

# TexGen: Text-Guided 3D Texture Generation with Multi-view Sampling and Resampling

Dong Huo<sup>1,3\*</sup>, Zixin Guo<sup>2</sup>, Xinxin Zuo<sup>3</sup>, Zhihao Shi<sup>3</sup>, Juwei Lu<sup>3</sup>, Peng Dai<sup>3</sup>, Songcen Xu<sup>3</sup>, Li Cheng<sup>1</sup>, and Yee-Hong Yang<sup>1</sup>

<sup>1</sup> University of Alberta, Canada

{dhuo, lcheng5}@ualberta.ca, yang@cs.ualberta.ca

<sup>2</sup> University of Toronto, Canada

zixin.guo@mail.utoronto.ca

<sup>3</sup> Huawei Noah’s Ark Lab

{xinxin.zuo1, zhihao.shi, juwei.lu, peng.dai, xusongcen}@huawei.com

*A high quality color photo of Tom Cruise*



**Fig. 1:** Given a 3D mesh, we present text-driven texture generation results from previous state-of-the-art approaches (TEXTure [28], Text2Tex [7], Fantasia3D [8], and ProlificDreamer [35]) as well as our proposed method.

**Abstract.** Given a 3D mesh, we aim to synthesize 3D textures that correspond to arbitrary textual descriptions. Current methods for generating and assembling textures from sampled views often result in prominent seams or excessive smoothing. To tackle these issues, we present TexGen, a novel multi-view sampling and resampling framework for texture generation leveraging a pre-trained text-to-image diffusion model. For view consistent sampling, first of all we maintain a texture map in RGB space that is parameterized by the denoising step and updated after each sampling step of the diffusion model to progressively reduce the view discrepancy. An attention-guided multi-view sampling strategy is exploited to broadcast the appearance information across views. To preserve texture details, we develop a noise resampling technique that aids in the estimation of noise, generating inputs for subsequent denoising steps, as directed by the text prompt and current texture map. Through an extensive amount of qualitative and quantitative evaluations, we demonstrate that our proposed method produces significantly better texture quality for diverse 3D objects with a high degree of view consistency and rich appearance details, outperforming current state-of-the-art methods. Furthermore, our proposed texture generation technique can also be applied

\* Work done during an internship at Huawei Noah’s Ark Lab

to texture editing while preserving the original identity. More experimental results are available at <https://dong-huo.github.io/TexGen/>.

## 1 Introduction

Generating high-quality 3D content is an essential component of visual applications in films, games, and upcoming AR/VR industries. While many prior works on 3D synthesis have focused on the geometric components of the assets, textures have garnered less attention which play a vital role in enhancing the realism of 3D assets. In this paper, we aim to realize automatic text-driven 3D texture synthesis for various meshes.

Recently, the research community has witnessed remarkable progress in text-to-image (T2I) generation [14, 15, 29, 31]. However, the generation of 3D assets still faces challenges due to the limited size of 3D datasets [6, 9, 36], characterized by overly simplified textures. To this end, existing methods have been leveraging the visual information encoded in the 2D image priors of pre-trained T2I diffusion models. A thread of studies, such as score distillation sampling (SDS) and variational score distillation (VSD) [8, 18, 26, 35], aim to distil the diffusion priors as score functions to directly optimize a 3D representation (*e.g.*, geometry and texture), ensuring that its rendered outputs align well with the high-likelihood image priors. Despite remarkable successes in 2D-to-3D conversion, there are noticeable shortcomings. Specifically, textures generated using score distillation pipelines tend to be over-saturated, as illustrated in Fig.1(d), or suffer from issues such as blurry edges and color artifacts, as seen in Fig.1(e).

Beyond score distillation methods, another thread of studies for texture synthesis involves mapping multi-view images, generated from the T2I models, onto a global UV texture. For example, TEXTure [28] and Text2tex [7] adopted an autoregressive image inpainting pipeline to progressively assemble multi-view images generated from T2I models. While these methods can produce high-fidelity textures for particular views, they often exhibit noticeable seams on the assembled texture map, as evident in Fig.1(b) and (c). This issue arises due to error accumulation stemming from the autoregressive view inpainting process, and the primary cause of this error lies in the limited guidance provided by previously observed views [5]. Considering the sequential characteristics of the denoising process, TexFusion [5] proposed to mitigate the inconsistency of different views at each sampling step of the diffusion model. They conducted an interlaced multi-view sampling in latent space, after which they decoded the latent map to RGB image for each view independently via the pre-trained stable diffusion VAE decoder. However, the view inconsistencies recurred after the view independent decoding. A shallow MLP was optimized in the RGB texture space to smooth out the inconsistency, which will result in over-smoothed textures.

In response to the challenges of view inconsistency and over-smoothness in texture generation, we introduce **TexGen**, a novel multi-view sampling and re-sampling strategy that directly generates view consistent RGB images with rich appearance details from pre-trained 2D T2I models for texture assembling. In

detail, we derive a UV texture map in the RGB space that will be iteratively updated during each denoising step, to gradually unveil texture details. At each denoising step, we predict the latent *denoised observations* of sequentially sampled views around the 3D objects. These denoised observations are then decoded and assembled onto the texture map, enabling direct generation of view-consistent RGB textures without the need for additional MLP optimization steps [5]. An attention-guided multi-view sampling mechanism is proposed to ensure better appearance consistency across views within each denoising step.

