GenesisTex2: Stable, Consistent and High-Quality Text-to-Texture Generation

Jiawei Lu[1,2*], **Yingpeng Zhang**[2*†], **Zengjun Zhao**[2], **He Wang**[3], **Kun Zhou**[1], **Tianjia Shao**[1†]

[1] State Key Lab of CAD&CG, Zhejiang University
[2] Tencent IEG
[3] AI Centre, Computer Science, University College London
lujiawei23@gmail.com, yingpengzhang@gmail.com, zengjunzhao@tencent.com,
he_wang@ucl.ac.uk, kunzhou@zju.edu.cn, tjshao@zju.edu.cn

Abstract

Large-scale text-guided image diffusion models have demonstrated remarkable results in text-to-image (T2I) generation. However, applying these models to synthesize textures for 3D geometries remains challenging due to the domain gap between 2D images and textures on a 3D surface. Early works that used a projecting-inpainting approach managed to preserve generation diversity, but often resulted in noticeable artifacts and style inconsistencies. While recent methods have attempted to address these inconsistencies, they often introduce other issues, such as blurring, over-saturation, or over-smoothing. To overcome these challenges, we propose a novel text-to-texture synthesis framework that takes advantage of pre-trained diffusion models. We introduce a local attention reweighing mechanism in the self-attention layers to guide the model in focusing on spatial-correlated patches across different views, thereby enhancing local details while preserving cross-view consistency. Additionally, we propose a novel latent space merge pipeline, which further ensures consistency across different viewpoints without sacrificing too much diversity. Our method significantly outperforms existing state-of-the-art techniques in terms of texture consistency and visual quality, while delivering results much faster than distillation-based methods. Importantly, our framework does not require additional training or fine-tuning, making it highly adaptable to a wide range of models available on public platforms.

Extended version — https://arxiv.org/abs/2409.18401

1 Introduction

Digital assets are essential for the gaming, film, and animation industries. The role of textures is pivotal, as they influence the visual effects and aesthetics. However, creating appealing textures takes considerable effort, even for professionals. Recently, diffusion models trained on billions of image-text pairs have enabled users to generate stunning images from text prompts. However, applying this approach to texture synthesis faces significant challenges, primarily due to: 1) a lack of high-quality text-labeled training data for textures and 2) a domain gap between 2D images and 3D surface textures. Therefore, most methods of text-guided texture generation circumvent the limitations by employing pretrained 2D text-to-image diffusion models. However, creating 3D consistent textures that maintain high quality remains a significant challenge, even with geometric guidance like Depth maps in ControlNet.

Figure 1: A gallery of high-quality and diverse textures generated by GenesisTex2.

Existing approaches typically navigate a trade-off between single-image quality and multi-view consistency, falling into two main categories. The first group optimizes an underlying 3D structure based on Score Distillation Sampling (Poole et al. 2022; Lin et al. 2023; Wang et al. 2024). However, these optimization-based methods are often time-consuming and struggle to match the diversity and quality of text-to-image generation. The second group generates images from various viewpoints to create the final texture in an optimization-free fashion. This can be achieved through sequential inpainting (Chen et al. 2023b; Richardson et al. 2023) or a multi-view diffusion approach (Liu et al. 2023c; Gao et al. 2024). Our method falls in this category.

We tackle the challenges of achieving both consistency and quality by introducing a cross-view local attention technique and a latent space merge pipeline specifically designed for the text-to-texture task, using only pretrained T2I models. For local attention, we input a 3D mesh and construct dense patch-level weight matrices based on the 3D locations of patches across different views. Patches that are closer in 3D receive higher weights, while farther ones get lower

*These authors contributed equally.
†Corresponding author.

weights. The weight matrices are then incorporated into the self-attention layers during diffusion to amplify or attenuate the effect of specific patches, thereby enhancing local details and improving the consistency of multi-view images.

Additionally, we design a latent space merge framework to ensure consistent and high-quality texture synthesis. Finally, we propose an efficient texture completion algorithm to fill uncolored UV pixels caused by self-occlusion. The algorithm approximates color dilation in surface space by discretizing the UV into sub-UV islands.

Our contributions can be summarized as follows:

- We propose a novel local attention mechanism for pretrained T2I models, which leverages 3D priors and establishes patch correspondences across different views.
- We design a framework that incorporates a latent merge pipeline and an efficient texture dilation algorithm in surface space, enabling a stable generation of consistent and high-quality textures.
- We have conducted extensive evaluations on a variety of 3D objects. The evidence demonstrates that our approach significantly surpasses the performance of the baseline methods by better preserving the generative potential of the original T2I models in aspects of details and color richness while maintaining multi-view consistency.

