

# Directional Texture Editing for 3D Models

Shengqi Liu<sup>1\*</sup> Zhuo Chen<sup>1\*</sup> Jingnan Gao<sup>1</sup> Yichao Yan<sup>1†</sup> Wenhan Zhu<sup>1</sup>  
Ke Gao<sup>2</sup> Jiangjing Lyu<sup>2</sup> Xiaokang Yang<sup>1</sup>  
<sup>1</sup> Shanghai Jiao Tong University <sup>2</sup> Alibaba Group

{lsqlsq, ningci5252, gjn0310, yanyichao, zhuwenhan823, xkyang}@sjtu.edu.cn

{gaoke.gao, jiangjing.ljj}@alibaba-inc.com

## Abstract

Texture editing is a crucial task in 3D modeling that allows users to automatically manipulate the surface materials of 3D models. However, the inherent complexity of 3D models and the ambiguous text description lead to the challenge in this task. To address this challenge, we propose ITEM3D, a **Texture Editing Model** designed for automatic 3D object editing according to the text Instructions. Leveraging the diffusion models and the differentiable rendering, ITEM3D takes the rendered images as the bridge of text and 3D representation, and further optimizes the disentangled texture and environment map. Previous methods adopted the absolute editing direction namely score distillation sampling (SDS) as the optimization objective, which unfortunately results in the noisy appearance and text inconsistency. To solve the problem caused by the ambiguous text, we introduce a relative editing direction, an optimization objective defined by the noise difference between the source and target texts, to release the semantic ambiguity between the texts and images. Additionally, we gradually adjust the direction during optimization to further address the unexpected deviation in the texture domain. Qualitative and quantitative experiments show that our ITEM3D outperforms the state-of-the-art methods on various 3D objects. We also perform text-guided relighting to show explicit control over lighting.

## 1. Introduction

Texture editing is an important task in 3D modeling that involves manipulating the surface properties of 3D models to create a visually appealing appearance according to the user’s ideas. With the increasing applications of 3D models in entertainment and e-shopping, how to automatically generate and edit the texture of a 3D model without manual

\*Equal contribution.

†Corresponding authors

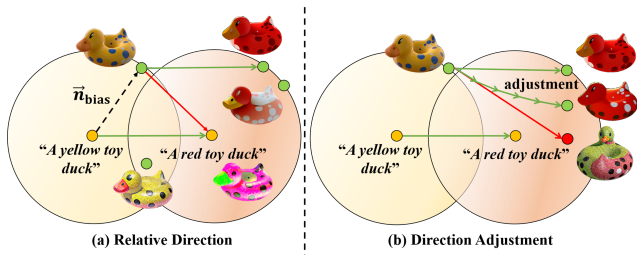


Figure 1. **Motivation.** (a) Previous methods [10, 38] with SDS loss to directly guide the optimization leads to ambiguous details due to the bias between texts and images (red line), while our method introduces the relative direction between source and target texts to the optimization process, eliminating the bias and improving the rendering results (green line). (b) The optimization in the texture domain gives rise to the deviation of the target direction (red line), thus we gradually adjust the direction to fine-tune the optimization (fold green line).

effort becomes an appealing task in the field of 3D vision. However, this task is challenging due to the complexity of 3D models and the special representation of the texture.

Recent advances have demonstrated the effectiveness of generative models in synthesizing high-quality appearances that are both visually pleasing and semantically meaningful. The use of generative adversarial networks (GANs) [1, 11, 47, 56] has shown promising results in producing textures with intricate patterns and complex structures. Other approaches, such as texture synthesis via direct optimization [12, 13, 44, 63] or neural style transfer [2, 18, 31, 54], have also been explored for their ability to generate textures with specific artistic styles. However, the capacity of these models is still unable to meet the need of real-world applications, which requires high-quality and diverse textures. Meanwhile, recent researches [10, 24, 27, 28, 30, 32, 42] on the diffusion models have emerged as a powerful new family of generative methods, which achieve impressive generation results in natural images and videos, inspiring us to introduce the awesome power into the task of 3D modeling.

However, directly applying the diffusion model to 3D objects is a non-trivial task due to the following reasons. 1) **The gap between the 3D representation and natural images.** Existing diffusion models are typically trained with natural images, making the pre-trained diffusion model lack prior knowledge in the 3D domain. Moreover, due to the complexity of the 3D model, it would be difficult to jointly edit shape, appearance, and shading, sometimes leading to conflicts in optimization goals. Therefore, directly editing the 3D representation may cause extreme semantic bias and sacrifice inherent 3D topology. 2) **The learning misdirection of text description.** It is hard for text prompts to exactly describe the target images at the pixel level, leading to an ambiguous direction when taking the rendered images as the bridge.

To solve these problems, we present an efficient model, dubbed **ITEM3D**, which can generate visually natural texture corresponding to the text prompt provided by users. Instead of directly applying the diffusion model to texture editing in the 2D space, we adopt rendered images as the intermediary that bridges the text prompts and the appearance of 3D models. Apart from the appearance, the lighting and shading are also key components influencing the rendering results. Therefore, we represent the 3D model into a triangular mesh, a set of disentangled materials consisting of the texture, and an environment map following *nvdiffrac* [36], which accomplishes the disentangled modeling of both appearance and shading.

To optimize a texture and an environment map with the diffusion model, a naive idea is to adopt the score distillation sampling (SDS) [38] like previous diffusion-based editing methods, which represents the absolute direction. Unfortunately, the utilization of absolute direction, such as SDS, often leads to noisy details and an inconsistent appearance, due to the ambiguous description of the text prompt for the target images. Inspired by the recent improvement [19], we replace the absolute editing direction led by the score distillation sampling with a relative editing direction determined by two predicted noises under the condition of the source text and the target text respectively, as illustrated in Fig. 1 (a). In this way, our model enables us to edit the texture in obedience to the text while bypassing the inconsistency problem by releasing the ambiguous description. It is ideal that the intermediate states between the source and target text can give relatively accurate descriptions for arbitrary rendered images during the optimization, like the green straight lines in Fig. 1 (b). However, the optimization in the texture domain actually shows the nonlinear distortion of the appearance in rendered images, leading to the deviation from the determined direction, like the red line in Fig. 1 (b). To reduce the deviation caused by the texture projection, we gradually adjust the editing direction during the optimization, as green fold lines shown in Fig. 1 (b).

With the advent of the textural-inversion model, it can be easy to automatically correct the description as the change of the texture and its rendered images.

Thanks to the proposed solutions, our method overcomes the challenges of domain gap and learning misdirection, fulfilling the requirements of texture editing. In summary, our contributions are:

- We design an efficient optimization pipeline to edit the texture and environment map obedient to the text instruction, directly empowering the downstream application in the industrial pipeline.
- We introduce the relative direction to the optimization of the textured 3D model, releasing the problem of noisy details and inconsistent appearance caused by the semantic ambiguity between the texts and images.
- We propose to gradually adjust the relative direction guided by the source and target texts which addresses the unexpected deviation from the determined direction caused by the nonlinear texture projection.