More importantly, we develop a Text&Texture-Guided Resampling approach for noise estimation which leverages the information from both the renders of view consistent texture map at each denoising step and the high-frequency priors from the pre-trained T2I model. Through the fusion of texture and text-guided noise estimation, our generated textures not only maintain view consistency but also exhibit a rich diversity of details.

In summary, our key contributions can be outlined as follows: (1) we propose a multi-view sampling and resampling framework for text-driven texture generation using pre-trained 2D T2I models; (2) in particular, our proposed attention-guided multi-view sampling as well as text&texture-guided noise resampling techniques ensure that both 3D view consistency and rich details are preserved in the generated textures; (3) we demonstrate the effectiveness of our approach in texturing diverse 3D objects, showcasing superior performance compared to state-of-the-art methods. It is noteworthy that our proposed framework can naturally support text-driven texture editing as well.

## 2 Related Work

### 2.1 Diffusion Models in 3D Domain

Inspired by the success of 2D image generation with diffusion models, researchers have also attempted to utilize diffusion models to generate 3D objects in the form of various representations, such as point clouds [19, 25, 39, 41], and neural fields [24, 34]. For example, Point-E [25] trains a diffusion model using a large synthetic 3D dataset to produce a 3D RGB point cloud conditioned on a synthesized single view from a text prompt. However, these works mainly focus on geometry generation and do not specifically tackle 3D texture synthesis. Yu *et al.* [38] trains a diffusion model for mesh texture generation of specific object categories. Although Shap-E [16] is proposed to directly generate the parameters of implicit functions that can be rendered as both textured meshes and neural radiance fields, it cannot generalize to incorporate arbitrary text prompts. Moreover, the generated textures tend to be over smoothed with rather low quality as compared with the generated images from the T2I model.

### 2.2 Lifting pre-trained 2D generative models to 3D

Initially, the process of distilling 3D objects from pre-trained 2D models has been enhanced by the development of joint text-image embedding, such as Contrastive Language-Image Pre-training (CLIP) [27]. For example, CLIP-Mesh [22]

learns to generate a mesh with the guidance of the CLIP text embedding and the corresponding image embedding of the diffusion model. However, since the CLIP guidance is rather sparse, the generated 3D models for CLIP-based approaches [17, 21, 30] are rather coarse.

Recently, researchers have leveraged large-scale 2D T2I diffusion models to distil individual 3D objects in the form of neural radiance fields. Among various distilling approaches, a dominant one is Score Distillation Sampling (SDS) [26]. SDS pioneered the approach with many follow-up works [8, 10, 18, 20, 32, 33, 35]. For example, Magic3D [18] proposed a coarse-to-fine strategy to improve generation quality. Latent-NeRF [20] performed distillation in the latent space of latent diffusion model (LDM) [29]. A crucial drawback of this line of work is that SDS typically requires strong guidance, resulting in low diversity and over-saturation of the generated textures. ProlificDreamer [35] proposed to address this issue with a Variational Score Distillation (VSD) algorithm that adopts a particle-based variational inference to estimate the distribution of 3D scenes instead of a single point as in SDS. Yet, it still suffers from issues like blurry edges and color artefacts.

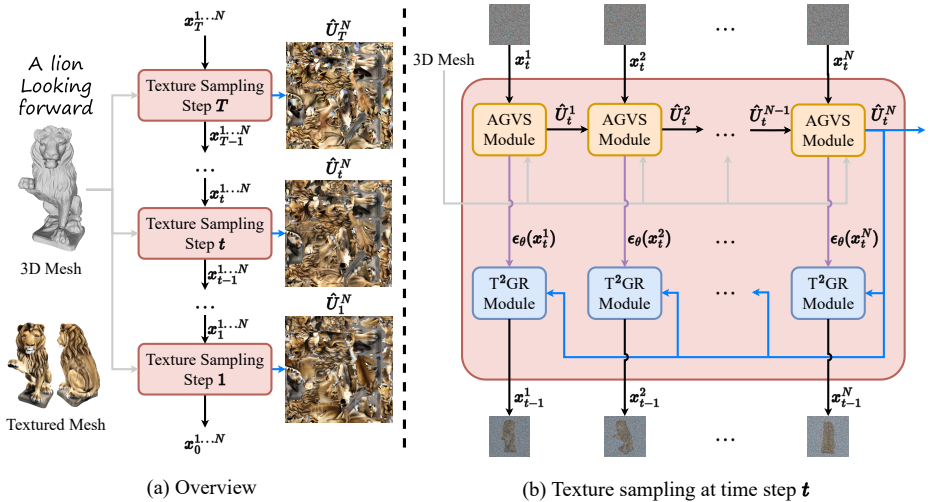
**Texture Synthesis with Multiview Denoising.** Instead of relying on the lengthy optimization of score distillation pipelines, an alternative research direction is directly leveraging the sampling process in diffusion models to synthesize UV textures. TEXTure [28] and Text2tex [7] adopt a depth-aware diffusion model [29, 40] to progressively paint the mesh surface from different views and aggregate the images generated from the T2I model of sampled views into the texture map. While rich textures and details can be faithfully synthesized, there were obvious seams on the assembled texture map due to error accumulation in the process of the autoregressive view update. To alleviate this problem, TexFusion [5] proposed to interleave texture assembling with denoising steps in different camera views and maintained a latent texture map at each sampling step. To convert latent features to RGB textures, they optimized an intermediate neural color field on the decoding of 2D renders of the latent texture which would wash out the rich details [23]. Our proposed approach distinguishes itself from previous methods with its ability to generate 3D-consistent textures while preserving rich details in the meantime.