2 Related Works

2.1 Text-to-Image Diffusion Models

Diffusion models are a class of generative models that use Markov chains to transform random noise into high-quality visuals sequentially. A pioneering work, GLIDE (Nichol et al. 2021), is the first to employ diffusion models for generating images in pixel space while supporting text conditioning by adopting classifier-free guidance. Following GLIDE (Nichol et al. 2021), Imagen (Saharia et al. 2022) integrates diffusion models for high-resolution text-guided image generation. DALLE-2 (Ramesh et al. 2022) leverages CLIP (Radford et al. 2021), a popular model that aligns texts and images to generate images from CLIP latent space. Stable diffusion is a landmark work built upon Latent Diffusion Model (LDM) (Rombach et al. 2022) trained on a large-scale text-image dataset (Schuhmann et al. 2022), which proposes to adapt the diffusion process in latent space to further reduce computational cost. Besides text conditioning, various flexible conditions have been introduced for image generation such as ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2024). These control methods aim to generate results that align with a given spatial condition, such as depth or normal images, which can be either predicted from input images or rendered from 3D meshes, supporting mesh-guided image generation.

2.2 Text-driven 3D Generation

Many recent studies (Jun and Nichol 2023; Hong et al. 2023; Huang et al. 2023; Xu et al. 2023; Nichol et al. 2022) attempt to replicate the success of 2D diffusion models in text-guided 3D content generation, after the supervision of text-paired 3D data. A common constraint of these methods lies in the scarcity of publicly available labeled 3D data. As such, rather than direct learning a 3D diffusion model, many works resort to using pretrained 2D image diffusion models for 3D tasks (Gao et al. 2024; Cao et al. 2023; Chen et al. 2023b; Liu et al. 2024, 2023a; Long et al. 2023; Shi et al. 2023). Pioneering works (Poole et al. 2022; Wang et al. 2023) suggest optimizing a 3D representation (e.g., NeRF) by distilling from 2D diffusion models. Subsequent research (Lin et al. 2023; Metzer et al. 2023) further improved these text-to-3D distillation methods in various aspects. A recent remarkable work (Wang et al. 2024) proposed a technique called Variational Score Distillation (VSD) that further enriches the details and diversity. Another line of work (Shi et al. 2023; Liu et al. 2023b; Tsalicoglou et al. 2023) typically fine-tune a multi-view diffusion model by incorporating camera directions to image diffusion models and simultaneously generate multi-view images. Zero-1-to-3 (Liu et al. 2023a) first attempts to leverage 3D data and camera parameters to fine-tune pretrained 2D diffusion models for 3D-consistent novel view synthesis. MVDream (Shi et al. 2023) and SyncDreamer (Liu et al. 2023b) share a similar idea to improve consistency by fine-tuning attention layers in 2D diffusion models using 2D and 3D data.

2.3 Mesh-guided Texture Synthesis

Beyond generating 3D objects using text prompts, creating textures for given meshes is also a critical and challenging task with various applications. Initial studies (Oechsle et al. 2019; Siddiqui et al. 2022; Yu et al. 2021; Chen, Yin, and Fidler 2022) have shown promising results using GANs. However, their application is limited to specific categories. In contrast, many recent works on mesh-guided text-to-texture synthesis have achieved broader applicability by leveraging large-scale pretrained diffusion models. These methods typically employ strategies such as sequential generation and inpainting (Chen et al. 2023b; Richardson et al. 2023; Cao et al. 2023), multi-view diffusion (Gao et al. 2024; Liu et al. 2023c) or score distillation (Chen et al. 2023a; Metzer et al. 2023; Youwang, Oh, and Pons-Moll 2023).

3 Method

Given a mesh $\mathcal{M}$ and a textual prompt $\mathcal{P}$, our goal is to produce a texture $\mathcal{T}$ that accurately depicts the prompt and suits the shape with high quality. An overview of our pipeline is shown in Fig. 2. In this section, we first introduce preliminaries on image space diffusion models and define notations for rendering. Next, we provide details on how to adapt local attention to the diffusion process to improve the local details in the generated images while preserving consistency. Then, we illustrate our latent merge pipeline, which is combined with the local attention mechanism and ensures the consistency. The final texture can be obtained by inverse rendering and merging the generated multi-view images.

3.1 Preliminary

2D Image Diffusion models In this paper, we employ Stable Diffusion (Rombach et al. 2022). Stable Diffusion is a latent diffusion model that operates in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, where $\mathcal{E}$ and $\mathcal{D}$ represent the encoder

