Spice·E: Structural Priors in 3D Diffusion using Cross-Entity Attention

Etai Sella*
Tel Aviv University
Tel Aviv, Israel
etaisella@gmail.com

Gal Fiebelman*
Tel Aviv University
Tel Aviv, Israel
galfiebelman@mail.tau.ac.il

Noam Atia
Tel Aviv University
Tel Aviv, Israel
noamatia@mail.tau.ac.il

Hadar Averbuch-Elor
Tel Aviv University
Tel Aviv, Israel
hadar.a.elor@gmail.com

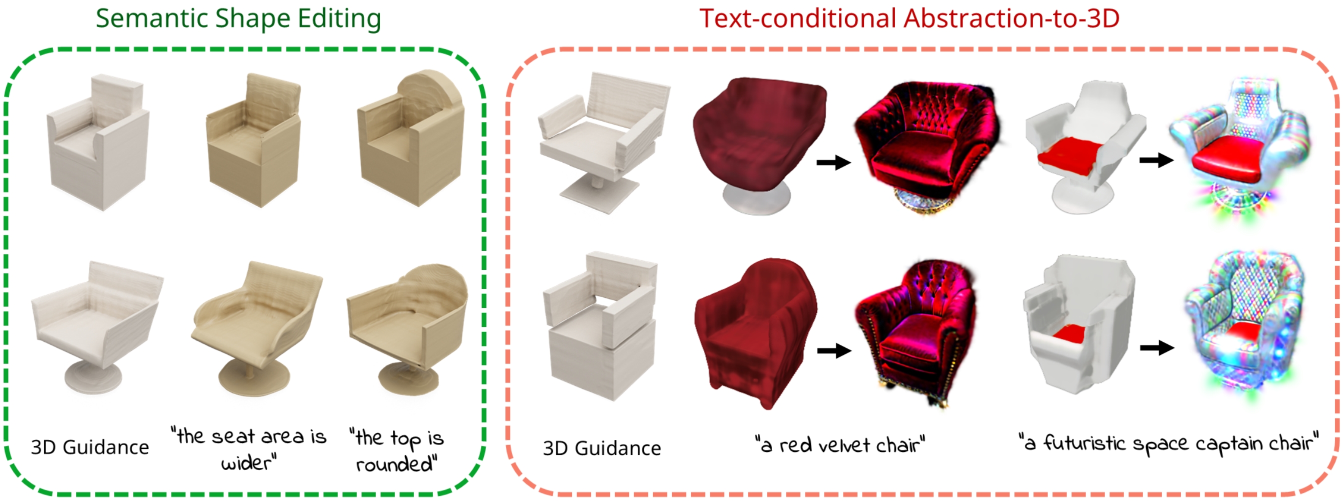

Figure 1: Our method adds structural guidance to 3D diffusion models. As illustrated above, this allows for generating text-conditional 3D shapes that enforce task-specific structural priors. For instance, input shapes can be semantically edited (left) and primitive-based abstractions can be transformed into high-quality textured shapes that conform with the target text (right). Our results can be optionally refined using an auxiliary process (represented by black arrows above).

ABSTRACT

We are witnessing rapid progress in automatically generating and manipulating 3D assets due to the availability of pretrained text-to-image diffusion models. However, time-consuming optimization procedures are required for synthesizing each sample, hindering their potential for democratizing 3D content creation. Conversely, 3D diffusion models now train on million-scale 3D datasets, yielding high-quality text-conditional 3D samples within seconds. In this work, we present Spice·E – a neural network that adds structural guidance to 3D diffusion models, extending their usage beyond text-conditional generation. At its core, our framework introduces a cross-entity attention mechanism that allows for multiple entities—in particular, paired input and guidance 3D shapes—to interact via their internal representations within the denoising network. We utilize this mechanism for learning task-specific structural priors in 3D diffusion models from auxiliary guidance shapes. We show that our approach supports a variety of applications, including 3D stylization, semantic shape editing and text-conditional abstraction-to-3D, which transforms primitive-based abstractions into highly-expressive shapes. Extensive experiments demonstrate that Spice·E achieves SOTA performance over these tasks while often being considerably faster than alternative methods. Importantly, this is accomplished without tailoring our approach for any specific task. We will release our code and trained models.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer graphics**; **Volumetric models**.

*Denotes equal contribution

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657461

KEYWORDS

Diffusion Models, 3D Generative AI, 3D Textual Editing, Conditional Generation

ACM Reference Format:
Etai Sella, Gal Fiebelman, Noam Atia, and Hadar Averbuch-Elor. 2024. Spice·E: Structural Priors in 3D Diffusion using Cross-Entity Attention. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657461

1 INTRODUCTION

Text-guided 3D generation has recently seen tremendous success, empowering us with the ability to convert our imagination into high-fidelity 3D models through the use of text [Lin et al. 2023; Poole et al. 2022; Wang et al. 2023a,b]. Consequently, there has been increasing interest in leveraging this generative power for editing existing 3D objects [Chen et al. 2023a; Metzer et al. 2023; Sella et al. 2023; Zhuang et al. 2023], a longstanding goal in computer vision and graphics [Igarashi et al. 2005; Lewis et al. 2023; Magnenat et al. 1988]. Unfortunately, these text-guided methods require timely optimization procedures for producing a single sample, as they rely on the guidance of pretrained 2D diffusion models such as Stable Diffusion [Rombach et al. 2022] over multiple rendered views, making them challenging to apply in practical settings.

In parallel with these advancements, million-scale 3D datasets pairing 3D data with text directly [Deitke et al. 2023a,b] have paved the way for the creation of powerful 3D diffusion models [Jun and Nichol 2023; Nichol et al. 2022]. These direct generative models can synthesize text-conditional 3D assets conveying complex visual concepts, and they achieve this in a matter of *seconds*, orders of magnitude faster than methods utilizing 2D diffusion models. However, they are inherently unconstrained and lack the ability to enforce structural priors while generating 3D samples, and thereby cannot be effectively utilized in the context of 3D editing applications.

Inspired by recent progress adding conditional control to 2D diffusion models [Zhang et al. 2023], we ask: How can we provide pretrained transformer-based 3D diffusion models with task-specific structural control? And importantly, how can we achieve such structural control while preserving the model's expressive power, and to do so *without* having access to (possibly) proprietary data or large computation clusters? This requires architectural modifications that maximize the utilization of pretrained weights during model finetuning on the one hand while still acquiring task-specific structural priors from auxilary guidance shapes on the other.

Accordingly, we present Spice·E (**S**tructural **P**riors **i**n 3D Diffusion using **C**ross-**E**ntity Attention)[1], a neural network that adds structural guidance to a 3D diffusion model. Our key observation is that the self-attention layers within transformer-based diffusion models can be modified to enable interaction between two different entities (*i.e.* 3D shapes) – one depicting the input and the other depicting the guidance entity. We introduce a cross-entity attention mechanism that mixes their latent representations by carefully combining their *queries* functions, which have recently been shown for being instrumental in modifying the structure of generated images [Cao et al. 2023b; Wu et al. 2023]. This operation allows for finetuning a 3D diffusion model to learn task-specific structural priors while preserving the model's generative capabilities. During inference, Spice·E receives a guidance shape in addition to a target text prompt, enabling the generation of 3D shapes conditioned on both high-level text directives and low-level structural constraints. The outputs of our system can be further refined by an auxiliary process (*i.e.*, [Yi et al. 2023]), which enhances the appearance and geometric details, albeit at the cost of increased processing time.

