a sample and the noise injected at a particular timestep determines how much information is destroyed at that timestep, and we would like a noise schedule that gradually destroys semantic information in the signal. In early experiments, we tried the cosine [39] and linear [22] schedules, as well as a schedule which we found visually to destroy information gradually: $\bar{\alpha}_t = e^{-12t}$ (see Figure 8). In these experiments, we found that the latter schedule performed better on evaluation metrics, and decided to use it for all future experiments.

We use similar Heun sampling hyperparameters as Point-E, but found that setting $s_{\text{churn}} = 0$ was a better choice for Shap-E, whereas $s_{\text{churn}} = 3$ was better for Point-E. Additionally, we found that, while our image-conditional models tended to prefer the same guidance scale as Point-E, our text-conditional models could tolerate much higher guidance scales while still improving on evaluations (Figure 9). We find that our best text-conditional Point-E samples are obtained using a scale of 5.0, while the best Shap-E results use a scale of 20.0.

A.5 Evaluation

When evaluating CLIP-based metrics, we render our models’ samples using NeRF at 128×128 resolution. We sample camera positions randomly around the z-axis, with a constant 30 degree elevation for all camera poses. We find that this works well in practice, since objects in our training dataset are usually oriented with the z-axis as the logical vertical direction.

B Overfitting in Text-Conditional Models

We observe that our text-conditional model begins to get worse on evaluations after roughly 600K iterations. We hypothesize that this is due to overfitting to the text captions in the dataset, since we did not observe this phenomenon in the image-conditional case. In Figure 10a, we observe that

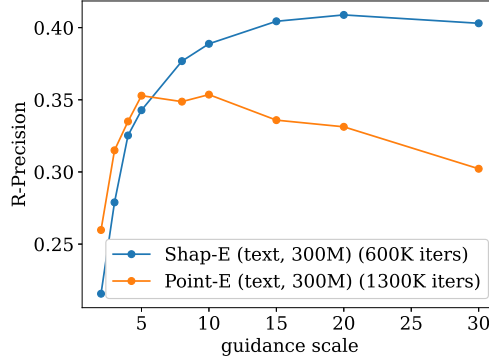
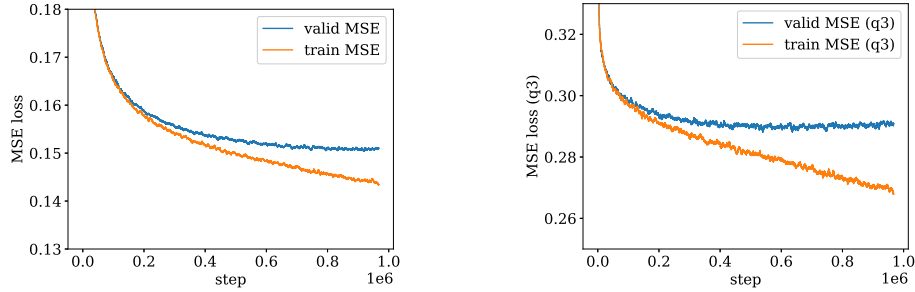


Figure 9: Evaluation sweep over guidance scale for text-conditional models. We find that Shap-E benefits from increasing guidance scale up to 20.0, whereas Point-E begins to saturate at lower guidance scales and then becomes worse.



(a) Train and validation loss averaged across all diffusion steps.

(b) Train/validation loss averaged over the noisiest quarter of the diffusion steps.

Figure 10: Training and validation losses for our text-conditional model. We find that this model overfits, and that the overfitting is stronger for the noisiest diffusion timesteps.

the training loss decreases faster than the validation loss, but that validation loss itself never starts increasing. Why, then, does the model get worse on evaluations?

To more deeply explore this overfitting, we leverage the fact that the diffusion loss is actually a sum of many different loss terms at different noise levels. In Figure 10b, we plot the training and validation losses over only the noisiest quarter of the diffusion steps, finding that in this case overfitting is more pronounced and the validation loss indeed starts increasing at about 600K iterations. Intuitively, conditioning information is more likely to affect noisier timesteps since less information can be inferred from the noised sample x_t . This supports the hypothesis that the overfitting is tied to the model’s understanding of the conditioning signal, although it may still be overfitting to other aspects of the data.

C Bias and Misuse

Biases present in our dataset are likely to impact the behavior of the models we develop. In Figure 11, we examine bias within our text-conditional model by providing it with ambiguous captions in which certain details, such as body shape or color, are left unspecified. We observe that the samples generated by the model exhibit common gender-role stereotypes in response to these ambiguous prompts.

Our models are not typically adept at producing photo-realistic samples or accurately following long and complex prompts, and this limitation comes with both benefits and drawbacks. On the positive side, it alleviates concerns regarding the potential use of our models to create convincing “DeepFakes”

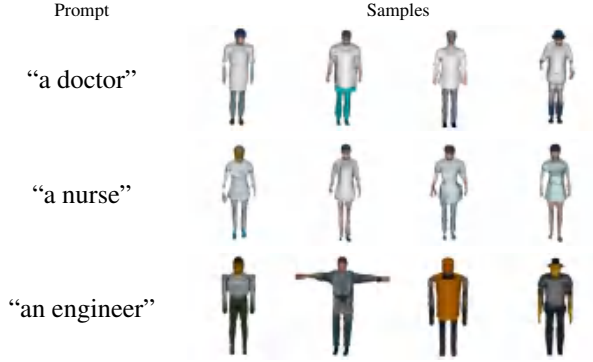


Figure 11: Examples where our text-conditional model likely exhibits biases from its dataset.