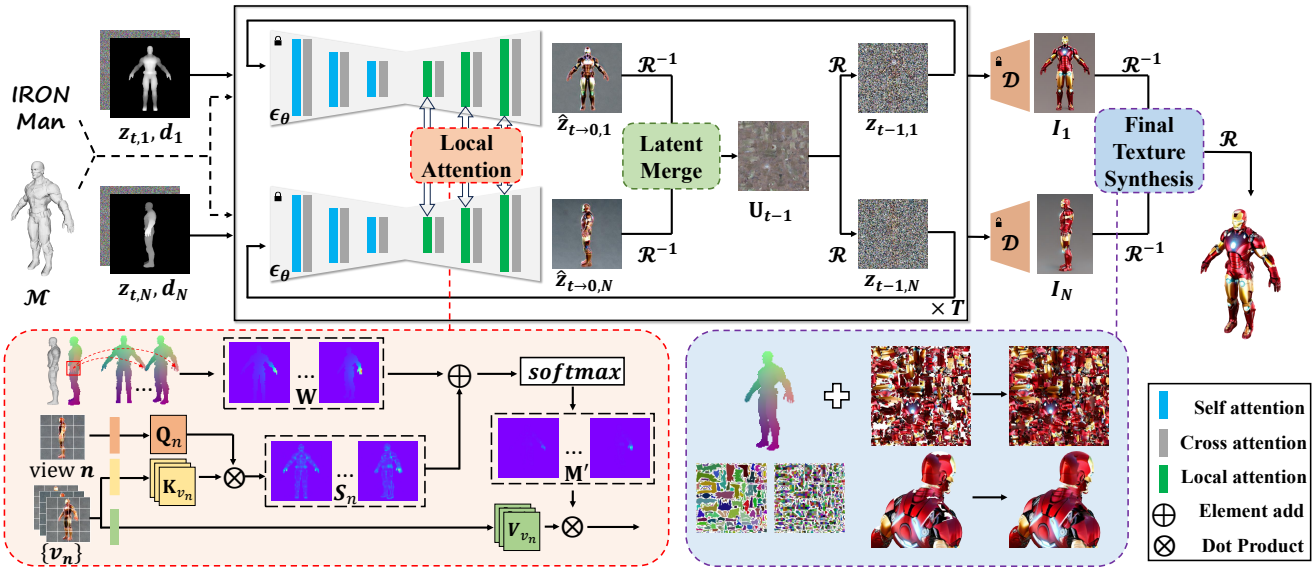

Figure 2: Given a mesh and a textual prompt, we aim to produce textures that well depict the prompt and suit the shape. To achieve this, we propose a local attention technique in Sec. 3.2, which enhances local details by reweighing the original self-attention layers based on the 3D shape. In addition, we introduce a framework for consistent texture synthesis in Sec. 3.3, enabling the stable generation of consistent and high-quality textures.

and decoder, respectively. For a given image I with its corresponding latent feature $\mathbf{z}_0 = \mathcal{E}(I)$, the DDPM forward process(Ho, Jain, and Abbeel 2020) iteratively adds Gaussian noise to $\mathbf{z}_0$.

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (1)$$

where $t = 1, ..., T$ is the time step, $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ is the conditional density of $\mathbf{z}_t$ given $\mathbf{z}_{t-1}$, and α_t is hyperparameter. In the DDPM backward process, a U-Net ϵ_θ is trained to predict the noise and $\mathbf{z}_{t-1}$ can be sampled based on $\mathbf{z}_t$ and prompt $\mathcal{P}$:

$$\mathbf{z}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{z}}_{t\to0} + \frac{(1-\bar{\alpha}_{t-1})(\sqrt{\alpha_t}\mathbf{z}_t + \beta_t\varepsilon_t)}{1-\bar{\alpha}_t}, \quad (2)$$

where α_t and $\beta_t = 1 - \alpha_t$ are pre-defined hyperparamters, $\hat{\mathbf{z}}_{t\to0}$ is the denoised estimation at time step t, $\epsilon_\theta(\mathbf{z}_t, t, \mathcal{P})$ is the predicted noise for $\mathbf{z}_t$, and $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$. We can sample $\mathbf{z}_0$ by iteratively performing denoising using Eq. 2 from the standard Gaussian noise $\mathbf{z}_T$ with DDPM sampling. The final generated image is then decoded by $\mathcal{D}(\mathbf{z}_0)$.

Rendering Representation In this paper, textures are defined in 2D image space using an injective UV parameterization of $\mathcal{M}$, represented as $UV : p \in \mathcal{M} \mapsto (u, v) \in [0, 1]^2$. This parameterization can be automatically constructed using tools like *xatlas* (Young 2016). We focus on synthesizing base color maps and disregard any shading effects. Given a mesh $\mathcal{M}$, a texture map $\mathcal{T}$ and a viewpoint $\mathbf{C}$, we use the rendering function $\mathcal{R}$ to get the rendered image $\mathbf{x} = \mathcal{R}(\mathcal{T}; \mathcal{M}, \mathbf{C})$. Conversely, the inverse rendering function $\mathcal{R}^{-1}$ is utilized to reconstruct the texture map from the rendered image: $\mathcal{T}' = \mathcal{R}^{-1}(\mathbf{x}; \mathcal{M}, \mathbf{C})$. For simplicity, we omit $\mathcal{M}$ and $\mathbf{C}$ for $\mathcal{R}$ and $\mathcal{R}^{-1}$ throughout this paper.

3.2 Local Attention

The attention layer is crucial in Stable Diffusion, featuring two types of attention mechanisms: 1) cross-attention, which measures the similarity between the latent features and text embeddings, and 2) self-attention, which can be viewed as patch matching and voting within a single image. In Stable Diffusion, each self-attention layer receives the deep spatial feature $\phi(\mathbf{z}_t)$ of the noisy latent $\mathbf{z}_t$, and linearly projects $\phi(\mathbf{z}_t)$ to the query, key, and value matrices $\mathbf{Q} = l_Q(\phi(\mathbf{z}_t))$, $\mathbf{K} = l_K(\phi(\mathbf{z}_t))$, $\mathbf{V} = l_V(\phi(\mathbf{z}_t))$, where l_Q, l_K, l_V are pre-trained linear networks for feature projection. The output of self-attention layers is given by $\mathit{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}) \cdot \mathbf{V}$, where d is a constant representing the dimension of deep features, we omit $\sqrt{d}$ for simplicity in this paper.