We show the effectiveness of our framework using different 3D editing tasks, such as semantic shape editing and text-conditional Abstraction-to-3D, which transforms a primitive-based abstract shape into a high-quality textured shape (see Figure 1 for an illustration of these tasks). We perform extensive experiments, demonstrating that our approach surpasses existing methods specifically targeting these tasks, while often being significantly faster.

2 RELATED WORKS

2.1 Text-guided Shape Manipulation

The emergence of powerful text–image representations, most notably CLIP [Radford et al. 2021], has driven progress in shape editing and manipulation via language prompts. Several methods use CLIP for stylizing input meshes, matching their 2D image projections with a target prompt [Chen et al. 2022; Michel et al. 2022]. CLIP guidance has also been exploited for generating rough un-textured shapes [Sanghi et al. 2022, 2023], for optimizing a neural radiance field (NeRF) [Mildenhall et al. 2021] depicting the 3D object [Jain et al. 2022; Lee and Chang 2022; Wang et al. 2022] and for deforming 3D meshes [Gao et al. 2023].

This progress has been further accelerated with the rise of diffusion models, which allow for generating diverse imagery conveying complex visual concepts. DreamFusion [Poole et al. 2022] introduced Score Distillation Sampling (SDS), a method that uses a 2D diffusion model to guide the optimization of a 3D model. SDS was later used in follow up text-to-3D works such as Prolific-Dreamer [Wang et al. 2023b], Score Jacobian Chaining [Wang et al. 2023a], DreamGaussian [Tang et al. 2023] and Magic3D [Lin et al. 2023], as well as image-to-3D techniques such as RealFusion [Melas-Kyriazi et al. 2023] and Magic123 [Qian et al. 2023]. In addition, this generative power has also been leveraged for editing existing 3D objects. Vox-E [Sella et al. 2023] and DreamEditor [Zhuang et al. 2023] have shown that it is possible to locally edit shapes using an SDS loss. LatentNeRF [Metzer et al. 2023] and later Fantasia3D [Chen et al. 2023a] propose a conditional text-to-3D variant, which is also provided with an input 3D shape.

However, these aforementioned works all require timely optimization for each individual sample, and hence they are challenging to apply in practical settings. Several methods have been proposed for texturing 3D meshes using image diffusion models while bypassing SDS [Cao et al. 2023a; Chen et al. 2023b; Richardson et al. 2023]. These methods, however, cannot modify the object's geometry and operate on a texture map representation, and not on the 3D representations directly.

[1] pronounced "spicy".

Methods performing text-guided shape manipulation without the use of pretrained text–image models are significantly less prevalent. Text2Shape [Chen et al. 2019] introduce a dataset tying 15K shapes from ShapeNet [Chang et al. 2015] with textual descriptions, utilized for text-to-3D generation and also later for manipulation [Liu et al. 2022]. ChangeIt3D [Achlioptas et al. 2022] introduce the ShapeTalk dataset, containing textual descriptions discriminating pairs of 3D shapes (also originating from ShapeNet), allowing for manipulating input shapes. LADIS [Huang et al. 2022] propose a disentangled latent representation which better localizes the 3D edits. We demonstrate that our technique allows for outperforming these prior 3D manipulation works, while enabling additional applications which are not necessarily restricted to specific domains.

2.2 Controllable Shape Representations

The problem of creating editable 3D representations has been extensively studied in recent years, not only in the context of text-guided techniques. DualSDF [Hao et al. 2020] represent shapes using two granularity levels, allowing to manipulate high resolution shapes through proxy primitive-based representations. Other works have shown that such primitive-based decompositions can also facilitate tasks such as shape completion [Ganapathi-Subramanian et al. 2018; Sung et al. 2015]. More recently, Tertikas *et al.* [2023] proposed PartNeRF which generates shapes that are an assembly of distinct parts, each parameterized with a neural radiance field. KeypointDeformer [Jakab et al. 2021] discover 3D keypoints, rather than shape primitives, which can be edited for deforming 3D shapes. Several works couple implicit 3D representations with 2D modalities, allowing for editing the 3D shapes from 2D inputs [Cheng et al. 2022; Zheng et al. 2023]. DIF [Deng et al. 2021] represents shapes using a template implicit field shared across a shape category and a 3D deformation field per shape. EXIM [Liu et al. 2023] introduces a hybrid representation composed of an explicit part that enables coarse localization and an implicit part that enables fine global geometric editing and color modifications. SPAGHETTI [Hertz et al. 2022b] propose a shape representation composed of Gaussian Mixture Models which allows for achieving part-level control. SALAD [Koo et al. 2023] later extend this framework to incorporate a diffusion neural network using a cascaded framework.

Several works edit shapes represented as neural fields by propagating edits from selected 2D projections [Liu et al. 2021; Yang et al. 2022]. Neutex [Xiang et al. 2021] represent appearance using 2D texture maps, allowing for editing textures using 2D techniques. Prior works have also shown that implicit neural fields can be coupled with an explicit mesh representation for editing them using as-rigid-as-possible deformations [Garbin et al. 2022; Xu and Harada 2022; Yuan et al. 2022]. Neural Shape Deformation Priors [Tang et al. 2022] predict a neural deformation field given a source mesh and target location of defined handles.

In this work, we propose to manipulate shapes via text-guidance in addition to various structural priors, offering a flexible interface that can operate in various settings. Our approach bears some similarity to SDFusion [Cheng et al. 2023], which enables conditional generation with multiple modalities including text. However, unlike SDFusion which requires training from scratch for each application, our work leverages pretrained text–3D diffusion models, allowing for a quick finetuning of these models without necessarily having access to the data or a vast number of high-end GPUs.

2.3 Conditional Generation with Diffusion Models

Many works are recently seeking new avenues for gaining control over the outputs generated by text-to-image diffusion models [Cao et al. 2023b; Geyer et al. 2023; Hertz et al. 2022a; Patashnik et al. 2023; Tumanyan et al. 2023; Wu et al. 2023]. ControlNet [Zhang et al. 2023] adds conditional control to 2D diffusion models, finetuning models to learn task-specific input conditions. They demonstrate image generation results using various conditions, including Canny edges and user scribbles. Our work is conceptually similar – we modify 3D diffusion models to learn task-specific structural priors.

To achieve structural control over the generation, we manipulate the internal representations of the denoising networks. Prior work have shown that manipulation of these representations, notably the cross-attention and self-attention layers, allows for effective editing of images and videos [Chefer et al. 2023; Geyer et al. 2023; Ruiz et al. 2023]. In particular, several works recently demonstrate that Query features roughly control the structure of the generated images [Cao et al. 2023b; Wu et al. 2023]. Cao *et al.* [2023b] have demonstrated that Query features in the self-attention layers play a pivotal role in modifying the structure of the generated image, showing that non-rigid manipulations can be obtained by querying fixed Keys and Values. Similarly, Wu *et al.* [2023] keep f_K and f_V frozen while finetuning spatio-temporal attention blocks for creating temporally-consistent videos. Inspired by these 2D techniques, our approach carefully mixes Query features belonging to different 3D shapes to learn task-specific structural priors in 3D diffusion models, which are composed of self-attention layers, unlike 2D diffusion models that also contain cross-attention layers.