## 2. Related Work

### 2.1. 3D Model Representation

**Neural Radiance Field.** Neural implicit representations employ differentiable rendering techniques to reconstruct 3D geometry along with its corresponding appearance from multi-view images. Neural Radiance Field (NeRF) [34] and followup approaches [7, 29, 35, 41, 53, 55, 57, 60] utilize volumetric representations and a neural encoder to compute radiance field. These methods explicitly represent the specular reflection as the appearance embedding on the surface, enabling the high-fidelity rendering results. While these approaches are capable of capturing complex view-dependent appearances, they often suffer from poor surface reconstruction and limited relighting capabilities. To maintain the geometry quality, surface-based methods [14, 37, 52] introduce an implicit surface to the volumetric representation, resulting in improved geometry compared to pure NeRF-based methods. Nevertheless, these representations require excessive computation resources and still face challenges in decoupling material components, lack of editability.

**Neural Inverse Rendering.** To solve the aforementioned problem, another category of methods focuses on the decomposition of rendering parameters. These methods introduce inverse rendering to estimate surface, materials, and lighting conditions simultaneously, achieving material editing and scene relighting. NeRV [46] and NeRD [3] decompose the scene into the shape, reflectance and illumination components, enabling rendering under varying lighting conditions. NeRFactor [62] introduces a two-stage method to separately reconstruct geometry and reflectance, simplifying reflectance optimization. *Nvdiffrac* [36] additionally introduces a differentiable model of environment lighting to

efficiently recover all-frequency lighting. Considering the balance between appearance and editability, our proposed method, ITEM3D, leverages this decomposed representation to enable efficient and natural texture editing while preserving the topology of objects.

## 2.2. 3D Text-guided Generation

**CLIP-based Generation.** With the advent of large text-image models, CLIP, recent works [21, 22, 33, 51] have led to impressive progress in 3D text-driven synthesis. The majority of methods adopt optimization procedures supervised by the CLIP similarity [39]. Specifically, CLIP-NeRF [51] proposes a unified framework to manipulate NeRF, guided by either a text prompt or an example image. Similarly, CLIP-Mesh utilizes an explicit textured mesh as a 3D representation, allowing for shape deformation along with corresponding texture modifications based on the input text. As the 3D-aware GAN gains popular, several works [11, 16, 25] focus on the 3D stylization by fine-tuning the 3D GAN under the guidance of the CLIP model. However, the main drawback of CLIP-guided generation is the lack of diversity due to the deterministic embedding of the target prompt.

**Diffusion-based Generation.** Apart from the CLIP-based method, the diffusion model [20] has recently inspired huge breakthroughs in 3D text-guided generation. These methods can be briefly classified into two categories, *i.e.*, 3D large generative models [9, 28, 45] and optimization for specific 3D representations [6, 10, 17, 32, 38, 49, 59]. Large models rely on large-scale 3D data to train a diffusion model aware to 3D objects. One of these related methods, Point-UV diffusion [58], proposes a two-stage architecture containing a point diffusion and a UV diffusion to generate textures for 3D meshes. Despite the fast inference, these methods faces the challenges of low quality and poor generalization for the complex objects, due to the amount and diversity of current 3D datasets.

Considering the limitation of data, more researchers focus on the optimization of 3D representation for the specific objects. Dreamfusion [38] initially proposes the SDS loss as a means to incorporate the 2D prior of Stable Diffusion [43] into the generation of 3D domains. Following Dreamfusion, several works [6, 32, 49, 59] incorporate the SDS loss to directly optimize NeRF-like representations. Different from the NeRF representations, DreamGaussian [48] adopts the recently advent Gaussian splatting as the 3D representation and introduces the 3D pre-trained diffusion model [28] to provide the additional geometry prior. Fantasia3D [10] leverages the disentangled modeling, and learns the geometry and appearance supervised by the score distillation sampling. However, these SDS-based methods often produce non-detailed and blurry outputs due to noisy gradients. In contrast, our ITEM3D utilizes the relative direction

to eliminate the semantic ambiguity of the target prompt towards the texture. Recently, several works [8, 17, 40] utilize 2D diffusion model [4, 61] to generate the initial image guidance and alternatively refine the 3D model. Instruct-NeRF2NeRF [17] introduces the NeRF optimization with training images that are gradually updated by the Instruct-Pix2Pix [4]. While their model can achieve high-quality rendering results, the training process is too lengthy as they need an updating strategy to maintain the 3D consistency of NeRF representation during the editing process. Similarly, Text2Tex [8] initializes one-view appearance for the given mesh by the Controlnet [61] and progressively generates the novel-view texture while refining the existed texture. In contrast, the decomposed 3D representation of our ITEM3D empowers us to directly optimize the disentangled materials, texture map and environment map, which significantly improves the training efficiency.

## 3. Method

### 3.1. Overview

Given a set of multi-view images  $\mathcal{I} = \{I_1, \dots, I_n\}$ , we first reconstruct the 3D model with both geometry and texture, and then edit the texture under the guidance of text prompts. To this end, we design a zero-shot differentiable framework that optimizes the disentangled materials of the object, *i.e.*, texture map and environment map. We leverage a differentiable rendering model to represent the 3D model as an accurate shape and surface materials with texture and environment map (Sec. 3.2). For further editing of appearance, we utilize the diffusion model to guide the direction of the texture optimization given the target text prompt. To solve the problem of ambiguous and noisy details, we introduce the relative direction of source texts and target texts into the optimization (Sec. 3.3). Moreover, we gradually adjust relative direction to address the challenges of deviation caused by the unbalanced optimization in the texture domain (Sec. 3.4). The overview of our method is demonstrated in Fig. 2.

### 3.2. 3D Model Representation

To accomplish disentangled editing, we decompose the 3D model into a triangular mesh, a set of spatially varying materials and an environment map. The material model we employ, denoted as  $(k_d, k_{orm}, n)$ , combines a diffuse term  $k_d$ , a specular term  $k_{orm}$ , and a normal term  $n$  following the physically-based (PBR) material model introduced by Disney [5]. This PBR material model can be seamlessly integrated into existing industry rendering engines without any change.

Additionally, for efficient representation of volumetric textures, we leverage a multi-layer perceptron (MLP) to encode all material parameters into a compact representation.