Previous works in zero-shot video editing(Yang et al. 2023, 2024; Khachatryan et al. 2023) have demonstrated that modifying the self-attention layers to incorporate cross-frame attention can help regularize style across multiple frames. In texture synthesis, a similar strategy for improving style consistency involves using features from other views as keys and values to perform cross-view attention, as in (Gao et al. 2024; Liu et al. 2023c). The cross-view attention for view n can be written as:

$$\mathit{cross\_view\_attn}(n) = \mathit{Softmax}(\mathbf{Q}_n\mathbf{K}_{v_n}^T)\mathbf{V}_{v_n}, \quad (3)$$

where v_n is a set of views that attend to the query view n. The *cross_view_attn* behaves as the original self-attention when v_n contains only n.

However, directly adopting this strategy in the diffusion process often leads to a decrease in color diversity and local details in the generated images, as demonstrated in Fig. 4.

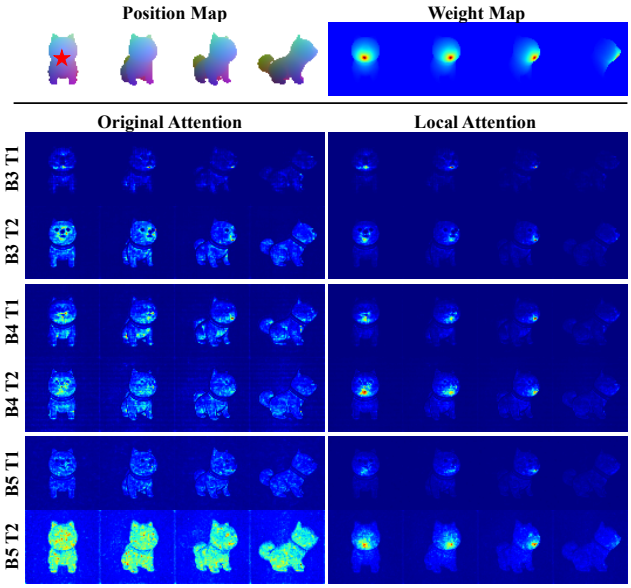

Figure 3: A visualization of attention maps concerning the query patch (in red star). The upper part illustrates the rendered position map and calculated weight map. The bottom part shows the attention map of different layers before and after reweighed by the weight map. $B\{i\}T\{j\}$ stands for the i-th Block and j-th Transformer layer in the output layers.

The root cause of the degradation lies in a reduction of variance in the cross-view attention mechanism, as the predicted feature embedding with the same underlying 3D structure can vary when viewed from different perspectives. This can result in a large attention weight for irrelevant patches, as illustrated in the visualization of attention maps in Fig. 3. In this situation, it becomes necessary to guide the attention module to give greater weight to the same surface area across different viewpoints. This requires considering the correlation of patches among multiple views. Fortunately, we have the input 3D proxy in the texture synthesis task, which naturally builds a strong semantic correspondence between patches of different views.

Inspired by (Hertz et al. 2022), which enables prompt-based image editing by modifying the cross-attention layers in diffusion models, we introduce an attention bias matrix $\mathbf{W}$ to reweigh the original attention produced by the pre-trained self-attention layers in Stable Diffusion. Similar to the attention mask mechanism that masks certain words in the cross-attention layers, $\mathbf{W}$, in our case, is used to emphasize or diminish the correlation between specific pairs of query-key patches within the self-attention layers. Unlike the previously mentioned cross-view ***global*** attention, we refer to our approach as cross-view ***local*** attention.

We now define the process for calculating the attention bias $\mathbf{W}$. Without loss of generality, let us consider the local attention of the n-th query view with attended views denoted as v_n. For simplicity, we will omit the subscript n until the end of this section. We render a set of position maps $\{\mathcal{O}\}$ by applying the rendering function $\mathcal{R}(\mathcal{V})$ to each view in v, where $\mathcal{V}$ denotes the vertex position of $\mathcal{M}$. Then, we calculate a distance matrix $\mathbf{d}$ based on the rendered position maps $\{\mathcal{O}\}$. Each entry of $\mathbf{d}$ can be calculated using Euclidean distance: $\mathbf{d}_{i,j} = \|\mathcal{O}_i - \mathcal{O}_j\|$ for any location $i \in \{1, ..., N_Q\}$ and $j \in \{1, ..., N_K\}$, where N_Q and N_K stands for the number of patches in query and key features, respectively. We do not use geodesic distance due to its significant computational cost, particularly for meshes with a large number of vertices. Furthermore, the precision of the distance calculations is inherently limited by the low resolution of the attention maps, making the choice of distance calculation method less critical.