Figure 12: Examples of generated 3D objects which could have adverse consequences if used in the real world without validation.

[68]. On the negative side, it raises potential risks when our models are used in conjunction with fabrication methods such as 3D printing to create tools and parts (e.g. Figure 12). In such scenarios, 3D objects generated by the model could be introduced into the real world without undergoing adequate validation or safety testing, and this could potentially be harmful when the produced samples do not adequately meet the desired prompt.

D Guidance in Image Space

While our diffusion models operate in a latent space, we find that it is possible to guide them directly in *image space*. During sampling, we have some noised latent x_t and a corresponding model prediction $x_0 = f(x_t)$. If we treat the model prediction as a latent vector and render it with NeRF to get an image I , we can compute gradients of any image-based objective function L like so:

$$\frac{\partial L}{\partial x_t} = \frac{\partial L}{\partial I} \frac{\partial I}{\partial x_0} \frac{\partial x_0}{\partial x_t}$$

Given this gradient, we can then follow the classifier guidance setup of Dhariwal and Nichol [9] to update each diffusion step in the direction of a scaled gradient $s \cdot \frac{\partial L}{\partial x_t}$.

To test this idea, we leverage DreamFusion [45] to obtain image-space gradients that incentivize rendered images to match a text prompt. Since DreamFusion requires a powerful text-to-image diffusion model, we use the 3 billion parameter GLIDE model [40]. We sample from our text-conditional Shap-E model using 1,024 stochastic DDPM steps. At each step, we use eight rendered views of the NeRF to obtain an estimate of the DreamFusion gradient. We then scale this gradient by a hyperparameter s before applying a guided sampling step. This process takes roughly 15 minutes on eight NVIDIA V100 GPUs.

In Figure 13, we explore what happens as we increase the DreamFusion guidance scale s while keeping the diffusion noise fixed. We observe in general that this text-conditional Shap-E model is not very good on its own with DDPM sampling, failing to capture the text prompts with $s = 0$. However, as we increase s , we find that the samples tend to approach something more closely matching the prompt. Notably, this is despite the fact that we do not use most of the tricks employed by DreamFusion, such as normals-based shading or grayscale rendering.

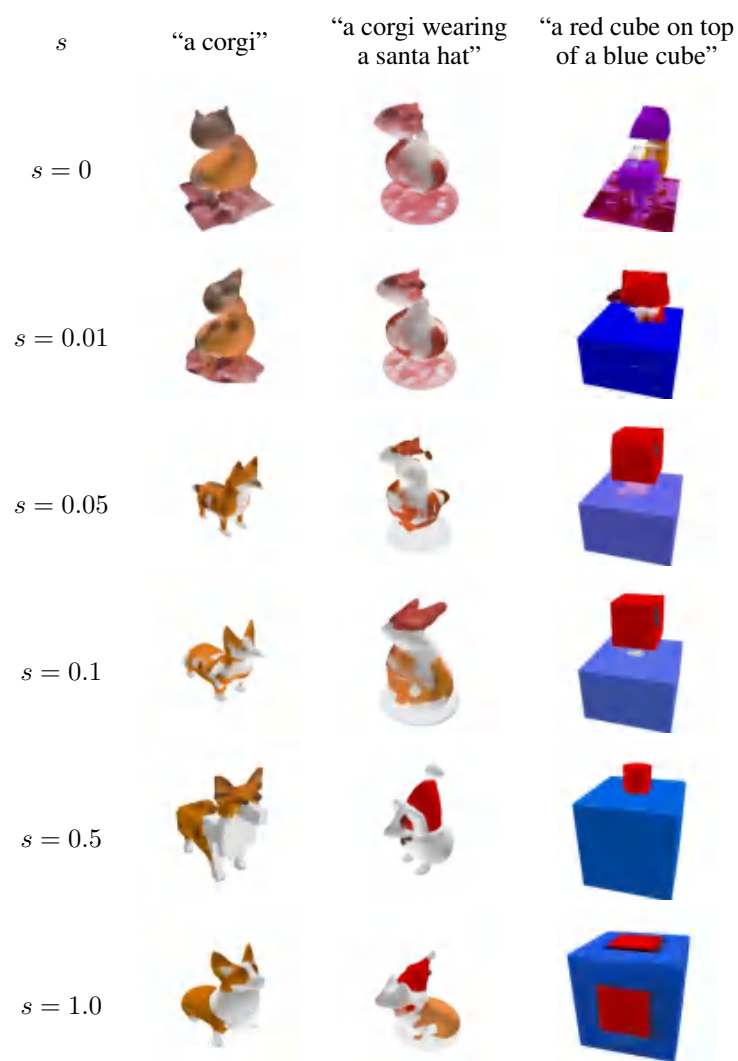


Figure 13: Using DreamFusion to guide our text-conditional Shap·E model.