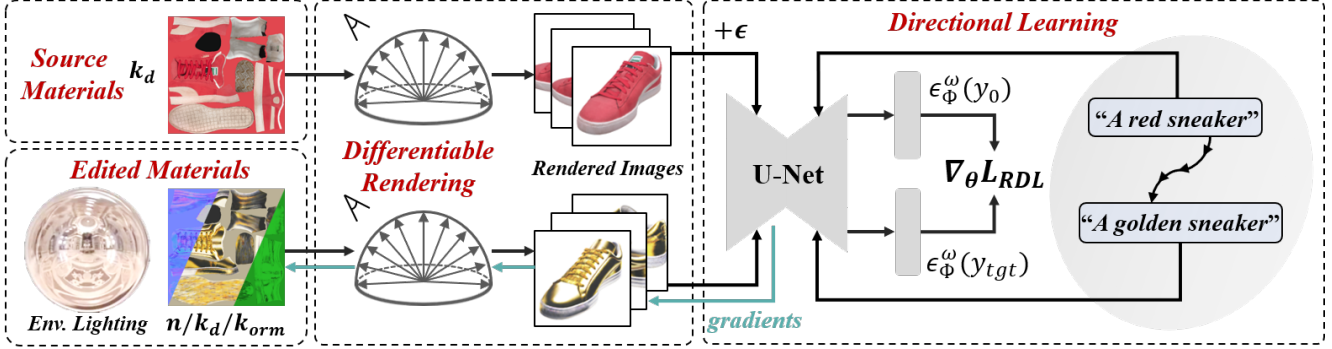


Figure 2. **Pipeline** of 3D model texture editing. We render the 3D model with mesh, texture, and environment map into 2D images which are then added with noise  $\epsilon$ . We further separately use the target text and the gradually adjusted source text as the conditions to predict the added noise via the U-Net. The difference between the two predicted noises serve as the relative direction to guide the optimization of the materials and environment map.

Specifically, given a world space position  $x$ , we compute the MLP texture, which includes the base color,  $k_d$ , the specular parameters  $k_{orm}$ , and a tangent space normal perturbation  $n$ . This mapping is formulated as  $x \rightarrow (k_d, k_{orm}, n)$ . With the introduction of neural texture representation, the textures are initialized by sampling the MLP on the mesh surface and then optimized efficiently, while maintaining a fixed topology.

Following the rendering equation of the image-based lighting model, we integrate all the incoming radiance  $L_i(\omega_i)$  from the environment map and compute the outgoing radiance  $L$  in direction  $\omega_o$  using the following equation:

$$L(\omega_o) = \int_{\Omega} L_i(\omega_i) f(\omega_i, \omega_o) (\omega_i \cdot \mathbf{n}) d\omega_i, \quad (1)$$

where the  $f(\omega_i, \omega_o)$  represents the BSDF function, and the integration domain is the hemisphere  $\Omega$  around the intersection normal  $n$ . For real-time rendering, we employ the split-sum approximation [23] and the integral radiance in Eq. (1) is approximated as:

$$L(\omega_o) \approx \int_{\Omega} f(\omega_i, \omega_o) (\omega_i \cdot \mathbf{n}) d\omega_i + \int_{\Omega} L_i(\omega_i) D(\omega_i, \omega_o) (\omega_i \cdot \mathbf{n}) d\omega_i. \quad (2)$$

The first term of this product only relies on the parameters  $(\omega_i \cdot \mathbf{n})$  and the roughness  $r$  of the BSDF  $f(\omega_i, \omega_o)$ , which is precomputed and stored in a 2D lookup texture. Meanwhile, the second term is the integral of the incident radiance with the GGX [50] normal distribution  $D$ , which is also precomputed and stored by a filtered cubemap following Karis [23].

In order to learn the environment lighting from 2D image observations, we employ a differentiable split sum shading model. We represent the environment light as a

high-resolution cube map, which is the trainable parameters in the shading model. Owing to the precomputation and lookup mechanism, the rendering process from the texture/normal/environment map to the 2D images can be greatly accelerated. This approach serves as a bridge that allows ITEM3D to directly optimize the texture and environment map through rendered 2D images, rather than optimizing the complex 3D representation. As a result, the number of parameters to be optimized is significantly reduced, leading to a more efficient editing process.

### 3.3. Relative Direction Based Optimization

To enable the appearance editing of 3D models using natural language, a direct idea is to utilize the diffusion model that has been pre-trained in 2D images as knowledge prior. Naively, we can use Score Distillation Sampling (SDS) loss,

$$\nabla_{\theta} \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = \mathcal{R}(\theta)) = \mathbb{E}_{t, \epsilon} \left[ w(t) (\epsilon_{\phi}^{\omega}(\mathbf{z}_t; y, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} \right], \quad (3)$$

where  $\mathbf{x}$  is the rendered images,  $t$  is the sampled time step,  $\mathbf{z}_t$  is the  $t$  time step latent,  $w(t)$  is the weighting function that equals  $\partial \mathbf{z}_t / \partial \mathbf{x}$ ,  $y$  is the text condition,  $\epsilon_{\phi}^{\omega}(\mathbf{z}_t; y, t)$  is the predicted noise through classifier-free guidance, and  $\epsilon \in N(0, I)$  is the noise added to the rendered images. The gradient of SDS loss gives an editing direction for our texture optimization, determined by the text prompt  $y$ . However, the SDS loss may harm the content of original images with noisy details, because the text prior typically cannot faithfully reflect the information of the image. It is known that the entropy of an RGB image is significantly larger than that of a text prompt. As a consequence, the misdescription inevitably arises when taking the text prompt as the prior to restore the high-quality image from the same-scale noise. Therefore, even for a text prompt  $y_0$  describing the original images, there exists a deviation related to the optimized texture  $\theta$  between the added noise  $\epsilon$  and the predicted noise

$\epsilon_\phi^\omega(\mathbf{z}_t; y_0, t)$ , which can be simply expressed as,

$$D_{\text{bias}}(\theta, \dots) \propto \|\epsilon_\phi^\omega(\mathbf{z}_t; y_0, t) - \epsilon\|. \quad (4)$$

Thus, the gradient leads to a bias term from the original input image, which can be expressed as,

$$\vec{n}_{\text{bias}} = \frac{\partial D_{\text{bias}}(\theta, \dots)}{\partial \theta} = (\epsilon_\phi^\omega(\mathbf{z}_t; y_0, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta}. \quad (5)$$

Moreover, for an arbitrary text prompt  $y_{\text{tgt}}$  describing the target editing texture, it could be considered that there exists a term of expected editing direction and a term of bias discussed above,

$$(\epsilon_\phi^\omega(\mathbf{z}_t; y_{\text{tgt}}, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} = \vec{n}_{\text{tgt}} + \vec{n}_{\text{bias}}. \quad (6)$$

As a result, the  $\vec{n}_{\text{bias}}$  gives rise to the misdirection for the optimization procedure.