Then, we compute $\mathbf{W}$ by:

$$\mathbf{W}_{i,j} = \begin{cases} 0, & \textit{if} \quad Q_i \in \textit{BG} \cap K_j \in \textit{BG} \\ -o\ln(1 + r\mathbf{d}_{i,j}), & \textit{if} \quad Q_i \in \textit{FG} \cap K_j \in \textit{FG} \\ -\infty, & \textit{else} \end{cases} \tag{4}$$

where o and r are hyper-parameters that determine the distribution of the attention bias, *BG* and *FG* refer to background and foreground patches, respectively. Intuitively, the attention bias approaches 0 for patch pair located at the same position in 3D and attenuates towards $-\infty$ as the distance increases. We do not reweigh attention between background patches, and to avoid extreme cases, we set a lower bound δ by applying a clamping operation: $\mathbf{W} = max(\mathbf{W}, \ln(\delta))$.

Given the original similarity $\mathbf{S} = \mathbf{Q}\mathbf{K}_v^T$ and attention bias $\mathbf{W}$, we can compute the reweighed attention matrix as follows:

$$\mathbf{M}' = \textit{Softmax}(\mathbf{S} + \mathbf{W}), \tag{5}$$

where each element $\mathbf{M}'_{i,j}$ is calculated by:

$$\mathbf{M}'_{i,j} = \frac{e^{\mathbf{W}_{i,j}} e^{\mathbf{S}_{i,j}}}{\sum_j e^{\mathbf{W}_{i,j}} e^{\mathbf{S}_{i,j}}}. \tag{6}$$

In this way, we manage to manipulate the attention maps by emphasizing on the correspondence of feature patches that are closer in 3D. We empirically find it helpful to replace the original similarity with the weight matrix, to improve the local consistency, i.e., $\mathbf{M}' = \textit{Softmax}(\mathbf{W})$. However, the replacement operation can lead to blurring and shape distortion in the late steps. Therefore, we limit the replacement strategy to the early stages of the diffusion process for rough consistency guidance.

3.3 Consistent Texture Synthesis

Latent merge pipeline Applying cross-view local attention in the diffusion process can improve the style consistency across different views, but it is insufficient for synthesizing 3D consistent views, i.e., two pixels projected to the same point in 3D have the same value. Directly merging these views will inevitably cause inconsistencies in the final texture, as shown in the first two rows of Fig. 6. We consider a latent space alignment strategy similar to (Liu et al. 2023c; Gao et al. 2024; Kim et al. 2024) for better cross-view consistency. However, the alignment operation can lead to an over-smoothed appearance and degradation in diversity due to a loss of variation in the alignment process, see Fig. 5. To

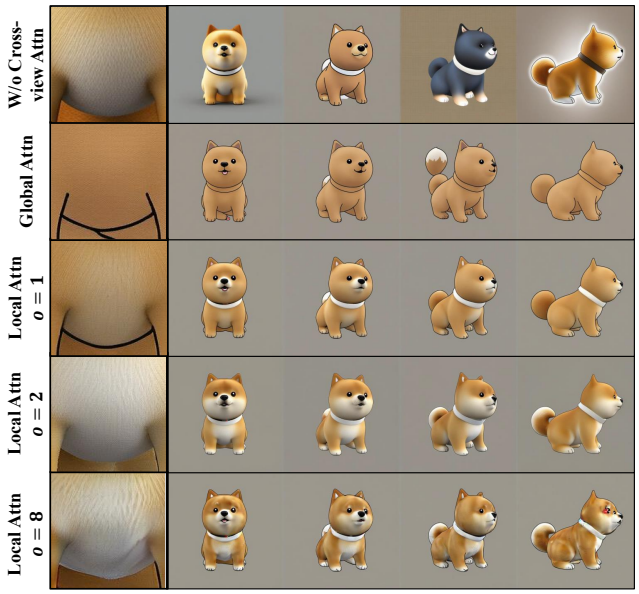

Figure 4: Results of different attention mechanisms for 4-view diffusion with prompt: *A cute shiba inu dog*. Images in row 1 are generated without cross-view attention and exhibit no consistency. Results using Global Attention (row 2) are consistent but lose color diversity and details. Images with Local Attention (row 3-5) show improvements in diversity and details, all while maintaining a significant level of cross-view consistency. We find that setting $o = 2$ achieves best diversity and quality.

overcome these issues while maintaining view consistency, we introduce a novel latent merge pipeline.

Specifically, we first initialize a set of noisy latents for each view by $\{\mathbf{z}_{T,n} \sim \mathcal{N}(0, \mathbf{I})\}_{n=1}^{N}$ and an initial latent texture $\mathbf{U}_T \sim \mathcal{N}(0, \mathbf{I})$ at the beginning of the denoising process. At each denoising step t, our goal is to predict a 3D consistent $\mathbf{z}_{t-1,n}$ from $\mathbf{z}_{t,n}$. We first obtain the denoised prediction $\hat{\mathbf{z}}_{t\to 0,n}$ in image space by:

$$\hat{\mathbf{z}}_{t\to 0,n} = (\mathbf{z}_{t,n} - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{z}_{t,n}, t, \mathcal{P}, d_n))/\sqrt{\alpha_t}, \quad (7)$$

where d_n is the depth condition for ControlNet at view n.