## 3 Proposed Method

### 3.1 Overview

In this section, we present an overview of our proposed multi-view sampling and resampling strategy to synthesize view-consistent textures from a pre-trained T2I diffusion model. We first introduce the sampling process of the Denoising Diffusion Implicit Models (DDIM) [31], which forms the basis of our texture sampling approach.

**DDIM Sampling.** Assuming we sequentially sample  $N$  distinct views around a 3D mesh, the DDIM sampling process for each sampled viewpoint  $i$  at the



**Fig. 2:** Overview of our proposed method, where **AGVS** and **T<sup>2</sup>GR** denote Attention-Guided View Sampling and Text&Texture-Guided Resampling, respectively. First of all, we sample  $N$  viewpoints across the objects. As shown in (a), our texture sampling strategy is an interleaved process of texture generation and diffusion denoising. Specifically, our texture sampling process is structured into  $T$  denoising steps of diffusion process, and a complete RGB texture map ( $\hat{U}_t^N$ ) is generated at the end of each step. As shown in (b), at denoising step  $t$ , each AGVS module receives noisy latent features  $x_t^i$  as input to sample an image and produce a **partial texture map**  $\hat{U}_t^i$ , along with noise estimation  $\epsilon_\theta(x_t^i)$ . The generated  $\hat{U}_t^i$  serves as guidance for sampling the subsequent view. Subsequently, a **complete texture map**  $\hat{U}_t^N$  is employed to refine the noise estimation of each view within T<sup>2</sup>GR modules, facilitating the prediction of noisy features for the ensuing denoising step ( $x_{t-1}^{1..N}$ ).

denoising step  $t$  can be described as follows:

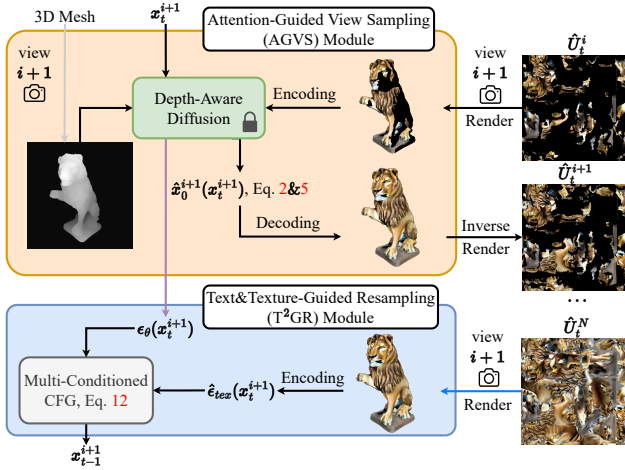
$$x_{t-1}^i = \sqrt{\alpha_{t-1}} \cdot \hat{x}_0^i(x_t^i) + \sqrt{1 - \alpha_{t-1}} \cdot \epsilon_\theta(x_t^i), \quad (1)$$

with

$$\hat{x}_0^i(x_t^i) = \frac{x_t^i - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(x_t^i)}{\sqrt{\alpha_t}}, \quad (2)$$

where  $x_t^i$  represents the noisy latent feature, and  $\epsilon_\theta(x_t^i)$  represents the estimated noise from the pre-trained diffusion model. At each denoising step  $t$ , we calculate  $\hat{x}_0^i(x_t^i)$ , representing the predicted  $x_0^i$  and dubbed as the *denoised observation* of  $x_t^i$ .  $\alpha_t$  is the total noise variance parameterized via denoising step  $t$ .

**Proposed Framework.** An overview of our proposed texture sampling is shown in Fig. 2(a). It leverages the sequential nature of the denoising process of the diffusion model and maintains the 3D consistency of the generated texture at each denoising step. We adopt the similar strategy of interlaced multi-view sampling as in [5]. But instead of relying on a post-processing to convert sampled view



**Fig. 3:** Details of denoising for view  $i+1$  at step  $t$ . The AGVS module is designed to generate denoised observation  $\hat{x}_0^{i+1}(x_t^{i+1})$  which will be assembled onto UV space to form intermediate texture  $\hat{U}_t^{i+1}$ . The attention guidance is omitted in the figure for simplification. After iterating over all sampled views starting from  $i=1$  to  $N$ , we obtain a complete texture map  $\hat{U}_t^N$  for each denoising step. Conditioned on the current aggregated texture map, the T<sup>2</sup>GR module will update the noise estimation  $\epsilon_\theta(x_t^i)$  with the multi-conditioned classifier-free guidance (CFG) to calculate the noisy latent feature  $x_{t-1}^{i+1}$  of the next denoising step.

consistent latents into RGB textures which resulted in over-smoothed textures, we directly enforce view consistent sampling in RGB texture space and develop an noise resampling strategy to retain rich texture details.