To address these issues, it is ideal to find the accurate editing direction  $\vec{n}_{\text{tgt}}$ , while the term of  $\vec{n}_{\text{bias}}$  is hard to estimate due to the diverse input images. To mitigate the gap, it is natural to take the text guidance as a relative direction rather than an absolute direction, enabling us to eliminate the term of  $\vec{n}_{\text{bias}}$ . The absolute direction of the source  $\vec{n}_{\text{src}}$  and the target  $\vec{n}_{\text{tgt}}$  can be expressed as,

$$\vec{n}_{\text{src}} = (\epsilon_\phi^\omega(\mathbf{z}_t; y_0, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} - \vec{n}_{\text{bias}}, \quad (7)$$

$$\vec{n}_{\text{tgt}} = (\epsilon_\phi^\omega(\mathbf{z}_t; y_{\text{tgt}}, t) - \epsilon) \frac{\partial \mathbf{x}}{\partial \theta} - \vec{n}_{\text{bias}}, \quad (8)$$

where  $\vec{n}_{\text{src}}$  is actually the  $\vec{0}$  giving no extra information to the input images. Inspired by the CLIP-directional loss improved by the StyleGAN-Nada [15] and the denoising loss proposed by the recent work [38], we utilize the difference between the source  $\vec{n}_{\text{src}}$  and the target  $\vec{n}_{\text{tgt}}$  as the relative direction of the target, which can be presented as,

$$\vec{n}_{\text{tgt}} = \vec{n}_{\text{tgt}} - \vec{n}_{\text{src}} = (\epsilon_\phi^\omega(\mathbf{z}_t, y_{\text{tgt}}, t) - \epsilon_\phi^\omega(\mathbf{x}, y_0, t)) \frac{\partial \mathbf{x}}{\partial \theta}. \quad (9)$$

Therefore, the final gradient utilized for optimizing the texture can be presented as,

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{RD}}(\phi, \mathbf{x} = \mathcal{R}(\theta)) = \\ \mathbb{E}_{t, \epsilon} \left[ w(t) (\epsilon_\phi^\omega(\mathbf{z}_t; y_{\text{tgt}}, t) - \epsilon_\phi^\omega(\mathbf{z}_t; y_0, t)) \frac{\partial \mathbf{x}}{\partial \theta} \right]. \end{aligned} \quad (10)$$

### 3.4. Direction Adjustment

Different from the gradual transition in the nature image domain, the optimization of the texture domain unfortunately shows an unexpected offset of the appearance in rendered images. The inherent reason is that the complexity of rendering leads to unbalanced optimization for the texture, with

some parts under-tuning and other parts over-tuning. This appearance offset can be seen in some parts of the rendered image, leading to the inconsistency between the source text and the rendered images in the median period of the optimization procedure. It is known that a source image with an inconsistent text description means an optimization misdirection which leads to an unknown change in the editing results. Similarly, if a rendered image during the intermediate optimization hops out of the direction between the source text and the target text, it can be considered as an inconsistent description of the source image when we take the current median point as a relative beginning point. The original editing direction is given by,

$$\vec{n}_{\text{ori}} = \vec{n}_{\text{tgt}} - \vec{n}_{\text{src}}. \quad (11)$$

If the optimization continues along the original direction, a more severe deviation of the material will be attached to the editing procedure.

To avoid the misdirection caused by the texture domain, we propose to adjust the editing direction, gradually changing the source text prompt during our optimization process of the texture map. The direction after adjustment  $\Delta \hat{T}_i$  can be represented as,

$$\Delta \hat{T}_i = \vec{n}_i - \vec{n}_{i-1} = \mathcal{B}(I_i) - \mathcal{B}(I_{i-1}), \quad (12)$$

where  $i$  is the optimization iteration and  $\mathcal{B}(\cdot)$  expresses the inverse text generated by a pre-trained language-image model BLIP-v2 [26]. As shown in the Fig. 1 (b), the direction is continually adjusted during the optimization so that the new global direction  $\vec{n}_i$  can be written as,

$$\vec{n}_i = \vec{n}_{\text{ori}} + \sum_{j \leq i} \Delta \hat{T}_j. \quad (13)$$

By adjusting the optimization direction step by step, we achieve more delicate and controllable editing, which can be seen in Sec. 4.4.

## 4. Experiments

### 4.1. Experiment Setup

**Dataset.** In the experiments, we evaluate our model on the NeRF Synthetic [34] dataset, Keenan’s 3D model repository and a set of real-world products. The NeRF synthetic dataset consists of 8 path-traced scenes with multi-view images which we reconstruct into our textural mesh-based representation via nvdiffr [36]. The Keenan’s 3D model repository and real-world dataset contains explicit mesh and texture map for each 3D model.

**Implementation Details.** We optimize 3D models on one RTX A6000 GPU with 48G memory. The procedure lasts about 600 iterations within 10 minutes for each 3D model.