We then apply inverse rendering to obtain the per-view partial latent textures by:

$$\hat{\mathbf{U}}_{t\to 0,n} = \mathcal{R}^{-1}(\hat{\mathbf{z}}_{t\to 0,n}). \quad (8)$$

Note that the partial textures do not exhibit 3D consistency at this time. One way is to aggregate them into a canonical one by averaging. However, trivially averaging the partial textures of different views can lead to a loss of high-frequency details and color diversity. Hence, we propose to merge them in a view-dependent way:

$$\hat{\mathbf{U}}_{t\to 0} = \frac{\sum_{n=1}^{N} \omega_{t,n}\mathcal{R}^{-1}(\mathbf{N}_n) \odot \hat{\mathbf{U}}_{t\to 0,n}}{\sum_{n=1}^{N} \omega_{t,n}\mathcal{R}^{-1}(\mathbf{N}_n)}, \quad (9)$$

where $\mathbf{N}_n$ is the cosine similarity map rendered at view point C_n with each pixel representing the cosine similarity between the normal vector of the 3D point and the reversed view direction. The term $\omega_{t,n}$ denotes the weight of view n at time step t. $\omega_{t,n}$ is set to 1 at time step T and is then linearly interpolated to $\max(|\cos\theta|^\gamma, \omega_{min})$ at time step t', where θ is the angle between C_n and C_0, and γ is a hyperparameter that balances the influence of different views. Intuitively, this approach ensures that at the beginning of the diffusion process, different views are merged with similar weights, promoting style consistency. As diffusion progresses, each texel becomes predominantly influenced by a single view, effectively preserving diversity and preventing the loss of high-frequency details.

After merging the denoised partial textures into a single one, we can update the latent texture $\mathbf{U}_{t-1}$ by adding back the variance with Eq. 2:

$$\mathbf{U}_{t-1} = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\hat{\mathbf{U}}_{t\to 0} + \frac{(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}(\sqrt{\alpha_t}\mathbf{U}_t + \beta_t\varepsilon_t). \quad (10)$$

The image space latent $\mathbf{z}_{t-1,n}$ for next step of $t-1$ can be then obtained by blending the rendered foreground latent $\mathcal{R}(U_{t-1}, C_k)$ with the image space latent $\hat{\mathbf{z}}_{t-1,n}$:

$$\mathbf{z}_{t-1,n} = \mathbf{M}_n \odot \mathcal{R}(\mathbf{U}_{t-1}; C_n) + (1-\mathbf{M}_n) \odot \hat{\mathbf{z}}_{t-1,n}, \quad (11)$$

where $\hat{\mathbf{z}}_{t-1,n}$ can be derived by Eq. 2, and $\mathbf{M}_n$ represents the binary foreground mask for viewpoint C_n.

The final denoised $\mathbf{z}_{0,n}$ of each view can be obtained by iterating the denoising steps. To alleviate aliasing caused by the reprojection of low-resolution latents, we do not perform latent merging in the last 5 steps.

Final Texture Synthesis To reconstruct the texture map, we first decode the latent of each viewpoint to generate multi-view images $\mathcal{I}_n$ by $\mathcal{D}(\mathbf{z}_{0,n})$. Subsequently, we finalize the texture by:

$$\mathcal{T}_{merge} = \frac{\sum_{n=1}^{N} \omega_n\mathcal{R}^{-1}(\mathbf{N_n}) \odot \mathcal{R}^{-1}(\mathcal{I}_n)}{\sum_{n=1}^{N} \omega_n\mathcal{R}^{-1}(\mathbf{N_n})}, \quad (12)$$

where $\mathbf{N_n}$ is the similarity mask at viewpoint C_n and $\omega_n = \max(|\cos\theta|^\gamma, \omega_{min})$.

After merging, the texture map still contains invalid pixels that fail to receive color from any perspective due to self-occlusion. A straightforward approach to address this issue is to expand the valid pixels on the texture map using a flood-fill technique within the image space. However, this naive flood-fill method may propagate colors from pixels that are not adjacent in the 3D space, leading to inaccuracies in the final texture map. An optimal solution involves using geodesic distance, but the computational cost is prohibitively high. Therefore, we introduce a fast texture completion method that approximates color propagation in surface space. The detailed algorithm could be found in the supplement.

4 Experiments

4.1 Implementation Details

We test our method on an NVIDIA A800 GPU, and the entire process was able to finish in 1 minute. The diffusion