In detail, as shown in Fig. 2(b) and Fig. 3, we execute the following two steps at each denoising step  $t$ . First, by progressively assembling the *denoised observations*  $\hat{x}_0^i(x_t^i)$  where  $i=1, \dots, N$ , employing an attention-guided multi-view sampling strategy (Sec. 3.2), we compute a view-consistent **noise-free** texture map  $\hat{U}(x_t^{1\dots N})$ . For brevity, we denote  $\hat{U}_t^i$  as the partial texture map  $\hat{U}(x_t^{1\dots i})$  where  $i < N$ , and  $\hat{U}_t^N$  as the complete texture map  $\hat{U}(x_t^{1\dots N})$ , both at denoising step  $t$ . Second, to retain rich texture details, we conduct a Text&Texture-Guided Resampling step to calculate the noisy latent feature for the next denoising step  $t-1$ , conditioned on the current texture map  $\hat{U}_t^N$  as well as the text-guided noise estimation from a pre-trained T2I model as elaborated in Sec. 3.3.

Following the DDIM sampling, we go through the above process with  $T$  denoising steps to arrive at the final generated texture map  $\hat{U}_1^N$ . We present the above-mentioned two major steps in the following sections.

### 3.2 Attention-Guided Multi-View Sampling

As highlighted in Sec. 1, conducting a full denoising process in sequence for each view generation, conditioned on previously observed views, can result in

noticeable seams due to limited guidance from previous views. To mitigate this issue, we generate an RGB texture map at each denoising step with the *denoised observation*. Since each denoising step of the diffusion model is conditioned on a complete texture map from the preceding denoising step, this significantly reduces view inconsistency.

In particular, we follow DDIM sampling, for each sampled view  $i$  at denoising step  $t$ , the *denoised observation*  $\hat{x}_0^i(x_t^i)$  in the latent space can be computed as in Eq. 2. The *denoised observation*  $\hat{x}_0^i(x_t^i)$  are then decoded into images  $I_t^i$  in the RGB space via the VAE decoder  $\mathcal{D}$  of the pre-trained stable diffusion [29],

$$I_t^i = \mathcal{D}(\hat{x}_0^i(x_t^i)). \quad (3)$$

Starting with the first viewpoint  $i = 1$ , we inverse render  $I_t^i$  into the UV texture space, obtaining the partial texture map  $\hat{U}_t^i$ . Then for the subsequent viewpoint  $i + 1$ , the prediction of its *denoised observation* will depend on the current partial texture map. More specifically, we render the partial texture map  $\hat{U}_t^i$  onto viewpoint  $i + 1$ , which is fed as input to the VAE encoder  $\mathcal{E}$  to obtain the latent features  $G_t^{i+1}$ :

$$G_t^{i+1} = \mathcal{E}(\text{Render}^{i+1}(\hat{U}_t^i)). \quad (4)$$

Referring to  $\mathcal{M}^{i+1}$  as the mask delineating regions observed for the first time at view  $i + 1$  in RGB space, we adopt the approach of blended latent diffusion [2] to fuse the encoded render  $G_t^{i+1}$  with the original noisy latents  $x_t^{i+1}$  using  $\mathcal{M}^{i+1}$ , which aims to solely generate unobserved regions while preserving observed ones. In particular, we align the **noise-free** render  $G_t^{i+1}$  to the same noise level as  $x_t^{i+1}$  by adding randomly sampled noise  $\epsilon$  before blending. This process can be expressed as follows:

$$x_t^{i+1} \leftarrow x_t^{i+1} \odot \mathcal{M}_\downarrow^{i+1} + (\sqrt{\alpha_t} \cdot G_t^{i+1} + \sqrt{1 - \alpha_t} \cdot \epsilon) \odot (1 - \mathcal{M}_\downarrow^{i+1}), \quad (5)$$

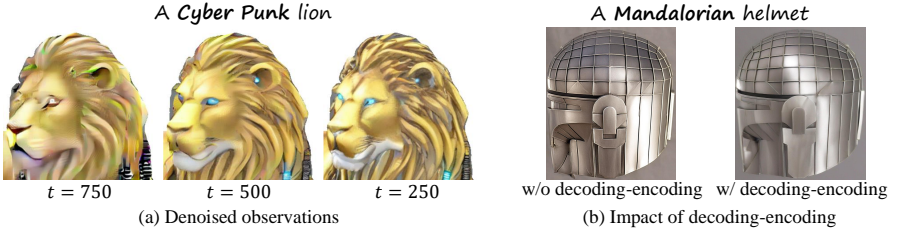
where  $\downarrow$  symbolizes downsampling to the resolution of latent features. The revised  $x_t^{i+1}$  is subsequently employed to compute the *denoised observation*  $\hat{x}_0^{i+1}(x_t^{i+1})$  for viewpoint  $i + 1$  at step  $t$ , in accordance with Eq. 2.