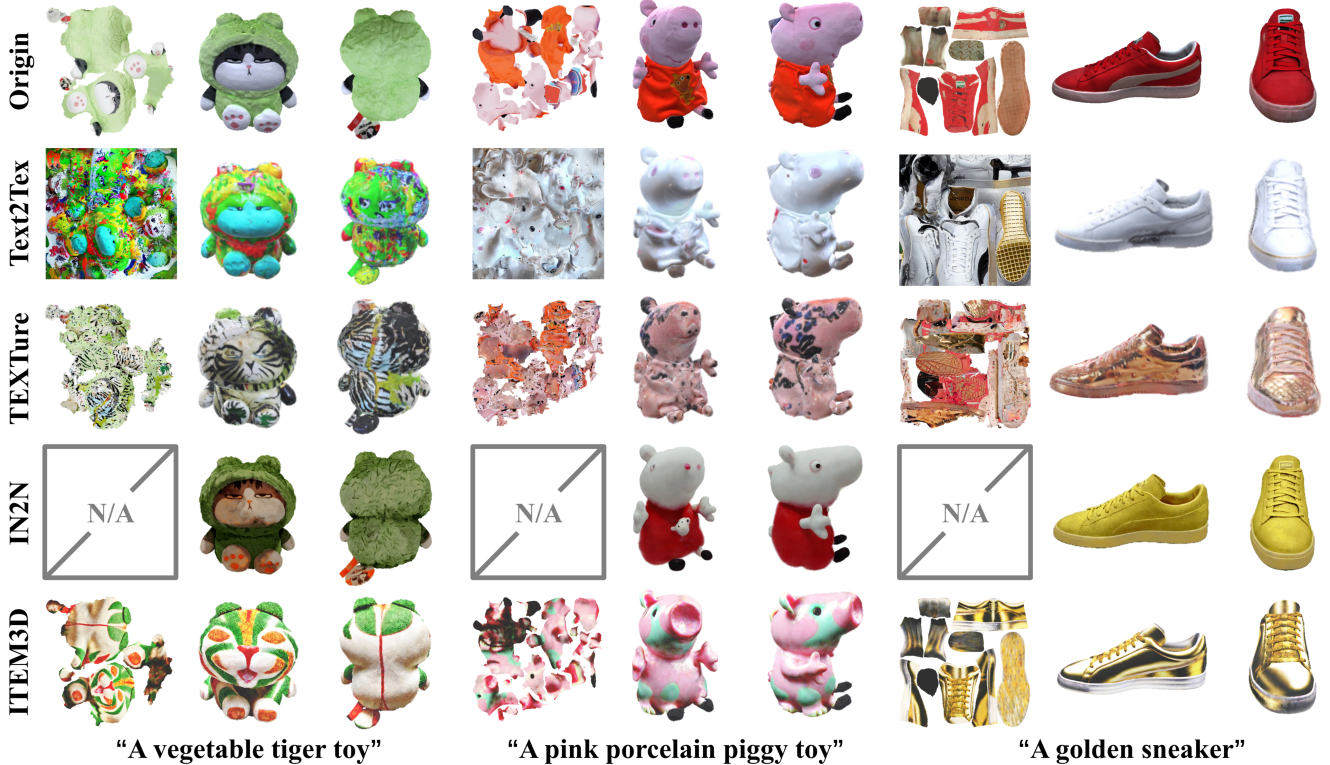


Figure 3. **Performance on real-world objects.** ITEM3D successfully transform the original cat toy into a vegetable tiger toy, the piggy doll into a porcelain pig, and the sneaker into a golden sneaker with remarkable quality. In contrast, Text2Tex [8] and TEXTure [42] both generate noisy textures resulting in low-quality rendering appearance. Instruct-NeRF2NeRF [17] achieves natural and text-consistent appearance, but it fails to edit the material of the objects. It is noted that the text instructions given to the Instruct-NeRF2NeRF are slightly different from other methods, i.e., “*Make it into a vegetable tiger toy*”, “*Make it into a porcelain piggy toy*” and “*Make it into a golden sneaker*”, due to its special requirement.

## 4.2. Qualitative Results

**Results on Real-world Objects.** We perform our ITEM3D, TEXTure [42] (SIGGRAPH 2023), Text2Tex [8] (ICCV 2023) and Instruct-NeRF2NeRF [17] (ICCV 2023) on a real-world dataset, and the results are depicted in Fig. 3. Our ITEM3D showcases remarkable texture editing capabilities for real products, including a shoe, a piggy doll, and a toy cat. For instance, when given the text prompt “*A vegetable tiger toy*”, ITEM3D effectively transforms the texture into that of a cute, furry tiger while preserving the structure of the original toy cat. Furthermore, when the prompts involve both material and texture descriptions, e.g., “*A golden sneaker*”, our model successfully bakes the golden material instead of the original material into the texture and shows a realistic shoe. Similarly, the example of “*A pink porcelain piggy toy*” also achieves realistic material editing along with consistent texture modifications as the provided text. Compared to our methods, Instruct-NeRF2NeRF [17] fails to learn the materials of porcelain and gold in the cases of the piggy toy and sneaker de-

spite the natural appearance. Meanwhile, it is difficult for Text2Tex [8] and TEXTure [42] to handle complex objects, resulting in low-quality appearances. Besides, Text2Tex suffers from the problem of multiple heads in the example of tiger toy, due to its patch-based texture learning. These results clearly demonstrate the generalization ability of ITEM3D in handling complex real-world objects.

**Comparison on Synthetic Objects.** We also compare ITEM3D with the SDS-based optimization method, and Instruct-NeRF2NeRF (IN2N) [17] in the NeRF synthetic dataset without explicit mesh and texture, as shown in Fig. 4. While the SDS-based method can edit textures along the direction of the text prompt, its rendered images show an unrealistic appearance and tend to overfit the text. In contrast, ITEM3D and IN2N can render realistic images with high quality while remaining consistent with the input text prompt. However, IN2N has certain limitations in comparison to our approach. For instance, when editing chairs, IN2N produces results with smooth textures, while our method captures more fine-grained details. Furthermore, when editing a ficus plant with the prompt “*Turn the*

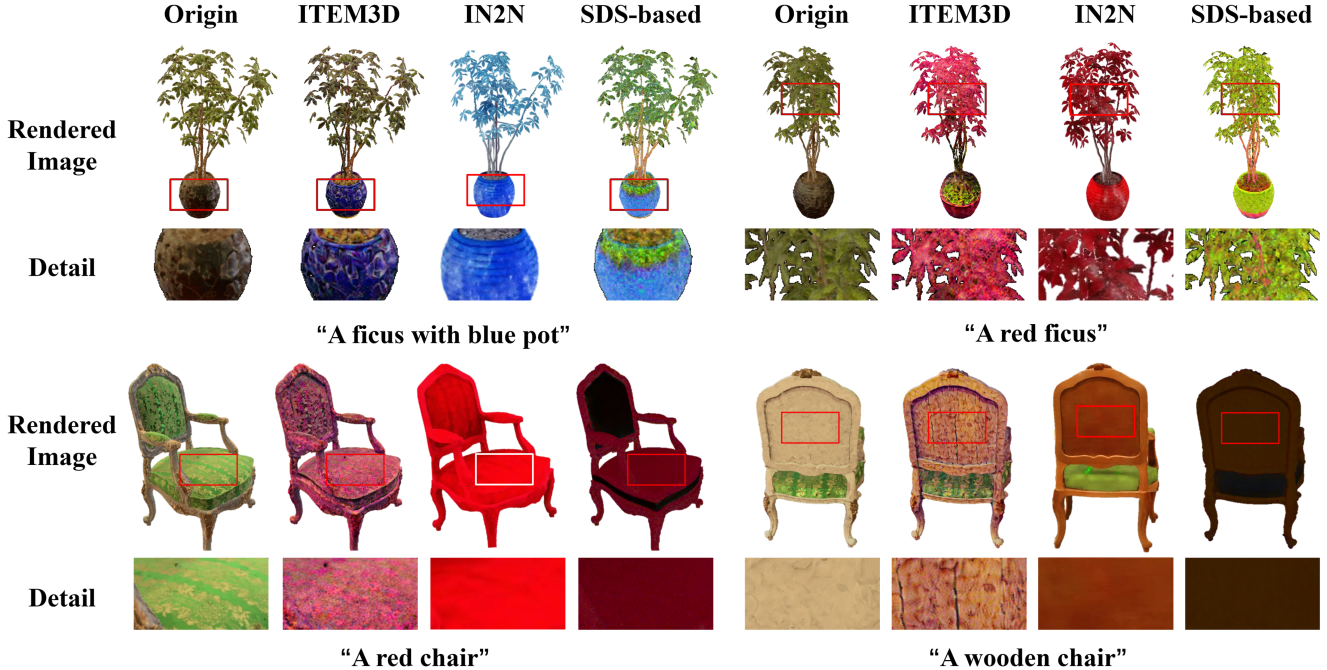


Figure 4. **Qualitative comparison** on NeRF synthetic dataset. We conducted an analysis of our method with the state-of-the-art approach, Instruct-NeRF2NeRF (IN2N) [17] and the simple SDS-based method. While both ITEM3D and IN2N demonstrate prompt-consistent editing, the SDS-based method fails to yield satisfactory outcomes. Conversely, IN2N exhibits a loss of the original chair patterns and an inability to faithfully represent natural wood patterns. Moreover, it lacks the necessary precision to accurately discern the edited object.