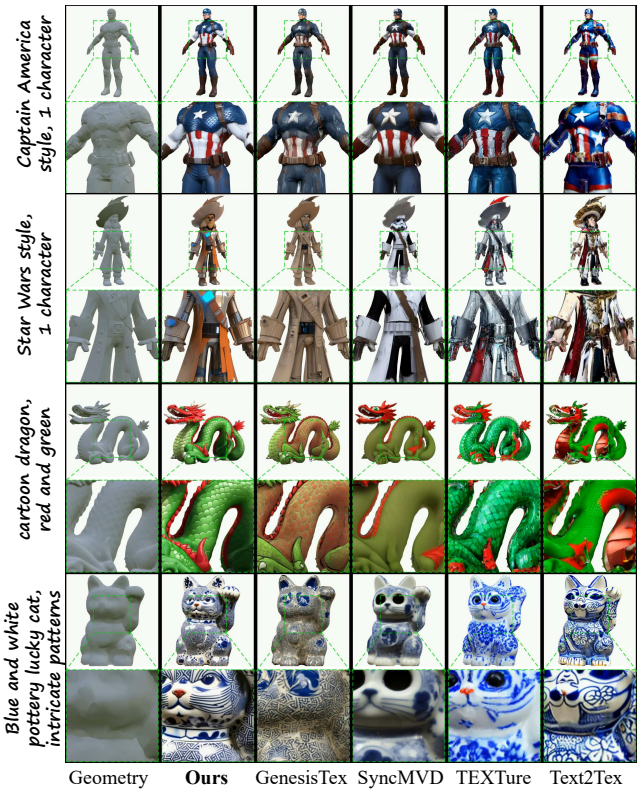

Figure 5: Qualitative comparison with different baselines.

process takes around 50s with 25 denoising steps at a resolution of 1280, and the final texture synthesis stage takes around 2s. The CFG scale is set to 12. We linearly interpolate the view-dependent weight ω for the first 8 steps. The parameters γ and ω_{min} are set to 8 and $1e-3$. We adopt SDXL (Podell et al. 2023) as our base model and ControlNet-Depth (Zhang, Rao, and Agrawala 2023) trained for SDXL for spatial control. We replace the self-attention layers in the output layers of SDXL with our proposed 3D-aware local attention mechanism, and we set o, r, δ as $2, 20$ and 0.1 to achieve the best performance.

Dataset The dataset for evaluation contains 35 meshes with 63 mesh-prompt pairs. These meshes are collected from the publicly open dataset, including objarverse (Deitke et al. 2023), shapenet (Chang et al. 2015), and Stanford 3D Scanning Repository (Turk and Levoy 1994). We use *Xatlas* (Young 2016) to automatically unwrap the UV for all meshes. We normalize all meshes to the range of $[-0.5, 0.5]$ and position the camera at a distance of 2 meters with the field of view set to 35 degrees. To balance time-cost and view coverage, we typically employ $N = 8$ fixed viewpoints at angles of $[0, 45, 90, 135, 180, 225, 270, 315]$ degrees, evenly distributed around the object of interest.

4.2 Comparisons

We conduct comparisons with four available methods for text-to-texture synthesis, including Text2Tex (Chen et al. 2023b), TEXTure (Richardson et al. 2023), SyncMVD (Liu et al. 2023c), GenesisTex (Gao et al. 2024). We have also compared our method with Meshy-3 (Meshy 2024), a state-of-the-art commercial software that supports generating textures for 3D models using text prompts. The comparison results with Meshy-3 are placed in the supplement. We strongly recommend that the reader check the appendix for more details.

Method	PS ↑ (%)	FID↓	KID↓	User study (%) D ↑	C ↑	Q ↑
Text2Tex	9.7	88.1	14.2	14.3	4.5	7.0
TEXTure	10.2	92.2	17.1	13.3	5.7	6.7
GenesisTex	17.1	77.0	9.5	10.8	12.7	11.4
SyncMVD	13.6	85.7	10.2	3.5	5.7	8.9
Ours	**49.4**	**66.4**	**7.3**	**58.1**	**71.4**	**66.0**

Table 1: Quantitative comparisons with baseline methods. Pick Score (PS), Diversity (D), Consistency (C), and Quality (Q). KID values are $\times 10^{-3}$.

Qualitative comparisons. We compare qualitatively with different baselines in Fig. 5. GenesisTex (Gao et al. 2024) produces visually reasonable renderings, but it tends to generate less diverse images. TEXTure (Richardson et al. 2023) and Text2Tex (Chen et al. 2023b) lack multi-view consistency since they operate on each view independently. SyncMVD (Liu et al. 2023c) yields visually consistent renderings. However, it tends to produce blurry results, as seen in the dragon and lucky cat in Fig. 5, since the latent averaging operation in their approach leads to a loss of high-frequency details and color diversity.

Quantitative comparisons. Following GenesisTex (Gao et al. 2024) and TexFusion (Cao et al. 2023), we report FID (Heusel et al. 2017) and KID (Bińkowski et al. 2018) scores. We generate depth maps as conditional images for all meshes by rendering them from 12 different viewpoints, each separated by 30-degree intervals. Using these depth maps and our textual prompts, we sample from pretrained image diffusion model to create a set of ground truth images. Additionally, we render meshes with textures generated by different methods using the same views to get the candidate set. We primarily focus on the foreground and set the background pixels of all images to white.

In addition, we also employ *Pick Score* (Kirstain et al. 2024) to evaluate the visual quality of our texture synthesis results. *Pick Score* is a CLIP-based scoring function trained on large-scale user preference regarding generated images paired with text prompts. For each mesh, we compute the average Pick Score using the same 12-view rendered images employed for calculating the FID, identifying the method with the highest score as the winning approach for that mesh and calculating the winning rate for each method.