3 METHOD

In this section, we introduce Spice·E, an approach for incorporating structural priors in pretrained 3D diffusion models. We first review concepts related to the self-attention layers within a transformer-based diffusion model (Section 3.1). We then introduce Cross-Entity Attention, the core component of our approach (Section 3.2). Finally, we describe how to apply it in a transformer-based 3D diffusion model (Section 3.3, Figure 2).

3.1 Preliminaries

We begin by describing the self-attention layers that compose the network blocks within a transformer-based diffusion model. At each timestep t, the noised latent code $\mathbf{z}_t$ is passed as input to the denoising network. For each self-attention layer l, the intermediate features of the network, denoted by $\phi_l(\mathbf{z}_t)$, are first projected to Keys (K), Queries (Q), and Values (V) using learned linear layers f_Q, f_K, f_V. Explicitly stated, $K = f_K(\phi_l(\mathbf{z}_t))$, $Q = f_Q(\phi_l(\mathbf{z}_t))$ and $V = f_V(\phi_l(\mathbf{z}_t))$.

The similarity between the Keys and Queries is initially computed, and then multiplied by the Values. Specifically, the pairwise dot product $Q \cdot K^T$ measures how relevant each key is to the corresponding query. This is then scaled by the square root of the key dimension d, normalised through a softmax function to obtain a

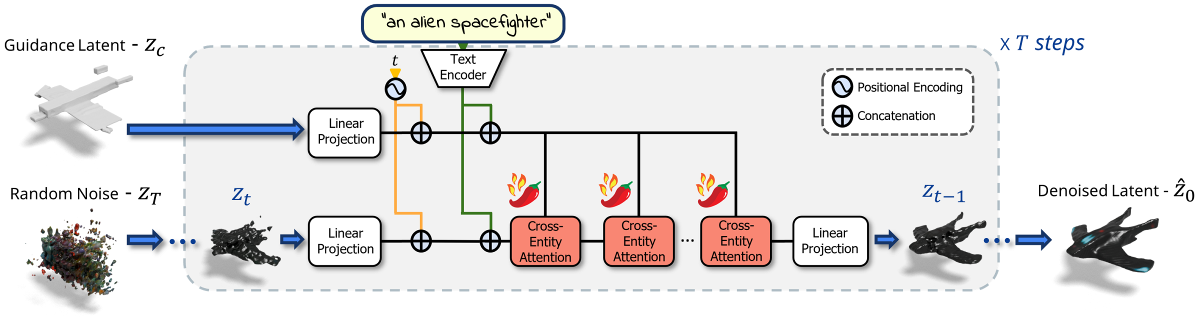

Figure 2: Finetuning 3D diffusion models with Spice·E. We finetune a transformer-based diffusion model [Jun and Nichol 2023], pretrained on a large dataset of text-conditional 3D assets, to enable structural control over the generated 3D shapes. The diffusion model (in gray) is modified to use latent vectors from multiple entities at each step t – a conditional guidance shape X_c encoded into the guidance latent Z_c and a noisy input latent Z_t. The self-attention layers are replaced with our proposed cross-entity attention mechanism. At inference time the fine-tuned diffusion model receives the guidance latent Z_c, random gaussian noise Z_T and a guidance text as input and over T steps gradually denoises the input to produce an output latent $\hat{Z}_0$. The output latent can be decoded into the output shape X_{out}, represented as either a neural radiance field or a signed texture field.

unit vector and finally aggregated to produce the attention function:

$$Attn(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right)V, \tag{1}$$

which is a weighted sum of V, with higher weights for values whose corresponding keys have a larger dot product with the query.

3.2 Cross-Entity Attention

Next we introduce the *Cross-Entity Attention* mechanism, our core technical contribution, illustrated in Figure 3. This mechanism modifies self-attention layers located within transformer-based diffusion models, allowing for latent vectors originating from multiple entities (*i.e.* 3D shapes) to interact. The input to our Cross-Entity Attention block is a pair of latent vectors $(\mathbf{z}, \mathbf{c})$, where $\mathbf{z}$ denotes a noised latent code and $\mathbf{c}$ denotes a conditional latent that encodes structural information we would like to add to the original network. In our setting, the original network is a transformer based diffusion model, pretrained on millions of 3D assets.

As we are interested in preserving the capabilities of the original network, we first apply the *zero-convolution* operator $\mathcal{Z}$ to $\mathbf{c}$. This is a 1 × 1 convolution layer with both weight and bias initialized to zeros, which was recently proposed for adding control to pretrained image diffusion models in ControlNet [Zhang et al. 2023]. Due to its zero initialization, it ensures that the network will not be effected by the conditional latent code when training (or finetuning) begins.

We define the cross-entity attention mechanism over the Queries of the latent vectors, as we are interested in manipulating the structure of the shape encoded within $\mathbf{z}$, while preserving its visual appearance. Formally, the noised latent code $\mathbf{z}$ is projected to $K = f_K(\phi(\mathbf{z}))$, $Q = f_Q(\phi(\mathbf{z}))$ and $V = f_V(\phi(\mathbf{z}))$, denoting $\phi(\mathbf{z})$ as the network's intermediate features. We then perform:

$$Q_\times = f_Q(\phi(\mathbf{z})) + f_{Q_c}(\mathcal{Z}(\phi_c(\mathbf{c}))), \tag{2}$$

where f_{Q_c} and $\phi_c(\mathbf{c})$ are a learned linear layer and intermediate features, initialized randomly.

The output of our cross-entity attention block is the attention function computed over these updated Queries. That is, the output of our proposed block is $\mathbf{z}_{out} = Attn(Q_\times, K, V)$. We allow all block parameters to optimize freely during model finetuning. Intuitively, our attention mechanism acts as a fully-functional self-attention block when finetuning begins. As finetuning progresses, the network gradually learns how to utilize information from the guidance shape at each layer. Note that this is in contrast to a more simple cross-attention mechanism, such as that used in [Loizou et al. 2023], which has no ability to retain a self-attention component.

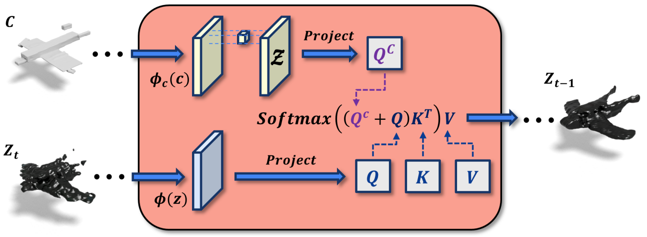

Figure 3: Cross-Entity Attention. Given a pretrained self-attention block, we add a conditional latent c originating from a different entity (*i.e.* 3D shape). Our proposed mechanism mixes the Queries features (after a zero-convolution operator $\mathcal{Z}$ is applied to c), allowing for incorporating structural priors from c.