*pot into a blue pot*”, IN2N faces the challenge of distinguishing between the pot and the plant, resulting in both elements being edited to the same blue color. The comparison indicates the effectiveness of the introduced relative direction and direction adjustment.

### 4.3. Quantitative Results

We perform a quantitative comparison with Text2Tex [8], TEXTure [42], and IN2N [17] on the editing quality, efficiency and user study as presented in Tab. 1.

**Editing Quality.** To evaluate the semantic consistency, we render  $512 \times 512$  RGB images after texture editing, and further compute the CLIP-Score of the rendered image and corresponding target text. CLIP-score contains two parts, *i.e.*, global score and directional score. Global score measures the similarity between the target text and the editing images, and directional score measures the similarity between two editing directions of text prompts and images, which can be expressed as,

$$\text{Score}_{\text{global}} = \frac{T_{\text{tgt}} \cdot I_{\text{tgt}}}{\|T_{\text{tgt}}\| \|I_{\text{tgt}}\|}, \quad \text{Score}_{\text{direction}} = \frac{\Delta T \cdot \Delta I}{\|\Delta T\| \|\Delta I\|}, \quad (14)$$

where  $T_{\text{tgt}}$  and  $I_{\text{tgt}}$  are the embedding of target text and edited image encoded by the CLIP encoder, and  $\Delta T$  and

$\Delta I$  are expressed as,

$$\Delta T = T_{\text{tgt}} - T_{\text{src}}, \quad \Delta I = I_{\text{tgt}} - I_{\text{src}}. \quad (15)$$

As shown, our method achieves better results than Text2Tex, TEXTure and IN2N in both global score and directional score, demonstrating the superiority of our method in the image quality and text consistency.

**Editing Efficiency.** Moreover, to demonstrate the editing efficiency of ITEM3D, we also compare the training time and GPU memory with Text2Tex, TEXTure and IN2N. Our findings demonstrate that ITEM3D exhibits comparable time efficiency to TEXTure, while outperforming Text2Tex and IN2N in texture editing. Moreover, our method ITEM3D consumes less GPU memory compared to other approaches.

**User Study.** Additionally, we perform a user study to further assess the quality of editing objects. Users are required to rate on a scale of 1 to 5, based on the following questions: (1) Are the edited objects realistic and natural (Photo-realism)? (2) Do the edited objects accurately reflect the target text’s semantics (Text Consistency)? The results demonstrate the superior quality with higher realism and more text consistency of our proposed method as compared to the baselines. The results of user study are also consistent to the qualitative analysis that IN2N achieves nat-

Method	Venue	Global Score $\uparrow$	Directional Score $\uparrow$	Training Time $\downarrow$	GPU Memory $\downarrow$	Photorealism $\uparrow$	Text Consistency $\uparrow$
Text2Tex [8]	ICCV 2023	0.294	0.121	15 min	10 GB	3.62	2.79
TEXTure [58]	SIGGRAPH 2023	0.299	0.097	<b>5 min</b>	12 GB	3.03	2.41
IN2N [17]	ICCV 2023	0.311	0.159	10 h	15 GB	4.26	3.36
ITEM3D (ours)	—	<b>0.333</b>	<b>0.174</b>	8 min	<b>9 GB</b>	<b>4.38</b>	<b>4.18</b>

Table 1. **Quantitative Comparisons.** We report two CLIP-based scores, *i.e.*, global score and directional score to evaluate the semantic quality of rendered images. We also compare the editing time and memory consumption of four methods. Besides, we conduct a user study with 33 participants. Each participant scores based on two evaluation criteria, *i.e.*, photorealism and text consistency.



Figure 5. **Ablation study** of direction adjustment. The results without adjustment show a wired appearance, *i.e.*, a pale body of the cattle and dual heads of the duck. However, with the application of gradual adjustment, the unrealistic artifacts are released, leading to a natural appearance.

ural image appearance but fails on text consistency, while Text2Tex and TEXTure have lower image quality.

#### 4.4. Direction Adjustment

In this section, we further study the necessity of direction adjustment. We perform the ablation study in Fig. 5. Without the adjustment for the relative optimization direction, the texture shows a wired change that the duck gradually generates two heads and the color seems partially yellow and partially red. When applying the gradual adjustment, the duck bypasses the unnatural change and smoothly achieves the target appearance. The example of cattle shows a similar trend. In this experiment, it can be noticed that there exists unbalanced optimization for different parts of the texture. The optimization scheme of simple pieces of texture converges quickly, while more complex modifications require longer time, which in turn over-tunes easy parts leading to poor results.

#### 4.5. Relighting Application

The disentangled representation of the environment map empowers ITEM3D to explicitly relight the 3D model. The results of relighting are demonstrated in Fig. 6. The Lego

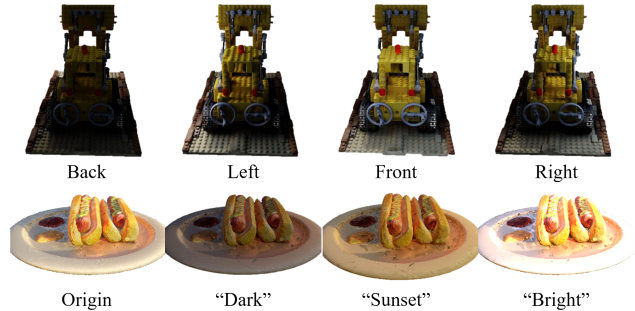


Figure 6. **Relighting results.** ITEM3D has capacity of explicit control over the lighting direction by manipulating the environment map. Furthermore, ITEM3D can relight the 3D model based on the text guidance via the optimization of the environment map.

cases show the capability to modify the lighting direction by directly manipulating the environment map. Moreover, the hotdog cases showcase the ability to guide relighting using text prompts, while maintaining the texture constant. During the text-guided relighting, we fix the texture and mesh components and focus on optimizing the environment map based on provided text prompts, such as “dark”, “sunset” and “bright light”. It proves that the diffusion model can utilize the lighting prior information from 2D images to guide the optimization for the environment map by leveraging the differential rendering process.