Furthermore, as indicated in Fig. 8(a), sequential generation across different viewpoints often falls short of ensuring consistent appearances, albeit without conspicuous seams. To address this, we introduce a novel attention-guided cross-view generation strategy. Drawing inspiration from the work of Cao *et al.* [4], we believe the Key and Value features in the self-attention module of the stable diffusion encapsulate the local contents and textures of generated images. In detail, we regard the front view as the reference view and propagate the Key and Value of the reference view to other views. The process can be outlined as follows:

$$\epsilon_\theta(x_t^{ref}), Q_t^{ref}, K_t^{ref}, V_t^{ref} \leftarrow \text{Unet}_\theta(x_t^{ref}), \quad (6)$$

$$\epsilon_\theta(x_t^i) \leftarrow \text{Unet}_\theta(x_t^i, K_t^{ref}, V_t^{ref}). \quad (7)$$

Herein,  $Q_t^{ref}$ ,  $K_t^{ref}$ , and  $V_t^{ref}$  denote the Query, Key, and Value features from the self-attention module of the reference view, respectively. In Eq. 7, the Key and



**Fig. 4:** (a) *Denoised observation*  $\hat{x}_0(x_t^i)$  of different denoising steps. The high-frequency information is gradually generated during sampling. (b) We claimed that the over-smoothness of directly using Eq. 8 for noise sampling is caused by repeatedly going through VAE decoder and encoder at each denoising step. For validation, we conducted an ablation for a simplified case of generating only a single viewpoint. It shows that the over-smoothness still existed even for single view generation. Mathematically, if we do not have encoding and decoding operation at each denoising step, single view sampling is exactly same as DDIM sampling.

Value features for each viewpoint are substituted with those from the reference view to calculate its estimated noise  $\epsilon_\theta(x_t^i)$ . Following this substitution, for each viewpoint  $i$ , the *denoised observation*  $\hat{x}_0^i(x_t^i)$  is updated in accordance with Eq. 2. As shown in Fig. 4(a), the texture details will gradually appear in the *denoised observation* as the diffusion process proceeds.

By sequentially applying the Eq. 3~Eq. 5 on all viewpoints with our proposed attention-guided cross-view generation, we obtain a complete, view-consistent and noise-free texture map  $\hat{U}_t^N$  for the current denoising step  $t$ .

### 3.3 Text&Texture-Guided Resampling (T<sup>2</sup>GR)

Upon obtaining the current texture map at step  $t$ , in this section we will present our Text&Texture-Guided Resampling (T<sup>2</sup>GR) approach for noise estimation to update the noisy latent features for the next denoising step  $t - 1$ .

As shown in Eq. 1, the derivation of denoised latents  $x_{t-1}^i$  depends on the estimated noise  $\epsilon_\theta(x_t^i)$  and the *denoised observation*  $\hat{x}_0^i(x_t^i)$ . Given that  $\hat{x}_0^i(x_t^i)$  is expected to exhibit view consistency maintained by the texture map  $\hat{U}_t^N$ , recalculating the noise map  $\epsilon_\theta(x_t^i)$  under the guidance of  $\hat{U}_t^N$  ensures to preserve the view consistency. Specifically, in Eq. 2 we set  $\hat{x}_0^i(x_t^i)$  equal to the current encoded render of the texture map  $\hat{U}_t^N$  at view  $i$ . From this, we derive the recalculated noise map  $\hat{\epsilon}_{tex}(x_t^i)$  as follows:

$$\hat{\epsilon}_{tex}(x_t^i) = \frac{x_t^i - \sqrt{\alpha_t} \cdot \mathcal{E}(\text{Render}^i(\hat{U}_t^N))}{\sqrt{1 - \alpha_t}}. \quad (8)$$

This recalculated noise map is then utilized in place of  $\epsilon_\theta(x_t^i)$  in Eq. 1 and Eq. 2 for the computation of  $x_{t-1}^i$ .

While the above noise map update strategy ensures view consistency, it tends to result in over-smoothed images (as shown in Fig. 4(b) and Fig. 8(b)). This is



primarily because the VAE encoder  $\mathcal{E}$  in the stable diffusion model compresses high-frequency details, referred to as *imperceptible details*, as noted by [29]. The repeated use of the encoder  $\mathcal{E}$  leads to an accumulation of this detail compression, affecting the overall image quality.

To avoid over-smoothness, we utilize the text-guided noise estimation which is not affected by the repeatedly encoding and decoding operation. Meanwhile, we take the current texture map as an additional condition to derive a multi-conditioned noise estimation formulation. The text-guided noise  $\epsilon_\theta(x_t^i|c)$  can be directly computed from the diffusion model and now we need to compute the texture-conditioned noise estimation which we denote as  $\epsilon_{tex}(x_t^i|\hat{U}_t^N)$ . By analyzing the formulation of  $\epsilon_\theta(x_t^i)$ , we see that it is essentially a weighted combination of conditional noise prediction  $\epsilon_\theta(x_t^i|c)$  and unconditional noise prediction  $\epsilon_\theta(x_t^i|\emptyset)$ , following the Classifier-Free Guidance (CFG) introduced in [15]:

$$\epsilon_\theta(x_t^i) = \epsilon_\theta(x_t^i|\emptyset) + \omega(\epsilon_\theta(x_t^i|c) - \epsilon_\theta(x_t^i|\emptyset)), \quad (9)$$

where  $c$  and  $\emptyset$  represent the text prompt and null-text prompt, respectively, and  $\omega$  is a user-specified weight. Similarly, the  $\epsilon_{tex}(x_t^i)$  is assumed to follow the same formulation of CFG,

$$\epsilon_{tex}(x_t^i) = \epsilon_\theta(x_t^i|\emptyset) + \omega(\epsilon_{tex}(x_t^i|\hat{U}_t^N) - \epsilon_\theta(x_t^i|\emptyset)). \quad (10)$$