3.3 Structural Priors in 3D Diffusion Models

In this section, we describe how our cross-entity attention mechanism can be integrated into transformer-based 3D diffusion models to enable structural control over the generated outputs. We use the recently proposed Shap·E [Jun and Nichol 2023] as a reference 3D diffusion model. Shap·E was trained on several million 3D assets, and is capable of generating diverse high-quality 3D objects conditioned on text prompts. For completeness, we briefly describe its architecture, which we modify for creating Spice·E.

Shap·E maps a 3D shape $\mathbf{X}$ to a latent representation $\mathbf{z} \in \mathbb{R}^{d \times d}$ via an encoder E. Specifically, we have $\mathbf{z} = E(\mathbf{X})$, with a latent dimension $d = 1024$. The input $\mathbf{X}$ is composed of both RGB point clouds and RGBA rendered images. The latent $\mathbf{z}$ can be linearly projected into the weights of either a NeRF or a signed texture field (STF) representation via a decoder D. Note that a STF, which is essentially a signed distance field that also provides appearance information, can be represented as a colored mesh, as further detailed in prior work [Gao et al. 2022; Shen et al. 2021]. For text-conditional generation, this latent representation, together with pre-pended tokens representing the CLIP text embedding and the timestep embedding, is fed to a transformer-based diffusion model. The diffusion model is trained following the setup of Ho *et al.* [2020], directly minimizing the error between the original and predicted (de-noised) latent code.

To generate shapes conditioned on structural priors (in addition to text prompts), we modify the system's input to also use a conditional guidance 3D shape $\mathbf{X}_c$. We freeze the encoder E, and fine-tune the pre-trained 3D diffusion generative model (modified as detailed below) on datasets of inputs and guidance shapes that are encoded with E. Each self-attention block is replaced with a cross-entity attention block. To avoid overfitting, we use constant intermediate features ϕ_c for each block, unlike $\phi_l(z)$ which are layer dependent.

During training, given an input latent representation $\mathbf{z}_0$ corresponding to an input 3D asset $\mathbf{X}_{in}$ (*i.e.* $\mathbf{z}_0 = E(\mathbf{X}_{in})$), noise is progressively added to it, producing a noisy latent $\mathbf{z}_t$, where t represents the number of timestamps noise is added. Given $\mathbf{z}_0$, a time step t, a text prompt c_{text} and a latent representation $\mathbf{z}_c$ corresponding to the 3D conditional guidance shape $\mathbf{X}_c$, our model $\mathcal{M}_\theta$ learns to directly predict the denoised input latent representation $\mathbf{z}_0$ by minimizing the same objective used in Shap·E:

$$\mathcal{L} = \mathcal{E}_{\mathbf{z}_0,t,c_{text},\mathbf{c}_0} ||\mathcal{M}_\theta(\mathbf{z}_t, t, c_{text}, \mathbf{z}_c) - \mathbf{z}_0||_2^2 \tag{3}$$

An overview of our training process is shown in Figure 2.

During inference our system is only provided with the guidance shape $\mathbf{X}_c$ encoded into the latent $\mathbf{Z}_c$ with E and a text prompt (c_{text}). We sample from $\mathcal{M}_\theta$, starting at a random noise sample $\mathbf{z}_T$. This sample is gradually denoised into $\hat{\mathbf{z}}_0$, which is then decoded into our 3D output $\mathbf{X}_{out}$, represented as either a NeRF or a STF, using D.

3.4 Optional Refinement

. Our outputs can be refined using an auxiliary unsupervised iterative process that uses 2D diffusion models. Specifically, we can replace the Shap·E initialization in GaussianDreamer [Yi et al. 2023] with Spice·E. GaussianDreamer then proceeds to optimize the Gaussians initialized according to our outputs using Score Distillation, producing more detailed Gaussian splats at the expense of time, specifically increasing generation time from roughly 20 seconds to 15 minutes. See Figures 1 and 5 for results before and after this optional refinement stage. Note that all other reported results are provided without refinement.

4 TASKS

We demonstrate the utility of Spice·E using three text-conditioned 3D-to-3D tasks: semantic shape editing (Section 4.1), text-conditional abstraction-to-3D (Section 4.2), and 3D stylization (Section 4.3).

For each task, we construct a dataset of latent representations and target text-prompts and fine-tune the pretrained 3D diffusion model following the procedure described in the previous section. In other words, we encode a set of input and conditional shapes $\{\mathbf{X}_{in}, \mathbf{X}_c\}$ via E to obtain a set of latent representations $\{\mathbf{z}_0, \mathbf{z}_c\}$ which are used together with their corresponding target text prompts $\{c_{text}\}$ for finetuning. Below, we describe the tasks and provide experimental details, as well as discuss alternative methods and evaluation metrics. Additional details and comparisons, including perceptual studies, are provided in the supplementary material.

4.1 Semantic Shape Editing

4.1.1 Task description. Several works have recently explored the problem of performing semantic fine-grained edits of shapes using language [2022; 2022]. For this task, the target text prompt describes desired semantic modifications to be performed over the input shape. For example, given an input chair, target texts include "the legs are thinner" or "there is a hole in the back".

4.1.2 Experimental details. . For this task, we use the ShapeTalk dataset proposed by Achlioptas *et al.* [2022]. This dataset contains pairs of *distractor* and *target* models (originating from ShapeNet) annotated with a textual annotation describing the shape differences from the distractor shape to the target one. For finetuning models on this task, we use distractor models as conditional guidance shapes $\mathbf{X}_c$ and target models as the input ones $\mathbf{X}_{in}$. We randomly replace 50% of the distractor models with the target ones to further enforce structural similarity to the target models. During inference, only the distractor model and the associated textual description are fed to Spice·E. We follow their setup, finetuning models for the *Table*, *Lamps*, and *Chair* categories and using their train/set splits. We perform additional filtering to these sets to ensure that the distractor and target models are sufficiently close, as we observe that many pairs are geometrically very different. This yields datasets containing approximately 15% of the shapes from the original ShapeTalk dataset (*i.e.* 8K pairs on average for training). See the supplementary material for details.

4.1.3 Alternative Methods. . We compare against ChangeIt3D [2022], which operates over point cloud representations. We use their outputs directly, as these are publicly available. In the supplementary material, we also perform a qualitative comparison with LADIS [2022] over results reported in their paper (as source code or trained models are not available we cannot conduct a quantitative evaluation).

4.1.4 Evaluation metrics. . We follow the evaluation protocol proposed by Achlioptas *et al.* [2022]. Specifically, we use the following metrics:

Linguistic Association Boost (LAB) uses their pretrained listener model for measuring the difference in the predicted association score between the input–output shapes and the target text prompt.

Geometric Difference (GD) uses a standard Chamfer distance to measure the geometric difference between the input and output shapes (scaled by 10^{-2} in comparison to the distances reported in [Achlioptas et al. 2022]), evaluating shape identity preservation.

localized-Geometric Difference (l-GD) uses a part-based segmentation model to only measure geometric differences in regions unrelated to the edit text.

Class Distortion (CD) uses their pretrained shape classifier for measuring the absolute difference of the shape category probability, comparing the input and output shapes.