## 5. Conclusion and Limitations

In conclusion, our ITEM3D model presents an efficient solution to the challenging task of texture editing for 3D models. By leveraging the knowledge from diffusion models, ITEM3D is capable to optimize the texture and environment map under the guidance of text prompts. To address the semantic ambiguity between text prompts and images, we replace the traditional score distillation sampling (SDS) with a relative editing direction. We further propose a gradual direction adjustment during the optimization procedure, solving the unbalanced optimization in the texture.

Despite the promising editing results, our ITEM3D still remains several limitations which should be solved in fu-



ture work. The major limitation is that our method may encounter challenges when dealing with complex and precise editing prompts. Another limitation is that it is hard to edit the lighting direction directly by the guidance of text prompt. Our further work aims to explore the UV mapping between neural texture and rendered images to achieve the disentangled and precise texture editing.

## References

- [1] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial GAN. In *ICML*, pages 469–477, 2017. [1](#)
- [2] Sema Berkiten, Maciej Halber, Justin Solomon, Chongyang Ma, Hao Li, and Szymon Rusinkiewicz. Learning detail transfer based on geometric features. *Comput. Graph. Forum*, 36(2):361–373, 2017. [1](#)
- [3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerf: Neural reflectance decomposition from image collections. In *ICCV*, pages 12684–12694, 2021. [2](#)
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. [3](#)
- [5] Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, pages 1–7, 2012. [3](#)
- [6] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023. [3](#)
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, pages 333–350, 2022. [2](#)
- [8] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. [3](#), [6](#), [7](#), [8](#)
- [9] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. [3](#)
- [10] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. [1](#), [3](#)
- [11] Zhuo Chen, Xudong Xu, Yichao Yan, Ye Pan, Wenhan Zhu, Wayne Wu, Bo Dai, and Xiaokang Yang. Hyperstyle3d: Text-guided 3d portrait stylization via hypernetworks. *arXiv preprint arXiv:2304.09463*, 2023. [1](#), [3](#)
- [12] Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, pages 1033–1038, 1999. [1](#)
- [13] Anna Frühstück, Ibraheem Alhashim, and Peter Wonka. Tiledgan: synthesis of large-scale non-homogeneous textures. *ACM Trans. Graph.*, 38(4):58:1–58:11, 2019. [1](#)
- [14] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. In *NeurIPS*, pages 3403–3416, 2022. [2](#)
- [15] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4):141:1–141:13, 2022. [5](#)
- [16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, pages 1–13, 2022. [3](#)
- [17] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. [3](#), [6](#), [7](#), [8](#)
- [18] Amir Hertz, Rana Hanocka, Raja Giryes, and Daniel Cohen-Or. Deep geometric texture synthesis. *ACM Trans. Graph.*, 39(4):108, 2020. [1](#)
- [19] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. *arXiv preprint arXiv:2304.07090*, 2023. [2](#)
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. [3](#)
- [21] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. [3](#)
- [22] Nikolay Jetchev. Clipmatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2109.12922*, 2021. [3](#)
- [23] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3):1, 2013. [4](#)
- [24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. [1](#)
- [25] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *CVPR*, pages 14203–14213, 2023. [3](#)
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [5](#)
- [27] 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 preprint arXiv:2211.10440*, 2022. [1](#)
- [28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *CVPR*, pages 9298–9309, 2023. [1](#), [3](#)
- [29] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for uncon-

- strained photo collections. *arXiv preprint arXiv:2008.02268*, 2020. [2](#)
- [30] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. [1](#)
- [31] Tom Mertens, Jan Kautz, Jiawen Chen, Philippe Bekaert, and Frédo Durand. Texture transfer using geometry correlation. In *Proceedings of the Eurographics Symposium on Rendering Techniques*, pages 273–284, 2006. [1](#)
- [32] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. [1](#), [3](#)
- [33] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, pages 13492–13502, 2022. [3](#)
- [34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 99–106, 2020. [2](#), [5](#)
- [35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. [2](#)
- [36] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *CVPR*, pages 8280–8290, 2022. [2](#), [5](#)
- [37] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. *arXiv preprint arXiv:2104.10078*, 2021. [2](#)
- [38] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. [1](#), [2](#), [3](#), [5](#)
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. [3](#)
- [40] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. [3](#)
- [41] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *arXiv preprint arXiv:2103.13744*, 2021. [2](#)
- [42] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. [1](#), [6](#), [7](#)
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. [3](#)
- [44] Omry Sendik and Daniel Cohen-Or. Deep correlations for texture synthesis. *ACM Trans. Graph.*, 36(5):161:1–161:15, 2017. [1](#)
- [45] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, pages 20875–20886, 2023. [3](#)
- [46] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, pages 7495–7504, 2021. [2](#)
- [47] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. IDE-3D: interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Trans. Graph.*, 41(6):270:1–270:10, 2022. [1](#)
- [48] Jiayang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. [3](#)
- [49] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. [3](#)
- [50] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the Eurographics Symposium on Rendering Techniques*, pages 195–206, 2007. [4](#)
- [51] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *CVPR*, pages 3835–3844, 2022. [3](#), [1](#)
- [52] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. [2](#)
- [53] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf-: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. [2](#)
- [54] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. Texture reformer: Towards fast and universal interactive texture transfer. In *AAAI*, pages 2624–2632, 2022. [1](#)
- [55] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *CVPR*, pages 8534–8543, 2021. [2](#)
- [56] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *CVPR*, pages 8456–8465, 2018. [1](#)
- [57] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *ICCV*, pages 5752–5761, 2021. [2](#)
- [58] Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, and Xiaojuan Qi. Texture generation on 3d meshes with point-uv diffusion. In *CVPR*, pages 4206–4216, 2023. [3](#), [8](#)

- [59] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. [3](#)
- [60] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. [2](#)
- [61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. [3](#), [1](#)
- [62] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970*, 2021. [2](#)
- [63] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *ACM Trans. Graph.*, 37(4):49, 2018. [1](#)

# Directional Texture Editing for 3D Models

## Supplementary Material

Although the main paper stands on its own, it is still worthwhile to provide more details. In this supplementary document, we provide

- More qualitative results and additional comparisons
- Failure cases in precise editing.
- Broader impacts of our ITEM3D.

The code of our framework used in our experiments will be made publicly available. Additionally, we also provide a supplementary video that offers a brief introduction of our work and showcases 3D results rendered by a camera following a circular trajectory.

### 6. Additional Qualitative Results

To begin with, we present additional qualitative results to showcase generalized editing capabilities of our ITEM3D. These results cover both real-world data and synthetic data. In Fig. 9, we conduct extensive experiments on real-world dataset, displaying multi-view rendered images and textures before and after editing. These results illustrate the generalization of our method for real-world applications. The 3D models of these objects are also shown in our supplementary video. In Fig. 10, we also include additional results of our ITEM3D on synthetic dataset, consisting of both mesh and rendered images.