Thus, to disentangle the texture-conditioned noise estimation  $\epsilon_{tex}(x_t^i|\hat{U}_t^N)$ , we subtract the null-text conditioned noise estimation from  $\epsilon_{tex}(x_t^i)$ . Here we have  $\epsilon_{tex}(x_t^i) = \hat{\epsilon}_{tex}(x_t^i)$  from Eq.8. The computation for the texture-conditioned noise estimation  $\epsilon_{tex}(x_t^i|\hat{U}_t^N)$  is as follows:

$$\epsilon_{tex}(x_t^i|\hat{U}_t^N) = \frac{1}{\omega}(\hat{\epsilon}_{tex}(x_t^i) - \epsilon_\theta(x_t^i|\emptyset)) + \epsilon_\theta(x_t^i|\emptyset). \quad (11)$$

In the end, we formulate our multi-conditioned CFG for final noise estimation, which is conditioned on both the textual prompt and texture map:

$$\epsilon_m(x_t^i) = \epsilon_\theta(x_t^i|\emptyset) + \omega_1(\epsilon_\theta(x_t^i|c) - \epsilon_\theta(x_t^i|\emptyset)) + \omega_2(\epsilon_{tex}(x_t^i|\hat{U}_t^N) - \epsilon_\theta(x_t^i|\emptyset)), \quad (12)$$

where  $\omega_1 + \omega_2 = \omega$ . We exploit a large  $\omega_2$  for early sampling steps, which will decrease linearly from  $\omega$  to 0 in the process of denoising. The comprehensive derivation of Eq. 12 can be found in the supplementary materials. Finally, we compute  $x_{t-1}^i$  for the subsequent denoising step by letting  $\epsilon_\theta(x_t^i) = \epsilon_m(x_t^i)$  in Eq. 1 and Eq. 2.

By combining the attention-guided multi-view sampling and text&texture-guided resampling, our proposed method can directly generate high-fidelity and view consistent texture map in the RGB space, without the need of an MLP post-processing as in TexFusion [5] which results in over-smoothed textures.

## 4 Experiments

### 4.1 Implementation Details

We employ the depth-aware diffusion model provided by ControlNet [40] as our T2I backbone with denoising steps  $T = 40$ . To render objects, we take eight



**Fig. 5:** Visual comparison of our proposed method against TEXTure [28] and Text2Tex [7].

different viewpoints around the object. The pose is sampled in spherical coordinates, with elevation angles being zero and azimuth angles uniformly sampled between  $[0^\circ, 360^\circ]$ . An additional top view is sampled. Additionally, we employ the Xatlas [37] tool to compute the UV atlas for a given mesh.

**Dataset.** Our experiments incorporate a diverse collection of 45 meshes, sourced from various datasets such as Objaverse [9] and ThreeDScans [1], with 2 to 3 distinct prompts for each mesh. Please refer to the supplementary for details.

## 4.2 Compared Methods

We conduct experimental comparison over several state-of-the-art approaches, including TEXTure [28], Text2Tex [7], Fantasia3D [8], ProlificDreamer [35] and TexFusion [5]. For TEXTure, Text2Tex, and Fantasia3D, we use their respective publicly available codebase. For ProlificDreamer, we adopt the implementation of ThreeStudio [11] and replaced its backbone with ControlNet [40] to recognize the depth. In the case of TexFusion, where the implementation is not available, our analysis is limited to a qualitative assessment using results extracted directly from the original paper. Notably, for all the compared approaches, the geometry remains fixed during texture generation.

## 4.3 Qualitative Comparison

We provide visual comparison in Fig. 5 and Fig. 6. Specifically, in Fig. 5, we showcase the robustness of our approach in addressing fragmented textures



**Fig. 6:** Visual comparison of our proposed method against Fantasia3D [8] and ProlificDreamer [35].



**Fig. 7:** Visual comparison of our proposed method against TexFusion [5]. The results of TexFusion are directly copied from its original paper.

against progressively texture assembling approaches, namely TEXTure [28] and Text2Tex [7]. This improvement is credited to our use of attention-guided view sampling combined with a distinct text&texture-guided resampling approach, which maintain view consistency at each denoising step to persistently enhance 3D consistency.

In Fig. 6, we compare with score distillation based approaches, namely Fantasia3D [8] and ProlificDreamer [35]. As demonstrated, Fantasia3D typically produces textures that are over-smoothed and over-saturated, while ProlificDreamer, though more detailed and contrasted, is marred by evident artifacts of

**Table 1:** Quantitative comparison on generated textures.

Methods	FID ↓	KID $\times 10^{-3}$ ↓	CLIPScore ↑
TEXTure	99.06	7.23	19.73
Text2Tex	109.94	7.17	21.26
Fantasia3D	108.58	7.52	21.14
ProlificDreamer	94.51	7.00	21.25
Ours	<b>84.65</b>	<b>4.27</b>	<b>22.83</b>

**Table 2:** User Study Preference: The entries in the table indicate our preference over other methods. A higher value represents a greater preference.