4.2 Text-conditional Abstraction-to-3D

4.2.1 Task description. . Primitive-based surface reconstruction is a longstanding problem in computer vision and graphics [Gal et al. 2007; Hao et al. 2020; Schnabel et al. 2009]. We explore this problem in the context of our framework. Specifically, given a proxy primitive-based abstract representation and a target text prompt, we are interested in generating a corresponding high-resolution 3D shape that conforms to the target text prompt while maintaining fidelity to the input abstract shape.

4.2.2 Experimental details. . We use 3D models from ShapeNet [Chang et al. 2015] annotated with textual descriptions for this task. Several methods provide means of abstracting shapes of a given category into an assembly of cuboid primitives [Sun et al. 2019; Tulsiani et al. 2017; Yang and Chen 2021]. Therefore, to create corresponding primitive-based shape representations, we utilize the trained *Airplane*, *Chair* and *Table* models given by Yang and Chen [2021]. We also use their splits for constructing train/test datasets.

4.2.3 Alternative Methods. . We compare against SketchShape, the variant from LatentNerf [Metzer et al. 2023] conditioned on coarse shapes, and Fantasia3D [Chen et al. 2023a] which can optionally use a guidance shape. Note that both of these methods are optimization-based, and therefore, are significantly slower at inference time.

4.2.4 Evaluation metrics. . We measure geometric differences (using the GD metric discussed in Section 4.1) between the input primitive-based proxy shape and the output shape to evaluate how well the model enforces the structural priors from the guidance abstract shape. Furthermore, we evaluate to what extent our results are faithful to the edit prompt using the following metrics:

CLIP Similarity (CLIP_{Sim}) measures the similarity between the output objects and the target text prompts, using the cosine-distance between their CLIP embedding.

CLIP Direction Similarity (CLIP_{Dir}), first introduced for evaluating image edits in StyleGAN-NADA [Gal et al. 2021], measures the cosine distance between the direction of the change from the input and output rendered images and the direction of the change from an input prompt to the edit prompt. To evaluate these CLIP based metrics, we render 20 images of both the output and guidance shapes from uniformly-distributed azimuth angles around the 3D object, and average over these angles.

4.3 3D Stylization

4.3.1 Task description. . This task aims at performing text-driven editing of an uncolored 3D asset. Following Michel *et al.* [2022], we define *style* as the object's texturing and fine-grained geometric details.

Table 1: Semantic Shape Editing Evaluation. Below we report performance over the ShapeTalk [Achlioptas et al. 2022] test set (averaging only over highly similar shapes, as discussed in Section 4.1). As illustrated above, our method yields significantly higher LAB scores, suggesting edits that are semantically more accurate, at the expense of slightly higher geometric differences.

Method	LAB↑	GD↓	l-GD ↓	CD↓
ChangeIt3D [Achlioptas et al. 2022]	0.27	**0.003**	**0.009**	**0.05**
Ours	**0.44**	0.007	0.013	**0.05**

4.3.2 Experimental details. . To construct a dataset for this task, we utilize the large-scale Objaverse [Deitke et al. 2023b] dataset. Each model in Objaverse is accompanied by metadata, which includes fields such as name, description, categories, and tags. For our purposes, we need text prompts that describe the object's style and overall appearance. We observed that using the available metadata directly (*e.g.* selecting specific fields) yields highly noisy target prompts. Therefore, we finetune the InstructBLIP [Dai et al. 2023] model to extract target prompts from the object's metadata and associated rendered imagery (see the supplementary for additional details); the model's outputs are used as the text prompts c_{text} for learning 3D stylization, along with the encoded 3D assets $\mathbf{z}_0$ and the uncolored assets $\mathbf{z}_c$. We construct a training dataset containing roughly 7.5K items overall.

4.3.3 Alternative Methods. . We compare against two gradient-based optimization techniques: Latent-Paint, the variant from LatentNerf [Metzer et al. 2023] that operates over 3D meshes directly (only modifying the object's texture), and Fantasia3D [Chen et al. 2023a]. For this task, we compare against two variants of Fantasia3D: One that only performs appearance modeling (henceforth denoted as Fantasia-Paint) and the full model, which also modifies the object's geometry. In the supplementary material, we also compare against Vox-E [Sella et al. 2023], a recent optimization-based method proposed for performing text-guided editing of 3D objects.

4.3.4 Evaluation metrics. . For this task, we use the same evaluation metrics discussed above in Section 4.2: CLIP_{Sim}, CLIP_{Dir} and GD, to evaluate both the fidelity to the edit and the guidance shape.

5 EXPERIMENTS

We present the results and comparisons for the tasks described above in Section 5.1. We then ablate the design choices for the cross-entity attention block in Section 5.2. Finally, we discuss limitations in Section 5.3. Additional results, comparisons and ablations can be found in the supplementary material.

5.1 Evaluation

5.1.1 Semantic Shape Editing. . Results for the semantic shape editing task are reported in Table 1. As illustrated in the table, our

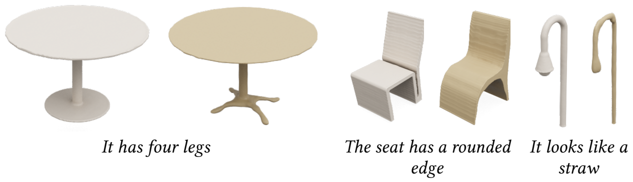

Figure 4: Semantic shape editing results are shown above (input guidance shape on the left and edited outputs on the right, shown in different colors for visualization purposes). As illustrated in the figure, our method can semantically edit input shapes according to target prompts, while preserving the shape's structure.

Table 2: Text-conditional Abstraction-to-3D Evaluation. Below we compare the performance of SketchShape [Metzer et al. 2023] and Fantasia3D [Chen et al. 2023a] against ours over the primitive-based shape conditioning task. As illustrated, our method can more faithfully preserve the input structure, while exhibiting significantly faster inference time. GD is not computed for SketchShape as it outputs a NeRF representation.

Method	CLIP_{Sim} ↑	CLIP_{Dir} ↑	GD ↓	Run Time
SketchShape	0.27	0.01	—	~ 15 minutes
Fantasia3D	0.27	0.01	0.06	~ 30 minutes
Ours	**0.28**	**0.03**	**0.01**	**~ 20 seconds**

edits better reflect the target text prompts, yielding an average LAB score of 0.44 versus 0.27 for ChangeIt3D. Both methods are capable of generating objects resembling their respective object categories, as illustrated by the low class distortion values. Our method yields slightly higher GD and l-GD scores. Generally, we observe that the outputs generated by ChangeIt3D often do not deviate significantly from the inputs (which is also consistent with the lower LAB scores). This is further illustrated in Figures 10 and 4.

5.1.2 Text-conditional Abstraction-to-3D. . Results for the text-conditional abstraction-to-3D task are reported in Table 2. As shown in the table, our generated 3D shapes can more faithfully preserve the abstract input guidance shapes, yielding better GD scores compared to Fantasia3D (GD is not computed for SketchShape as it outputs a NeRF representation). While Fantasia3D was not trained with any geometric supervision (thus explaining this lower GD score), we believe this metric is important in emphasizing that prior work are not suitable for this task. Note that our method also maintains high fidelity to the text prompts, outperforming both methods over CLIP_{dir} while achieving comparable CLIP_{sim} scores, all while exhibiting significantly faster inference times. See Figure 5 for a qualitative comparison, and additional results in Figure 7.