### 7. Compare with Baselines

To show the superiority of 3D consistency in 3D representations and knowledge prior in diffusion models, we also compare our approach with two baselines: CLIP-NeRF and a 2D baseline called ControlNet. We present the results of this comparison in Fig. 7.

### 8. Failure Case

During the experiments, we have observed that our method fails to precisely editing the correct part of the objects for complex editing. As shown in Fig. 8, we attempted to add a mustache to the cow. While it did generate a black mustache, it inadvertently altered the mustache located on the cow's body instead of its face.

### 9. Broader Impacts

Our ITEM3D allows for delicate editing of material properties of 3D objects while maintaining their geometric structure and UV map. This approach has broad applications in the gaming and e-shopping industries, where object materials can be edited through textual descriptions to create various 3D objects. Our method can directly edit the object

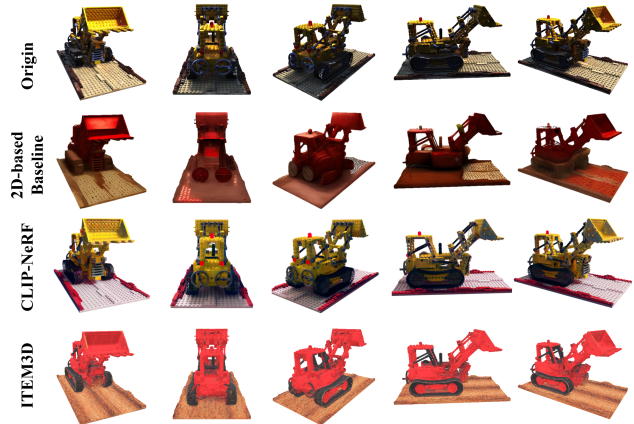


Figure 7. **Comparison with 2D baseline [61] and CLIP-NeRF [51].** We compare the performances of three methods under the prompt of “A real red excavator”. The 2D baseline, ControlNet, successfully edits the input views into realistic red excavator, however, the multi-view results lack 3D consistency. CLIP-NeRF maintains 3D consistency, while failing to edit the color of excavator. Our ITEM3D can both edit the texture according to the prompt and keep the inherent 3D consistency. The collection of five-view images currently exhibits distinct patterns and slight variations in the chair’s color.



“A milk cow toy with a black mustache on the face”

Figure 8. **Failure case.** Our method encounters challenges when dealing with complex and precise editing prompts. Our approach inadvertently alters the mustache located on the cow’s body instead of its face.

materials through textual descriptions, seamlessly incorporated into the industrial modeling pipeline. However, it is crucial to know that our method carries potential risks of deception. Our work may be used in product forgery which could become a threat to personal credit. We do not allow any application of our work to such malicious acts.



Figure 9. **Additional results** on real-world dataset. We show both textures and multi-view rendered images before and after editing.

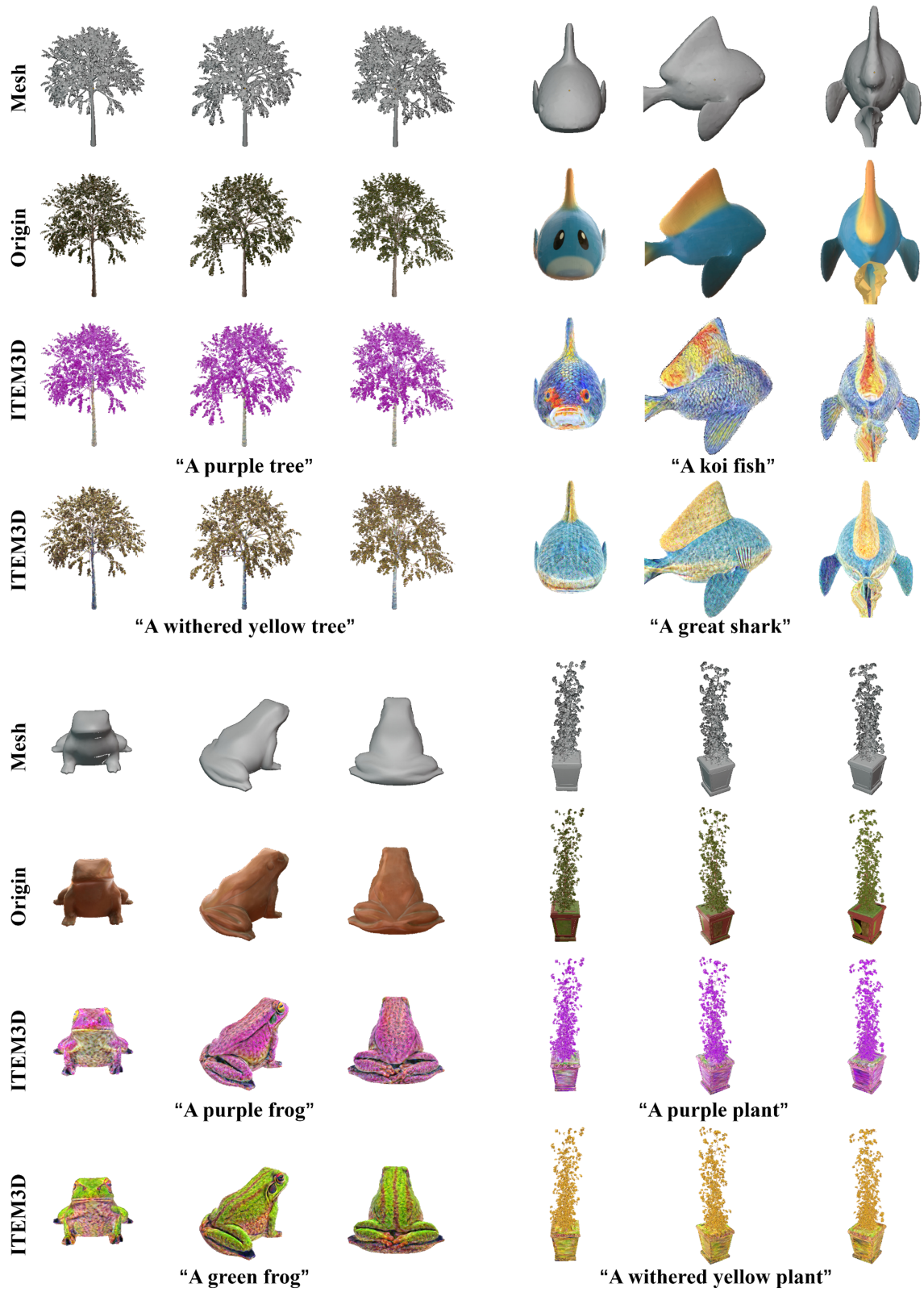


Figure 10. **Additional results** on synthetic dataset. The results of both mesh and rendered images are presented. We show the multi-view rendered images of two text prompts for each 3D object.