	TEXTure ↑	Text2Tex ↑	Fantasia3D ↑	ProlificDreamer ↑
Ours	64.72%	71.46%	70.97%	69.18%

blurry edges. In contrast, our method surpasses these distillation-based methods by generating more realistic high-quality results.

**Comparison with TexFusion [5].** We also present a qualitative comparison of our method with TexFusion [5] in Fig. 7. TexFusion employed instantNGP [23] to mitigate inconsistencies post-decoding of latent features into RGB space, which often led to over-smoothed results. In contrast, our method effectively generates textures that are consistent across views and retain rich details. Please refer to supplementary materials for more visual results.

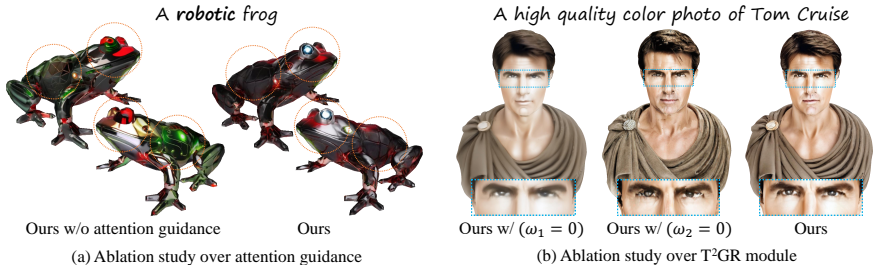
#### 4.4 Quantitative Comparison

**Evaluation Metrics.** For quantitative evaluation of the generated texture, we employ two widely used image quality and diversity evaluation metrics, including Frchet Inception Distance (FID) [13] and Kernel Inception Distance (KID) [3]. These metrics are instrumental in measuring the distribution similarity between two sets of images. For each comparison method, we render a set of images by uniformly sampling 32 different views of the generated textured mesh. To establish a ground truth image set, we follow the approach outlined by Cao *et al.* [5] which used a depth-conditioned ControlNet to synthesize images conditioned on rendered depth maps and corresponding textual prompts. The background pixels have been removed from all images to mitigate the influence caused by unconstrained background. Additionally, we incorporate the CLIPScore metric [12] to assess the congruence and resemblance between the generated images and their associated text prompts. Specifically, for each method, we calculate the average CLIPScore across all rendered images relative to the given text prompts.

We present the quantitative evaluations of the above-mentioned methods on FID, KID and CLIPScore in Tab. 1. Notably, our approach demonstrates superior performance, outstripping the other methods by at least 10.4% in FID and 39.0% in KID. The figures showcase our method’s capability to generate textures that not only are more realistic but also exhibit a wide variety of appearances across diverse objects.

**Table 3:** Ablation study over attention guidance from attention-guided cross-view generation and T<sup>2</sup>GR.

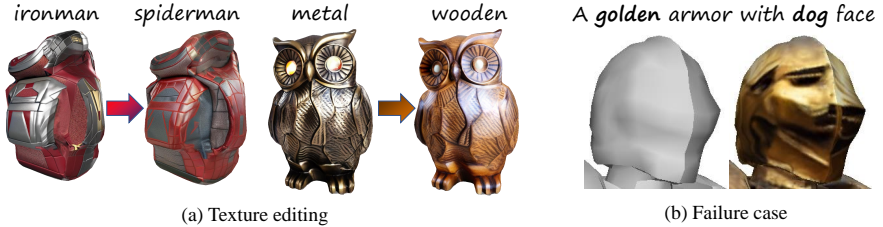
Methods	FID ↓	KID $\times 10^{-3}$ ↓
Ours w/o attention guidance	98.62	5.01
Ours w/ ( $\omega_1 = 0$ )	95.97	4.60
Ours w/ ( $\omega_2 = 0$ )	99.46	5.22
<b>Ours</b>	<b>84.65</b>	<b>4.27</b>

**Fig. 8:** Visual comparison of ablation study over (a) attention guidance from attention-guided cross-view generation and (b) T<sup>2</sup>GR module. Without attention guidance, the frog has different appearance patterns and color tones over different sides.

**User Study.** To analyze the quality of the generated textures and their fidelity to the corresponding text prompts, we conducted a detailed user study of our method against four baseline methods. We randomly select 40 meshes from our collected data and feed them along with a text prompt as the input for each method. For each of these 40 selections, we generate 360° rotating view videos using both our method and one of the baseline methods and display them side-by-side. Participants in the study are then requested to select the video that not only better matched the given caption but also exhibited superior quality. The user study yielded a dataset of 2,480 responses from 62 participants. We report the user preferences in Tab. 2. The results indicate that our method is notably more effective in producing high-quality textures that are preferred by human evaluators.