5.1.3 3D Stylization. . Results for the 3D stylization task are reported in Table 3. As illustrated, our edits capture the target text prompt well, yielding results comparable with Latent-Paint and Fantasia3D, while being orders of magnitude faster. Additional results can be seen in Figure 8.

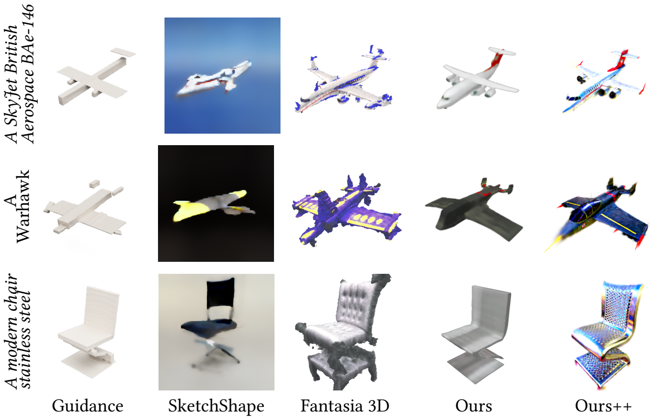
A SkyJet British Aerospace BAe-146
A Warhawk
A modern chair stainless steel
Guidance
SketchShape
Fantasia 3D
Ours
Ours++

Figure 5: Text-conditional Abstraction-to-3D Comparison. We compare to the results obtained using SketchShape [Metzer et al. 2023] and Fantasia3D [Chen et al. 2023a]. Methods are provided with a proxy cuboid-based abstract shape with a target prompt (left). As illustrated in the figure, our results better preserve the structure of the abstract guidance shape, while conveying the target text prompt. In the rightmost column (denoted as "Ours++"), we present results obtained after the optional refinement stage.

Table 3: 3D Stylization Evaluation. We compare against Latent-Paint [Metzer et al. 2023] and two versions of Fantasia3D [Chen et al. 2023a] (with and without geometry modeling) over the 3D stylization task. As illustrated below, our edits are comparable with prior work and can be achieved orders of magnitude faster. GD is computed only for methods that can modify the geometry of the shape.

Method	CLIP_{Sim} ↑	CLIP_{Dir} ↑	GD ↓	Run Time
Latent-Paint	0.27	**0.01**	—	~ 15 minutes
Fantasia3D-Paint	**0.28**	**0.01**	—	~ 15 minutes
Fantasia3D	**0.28**	**0.01**	0.06	~ 30 minutes
Ours	0.27	**0.01**	**0.01**	**~ 20 seconds**

5.1.4 Additional Experiments. . To better illustrate what differences in CLIP-based metrics mean, we perform two additional experiments: (i) *No Operation* baseline, measuring the CLIP_{Sim} of the guidance shape to the target text, and (ii) *Oracle*, measuring CLIP_{Sim} on the ground-truth shape and CLIP_{Dir} in the direction pointing from the guidance shape to the ground truth shape. These provide a lower and upper bound over these metrics in our setting.

For both the text-conditional abstraction-to-3D and the 3D stylization tasks, the *No Operation* baseline produces lower CLIP_{Sim} scores of 0.25 and 0.24 (for abstraction-to-3D and 3D stylization, respectively) and a CLIP_{Dir} of 0.0, while the *Oracle* produces CLIP_{Sim} scores of 0.28 (for both tasks) and CLIP_{Dir} scores of 0.02 and 0.05 (for abstraction-to-3D and 3D stylization, respectively). Indeed, for

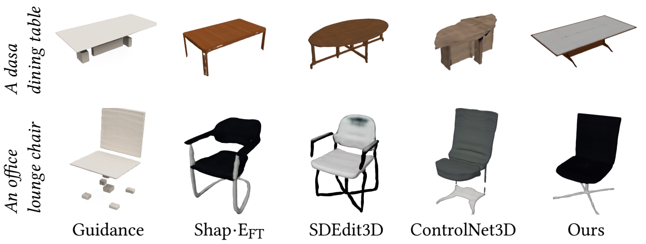

Figure 6: Qualitative ablation results, obtained for test shapes from the text-conditional abstraction-to-3D task. We compare our cross-entity attention mechanism (right) with several baselines, detailed in Section 5.2. As illustrated above, our approach allows for generating 3D shapes that conform to the guidance structure significantly better than baseline methods, while remaining faithful to the target text prompt.

both tasks, the performance of our method, as well as competing methods, all fall in the range of the upper and lower bounds given by these baselines, with the CLIP$_{Dir}$ metric suggesting room for further improvement by future work.

5.2 Ablations

Next we ablate our cross-entity attention mechanism, demonstrating that comparable structural control cannot be achieved with baseline methods. We compare to the following baselines: (i) Shap·E$_{\text{FT}}$, the original Shap·E model finetuned on each dataset with text guidance only (no structural guidance is added). (ii) SDEdit3D, inspired by the image editing technique SDEdit [Meng et al. 2021], which uses the Shap·E$_{\text{FT}}$ models. During inference, noise is added to the conditional latent, and it is denoised with the target textual prompt. (iii) CrossOnly, an ablated version of our framework that uses Cross-Attention instead of our Cross-Entity Attention mechanism, i.e. using only the conditional queries. (iv) ControlNet3D, inspired by the network architecture used in ControlNet [Zhang et al. 2023], which freezes and clones the original network blocks of Shap·E, creating a frozen and trainable copy of it. The guidance shape is passed through the trainable copy with intermediate outputs added to the appropriate frozen copy blocks as residuals through a zero-convolution (see the supplementary material for more details).

We conduct experiments over the text-conditional abstraction-to-3D task. As illustrated in Figure 6, these baselines methods cannot faithfully preserve the conditional guidance shape. For instance, the ControlNet3D results are of significantly lower quality in comparison to our method. We attribute this visual gap to the much larger number of parameters that need to be optimized in comparison to our method (50M vs. 330M additional parameters), making this method more prone to overfitting on the relatively small datasets we use (i.e. resulting in the model forgetting its pretrained knowledge). Quantitatively, the baselines yield significantly worse GD scores: 0.06 (Shap·E$_{\text{FT}}$), 0.05 (SDEdit3D), 0.03 (CrossOnly) and 0.03 (ControlNet3D), compared to 0.01 for our approach, further showing that their outputs strongly deviate from the guidance shapes.

In the supplementary material, we also conduct additional perceptual studies to evaluate user's preference of our results over the ControlNet3D baseline. We also conduct additional ablations to motivate our design choices. In particular, we modify our cross-entity attention mechanism in various ways, including removing the zero-convolution operators and performing cross-attention over the Keys or Values. These ablations demonstrate that our proposed cross-entity mechanism allows for better preserving the structure of the guidance shape in comparison to other possible modifications.

5.3 Limitations

Our method allows for learning various types of structural priors for generating text-conditional shapes guided by 3D inputs, but there are several limitations to consider, as also shown in Figure 9. First, our approach inherits limitations from diffusion-based techniques and in particular from Shap·E, which we build our method upon. While Shap·E can generate diverse high-quality 3D shapes, it still struggles to bind multiple attributes to objects, limiting the scope of possible object edits. Furthermore, we observe that highly complicated shapes are not often successfully encoded, leading to noisy data used for both training and evaluation. As our approach can be added on top of other transformer-based 3D diffusion models, we expect that with the emergence of stronger backbones, more powerful edits can be achieved.

Additionally, our approach does not offer explicit control over the tradeoff between the fidelity to the input guidance shape and the consistency with the target prompt. This often leads to results which are either not functionally plausible (for instance, see the ping pong table on the top row of Figure 7 where the table's legs make it challenging for the table to correctly function as intended) or conversely do not sufficiently preserve the guidance structure.

6 CONCLUSION

In this work, we presented Spice·E, a new approach for adding structural control to 3D diffusion models. We demonstrated that our method facilitates several text-conditional 3D editing tasks, without the need for tailoring the network architectures or training objectives. Our work represents a step towards the goal of democratizing 3D generation, making 3D object editing more accessible to non-experts by providing them with task-specific structural control within seconds. Technically, we introduced the cross-entity attention mechanism, which allows for mixing latent representations corresponding to different 3D shapes while preserving the capabilities of the pretrained 3D diffusion model. We believe that our mechanism could potentially improve a wide variety of applications where guidance is injected into a generative framework, beyond the realm of 3D shape generation and manipulation.

ACKNOWLEDGMENTS

We thank Peter Hedman, Ron Mokadi, Daniel Garibi, Itai Lang and Or Patashnik for helpful discussions. This work was supported by the Israel Science Foundation (grant no. 2510/23) and by the Alon Scholarship.

REFERENCES

Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. 2022. ChangeIt3D: Language-Assisted 3D Shape Edits and Deformations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023b. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. *arXiv preprint arXiv:2304.08465* (2023).

Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. 2023a. TexFusion: Synthesizing 3D Textures with Text-Guided Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4169–4181.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An Information-Rich 3D Model Repository. *arXiv preprint arXiv:1512.03012* (2015).

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023b. Text2Tex: Text-Driven Texture Synthesis via Diffusion Models. *arXiv preprint arXiv:2303.11396* (2023).

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. Text2Shape: Generating shapes From Natural Language By Learning Joint Embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 100–116.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-Quality Text-to-3D Content Creation. *arXiv preprint arXiv:2303.13873* (2023).

Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. 2022. Tango: Text-Driven Photorealistic and Robust 3D Stylization via Lighting Decomposition. *arXiv preprint arXiv:2210.11277* (2022).

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. 2023. SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 4456–4465.

Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. 2022. Cross-Modal 3D Shape Generation and Manipulation. In *European Conference on Computer Vision*. Springer, 303–321.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2023a. Objaverse-xl: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023b. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 13142–13153.

Yu Deng, Jiaolong Yang, and Xin Tong. 2021. Deformed Implicit Field: Modeling 3D Shapes with Learned Dense Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 10286–10296.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. Stylegan-nada: Clip-guided Domain Adaptation of Image Generators. *arXiv preprint arXiv:2108.00946* (2021).

Ran Gal, Ariel Shamir, Tal Hassner, Mark Pauly, and Daniel Cohen-Or. 2007. Surface Reconstruction Using Local Shape Priors. In *Symposium on Geometry Processing.* 253–262.

Vignesh Ganapathi-Subramanian, Olga Diamanti, Soeren Pirk, Chengcheng Tang, Matthias Niessner, and Leonidas Guibas. 2018. Parsing Geometry using Structure-Aware Shape Templates. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 672–681.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images. *Advances In Neural Information Processing Systems* 35 (2022), 31841–31854.

William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. 2023. TextDeformer: Geometry Manipulation using Text Guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. 2022. VolTeMorph: Realtime, Controllable and Generalisable Animation of Volumetric Representations. *arXiv preprint arXiv:2208.00949* (2022).

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arXiv:2307.10373* (2023).

Zekun Hao, Hadar Averbuch-Elor, Noah Snavely, and Serge Belongie. 2020. Dualsdf: Semantic Shape Manipulation Using a Two-Level Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7631–7641.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022a. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626* (2022).

Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. 2022b. Spaghetti: Editing Implicit Shapes Through Part Aware Generation. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–20.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Ian Huang, Panos Achlioptas, Tianyi Zhang, Sergey Tulyakov, Minhyuk Sung, and Leonidas Guibas. 2022. LADIS: Language Disentanglement for 3D Shape Editing. *arXiv preprint arXiv:2212.05011* (2022).

Takeo Igarashi, Tomer Moscovich, and John F Hughes. 2005. As-Rigid-as-Possible Shape Manipulation. *ACM transactions on Graphics (TOG)* 24, 3 (2005), 1134–1141.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 867–876.

Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. 2021. KeyPointDeformer: Unsupervised 3D Keypoint Discovery for Shape Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12783–12792.

Heewoo Jun and Alex Nichol. 2023. Shap-E: Generating Conditional 3D Implicit Functions. *arXiv preprint arXiv:2305.02463* (2023).

Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, and Minhyuk Sung. 2023. Salad: Part-level Latent Diffusion for 3D Shape Generation and Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 14441–14451.

Han-Hung Lee and Angel X Chang. 2022. Understanding Pure Clip Guidance for Voxel Grid NeRF Models. *arXiv preprint arXiv:2209.15172* (2022).

John P Lewis, Matt Cordner, and Nickson Fong. 2023. Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 811–818.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-resolution Text-To-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 300–309.

Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. 2021. Editing Conditional Radiance Fields. In *Proceedings of the IEEE/CVF international conference on computer vision.* 5773–5783.

Zhengzhe Liu, Jingyu Hu, Ka-Hei Hui, Xiaojuan Qi, Daniel Cohen-Or, and Chi-Wing Fu. 2023. EXIM: A Hybrid Explicit-Implicit Representation for Text-Guided 3D Shape Generation. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–12.

Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. 2022. Towards Implicit Text-Guided 3D Shape Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 17896–17906.

Marios Loizou, Siddhant Garg, Dmitry Petrov, Melinos Averkiou, and Evangelos Kalogerakis. 2023. Cross-Shape Attention for Part Segmentation of 3D Point Clouds. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, e14909.

Thalmann Magnenat, Richard Laperrière, and Daniel Thalmann. 1988. *Joint-dependent Local Deformations for Hand Animation and Object Grasping*. Technical Report. Canadian Inf. Process. Soc.

Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg Reconstruction of any Object from a Single Image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 8446–8455.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *arXiv preprint arXiv:2108.01073* (2021).

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12663–12673.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2022. Text2Mesh: Text-Driven Neural Stylization for Meshes. In *CVPR*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv preprint arXiv:2212.08751* (2022).

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-Level Shape Variations with Text-to-Image Diffusion Models. *ICCV* (2023).

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv preprint arXiv:2209.14988* (2022).

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. 2023. Magic123: One Image to High-Quality 3D Object Generation using both 2D and 3D Diffusion Priors. *arXiv preprint arXiv:2306.17843* (2023).

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning.*

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. *arXiv preprint arXiv:2302.01721* (2023).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. Clip-Forge: Towards Zero-Shot Text-to-Shape Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18603–18613.