## 4.5 Ablation Study

We first visually evaluate the impact of our proposed attention-guided cross-view generation as shown in Fig. 8(a). The results demonstrate that our proposed method with attention guidance is able to generate textures which have a consistent appearance in different viewpoints. We also evaluate the impact of the T<sup>2</sup>GR by keeping  $\omega_1 = 0$  and  $\omega_2 = 0$  in Eq. 12, respectively. As shown in Fig. 8(b), the first figure with  $\omega_1 = 0$  lacks high-frequency details and tends to be over-smoothed, while the middle figure with  $\omega_2 = 0$  lacks texture guidance and the assembled texture from all viewpoints is fragmented. We also evaluate the generation quality using FID and KID in Tab. 3, which shows that our method



**Fig. 9:** (a) Applications of our proposed texture sampling strategy for text-driven texture editing. (b) Our method struggles to generate correct textures when text prompts are far from the semantics of the mesh geometry.

with attention-guided cross-view generation and T<sup>2</sup>GR outperforms other variants by a large margin. More ablation studies about the impact of reference views are shown in the supplementary materials.

#### 4.6 Application and Failure Case

Our proposed texture sampling scheme can also be applied to texture editing, as shown in Fig. 9(a). It shares the same pipeline with texture generation, but here we replace the depth-aware ControlNet with the MultiControlNet [40] that combines both the depth-guided and edge-guided generation to preserve the original identity, where the Canny edges are extracted from the generated views. Fig. 9(b) show a failure case of our method caused by significant semantic mismatch between text prompts and mesh geometry.

## 5 Conclusion

In this paper, we present TexGen, a novel texture sampling strategy for text-driven texture generation on 3D meshes, leveraging depth-aware diffusion models. To address the significant challenges in producing textures that are consistent across views and rich in detail, we first propose to maintain a time-dependent texture map that evolves with each denoising step to progressively reduce the view discrepancy. Specifically, at each denoising step, the texture is assembled from the *denoised observations* of sampled views under our attention-guided multi-view sampling process. It is then utilized in our text&texture-guided noise resampling procedure to further guide the estimated noise fed into the next denoising step. The effectiveness of our method is evident in its ability to generate superior-quality textures for diverse 3D objects as well as in its adaptability for texture editing purposes. As for limitations, the overall quality of the generated 3D textures still exhibits a gap when compared to 2D image generation. Striking a balance between 3D consistency and the generation quality of specific views remains a challenge. Additionally, we didn't consider the disentanglement of material and lighting from the generated textures, which is leaved as future work to explore.

**Acknowledgement.** We gratefully acknowledge the support of MindSpore (<https://www.mindspore.cn/>), CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research. Huo and Yang would like to thank the financial support from the Natural Sciences and Engineering Research Council of Canada and the University of Alberta.

## References

1. Three d scans. <https://threedscans.com> (2012)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. *ACM Transactions on Graphics (TOG)* **42**(4), 1–11 (2023)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* (2018)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465* (2023)
5. Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Textfusion: Synthesizing 3d textures with text-guided image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4169–4181 (2023)
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015)
7. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396* (2023)
8. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873* (2023)
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13142–13153 (2023)
10. Guo, Y., Zuo, X., Dai, P., Lu, J., Wu, X., Yan, Y., Xu, S., Wu, X., et al.: Decorate3d: text-driven high-quality texture generation for mesh decoration in the wild. *Advances in Neural Information Processing Systems* **36** (2024)
11. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voletti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. <https://github.com/threestudio-project/threestudio> (2023)
12. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022)

16. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
17. Lei, J., Zhang, Y., Jia, K., et al.: Tango: Text-driven photorealistic and robust 3d stylization via lighting decomposition. *Advances in Neural Information Processing Systems* **35**, 30923–30936 (2022)
18. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 300–309 (2023)
19. Luo, S., Hu, W.: Diffusion probabilistic models for 3d point cloud generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2837–2845 (2021)
20. Metzger, G., Richardson, E., Patashnik, O., Giryas, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12663–12673 (2023)
21. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13492–13502 (2022)
22. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: *SIGGRAPH Asia 2022 conference papers*. pp. 1–8 (2022)
23. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* **41**(4), 1–15 (2022)
24. Nam, G., Khlifi, M., Rodriguez, A., Tono, A., Zhou, L., Guerrero, P.: 3d-ldm: Neural implicit 3d shape generation with latent diffusion models. arXiv preprint arXiv:2212.00842 (2022)
25. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
26. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
28. Richardson, E., Metzger, G., Alaluf, Y., Giryas, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
30. Sanghi, A., Fu, R., Liu, V., Willis, K.D., Shayani, H., Khasahmadi, A.H., Sridhar, S., Ritchie, D.: 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*. pp. 18339–18348 (2023)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
32. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023)



33. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023)
34. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023)
35. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
36. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023)
37. Young, J.: xatlas. In: [github.com/jpcy/xatlas](https://github.com/jpcy/xatlas) (2016)
38. Yu, X., Dai, P., Li, W., Ma, L., Liu, Z., Qi, X.: Texture generation on 3d meshes with point-uv diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4206–4216 (2023)
39. Zeng, X., Vahdat, A., Williams, F., Gojcic, Z., Litany, O., Fidler, S., Kreis, K.: Lion: Latent point diffusion models for 3d shape generation. arXiv preprint arXiv:2210.06978 (2022)
40. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
41. Zhou, L., Du, Y., Wu, J.: 3d shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5826–5835 (2021)