Aditya Sanghi, Rao Fu, Vivian Liu, Karl DD Willis, Hooman Shayani, Amir H Khasahmadi, Srinath Sridhar, and Daniel Ritchie. 2023. CLIP-Sculptor: Zero-Shot Generation of High-Fidelity and Diverse Shapes From Natural Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18339–18348.

Ruwen Schnabel, Patrick Degener, and Reinhard Klein. 2009. Completion and Reconstruction with Primitive Shapes. In *Computer Graphics Forum*, Vol. 28. Wiley Online Library, 503–512.

Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. 2023. Vox-E: Text-guided Voxel Editing of 3D Objects. *arXiv preprint arXiv:2303.12048* (2023).

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep Marching Tetrahedra: a Hybrid Representation for High-Resolution 3D Shape Synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.

Chun-Yu Sun, Qian-Fang Zou, Xin Tong, and Yang Liu. 2019. Learning Adaptive Hierarchical Cuboid Abstractions of 3D Shape Collections. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

Minhyuk Sung, Vladimir G Kim, Roland Angst, and Leonidas Guibas. 2015. Data-Driven Structural Priors for Shape Completion. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–11.

Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. 2022. Neural Shape Deformation Priors. *Advances in Neural Information Processing Systems* 35 (2022), 17117–17132.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653* (2023).

Konstantinos Tertikas, Despoina Paschalidou, Boxiao Pan, Jeong Joon Park, Mikaela Angelina Uy, Ioannis Emiris, Yannis Avrithis, and Leonidas Guibas. 2023. Generating Part-Aware Editable 3D Shapes without 3D Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4466–4478.

Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. 2017. Learning Shape Abstractions by Assembling Volumetric Primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2635–2643.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. Clip-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3835–3844.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023a. Score jacobian chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12619–12629.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213* (2023).

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-Video: One-shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.

Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. 2021. Neutex: Neural Texture Mapping for Volumetric Neural Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7119–7128.

Tianhan Xu and Tatsuya Harada. 2022. Deforming Radiance Fields with Cages. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 159–175.

Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. 2022. Neumesh: Learning Disentangled Neural Mesh-Based Implicit Field for Geometry and Texture Editing. In *European Conference on Computer Vision*. Springer, 597–614.

Kaizhi Yang and Xuejin Chen. 2021. Unsupervised Learning for Cuboid Shape Abstraction via Joint Segmentation from Point Clouds. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. GaussianDreamer: Fast Generation from Text to 3D Gaussian Splatting with Point Cloud Priors. *arXiv preprint arXiv:2310.08529* (2023).

Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. 2022. NeRF-editing: Geometry Editing of Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18353–18364.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-To-Image Diffusion Mdels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. 2023. Locally Attentional SDF Diffusion for Controllable 3D Shape Generation. *arXiv preprint arXiv:2305.04461* (2023).

Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. 2023. DreamEditor: Text-Driven 3D Scene Editing with Neural Fields. *arXiv preprint arXiv:2306.13455* (2023).

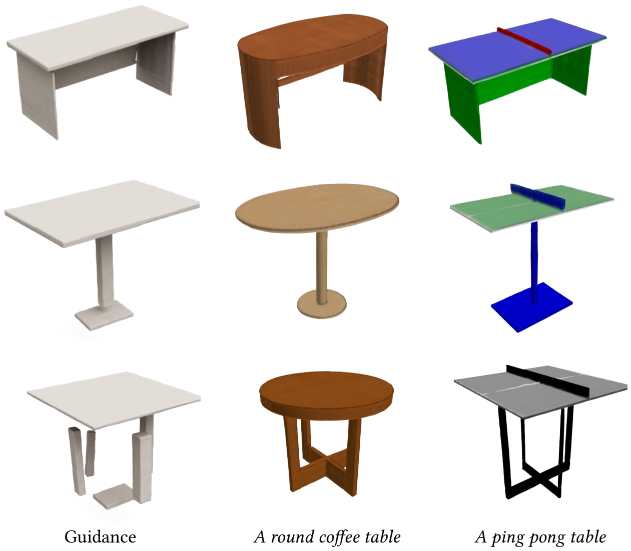

Figure 7: Text-conditional abstraction-to-3D results for test shapes from the *Table* category. The leftmost column displays the guidance input — a proxy cuboid-based shape. The remaining columns showcase our results over two different target text prompts.

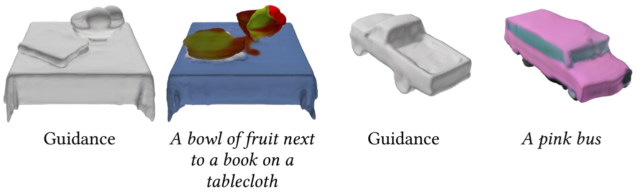

Figure 9: Limitations. Above, we present two failure cases. These likely result from incorrect multiple attribute binding (the fruit bowl and the book colored similarly) or insufficient preservation of the guidance structure (changing the guidance pickup truck into a bus and switching the back of the guidance truck to the front of the bus).

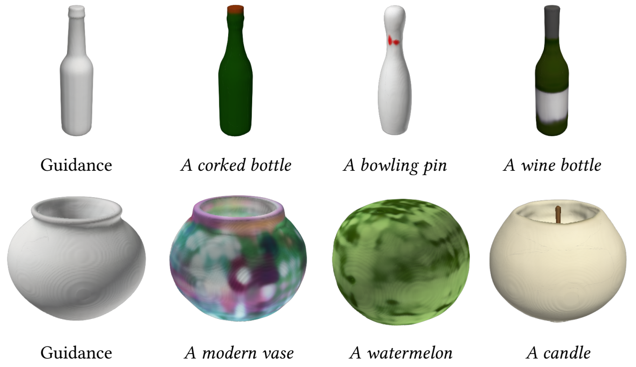

Figure 8: 3D stylization results results are shown above. The leftmost column displays the guidance input — an uncolored 3D asset. The remaining columns showcase how the guidance input is styled according to the target text prompt.

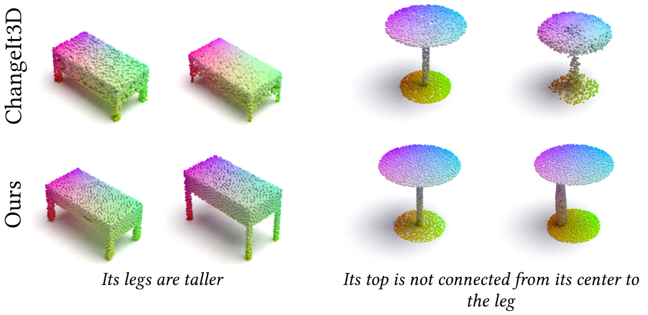

Figure 10: Semantic Shape Editing Comparison. We compare to prior work performing semantic shape editing above. As ChangeIt3D [Achlioptas et al. 2022] operates over a point cloud representation, we show input point clouds on the left and edited point clouds on the right. For our results, we visualize the point clouds after shape encoding, hence our inputs are not identical to theirs. As illustrated in the figure, our method can perform more significant edits, yielding edited shapes that better reflect the target prompts.