We also conducted a user study to analyze the results in three aspects: 1) consistency, 2) diversity, and 3) overall quality. We rendered the results of different methods into videos that show the textured object from a 360° rotating view. We randomly selected 15 meshes for each

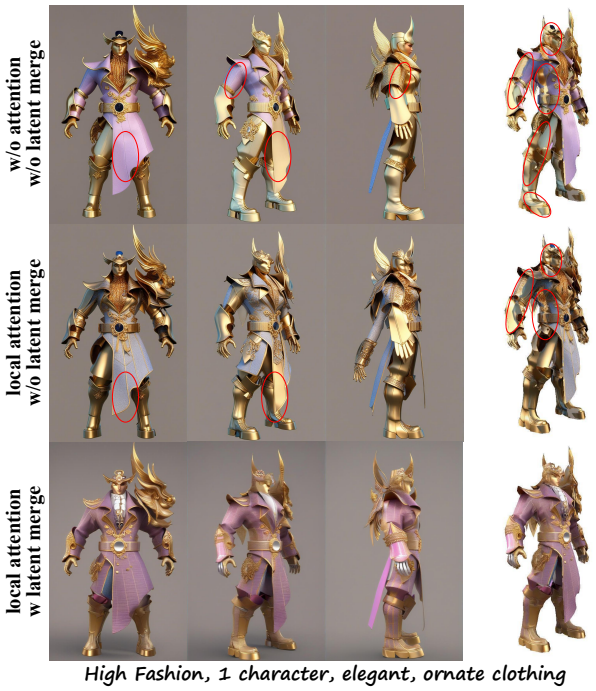

Figure 6: Ablation results on local attention and latent merge. The left three columns show the generated images, and the last column depicts the rendered result with synthesized texture.

questionnaire and asked the participants to judge which method matches best for each aspect. Finally, we collected 30 valid responses from both professional artists and non-professionals. The complete quantitative results can be found in Tab 1. Our method achieves the highest pick score and is preferred by most human evaluators in terms of consistency, diversity, and overall quality.

4.3 Ablation Studies

Effectiveness of local attention To investigate the impact of the cross-view local attention, we visualize the decoded multi-view images of different attention strategy in Fig. 4 and Fig. 6. Fig. 4 illustrates an example with the prompt *A cute shiba inu dog*. We can discover that the color and pattern of the dog varies a lot across different viewpoints without any cross-view constraint. With global attention, the query view attends to all views in the attention layer and brings higher consistency, but at a cost of losing image details and variance. Our proposed geometry-aware local attention amplifies the local attention on pixels that are closer in 3D, which not only leads to vivid color and fine-grained details, but also preserves cross-view consistency. Similar in Fig. 6, the cross-view images are more consistent with local attention than the baseline without cross-view attention.

Effectiveness of latent merge pipeline We ablate the latent merge pipeline to evaluate the effectiveness of our latent merge strategy in generating consistent textures. As shown in the last column of Fig. 6, the full pipeline with latent merge exhibits the best consistency compared with baselines in the final renderings. Note how the full method achieves the best multi-view consistency and generates rich details, while the baselines without latent merge exhibit severe inconsistencies.

4.4 More Applications

Our method is designed to be fully compatible with existing Stable Diffusion models without the need for additional training. This makes it readily applicable to a wide range of models available on platforms such as Civitai (civitai 2024) and HuggingFace (Huggingface 2024). Furthermore, our pipeline can be seamlessly integrated with auxiliary models tailored for Stable Diffusion, thereby enriching its versatility in practical scenarios. For instance, we can incorporate the IP-Adapter into our framework to facilitate image-guided texture synthesis, and leverage various LoRAs to achieve distinct artistic styles. The results with LoRAs and IP-Adapters can be found in the supplement.

5 Discussions

Failure Cases Our algorithm employs texture dilation to fill the fully-occluded regions, which may sometimes produce overly smoothed results in areas that should have complex textures. Additionally, the Janus effect is a challenge inherent to methods that utilize pretrained 2D image diffusion models. While this issue is alleviated through the proposed local attention and perspective prompts (as seen in DreamFusion), the inherent bias presented in 2D image diffusion models can still result in unwanted anatomical features.

Limitation A common limitation in the field of texture synthesis using pretrained 2D diffusion models is the imperfect alignment between the mesh and texture. This issue is largely due to the limited control capabilities of the currently available ControlNets. It could be improved along with the development of more powerful control models. Another common limitation in this field is the baked-in lighting effect, which we plan to address in future work.

6 Conclusions

In this article, we propose a pipeline designed to generate consistent and high-quality textures for 3D meshes using textual prompts. Our method leverages pretrained Stable Diffusion models without any further training or fine-tuning. This makes it highly versatile, capable of handling a wide range of geometry and texture types, and easily adaptable to various models on model-sharing platforms. We believe this work will advance AI-based texturing and open up new possibilities for 3D content generation.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments. This work is supported by NSF China under Grants 62322209 and 62421003, and the 100 Talents Program of Zhejiang University.

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Cao, T.; Kreis, K.; Fidler, S.; Sharp, N.; and Yin, K. 2023. TexFusion: Synthesizing 3D Textures with Text-Guided Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4169–4181.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, D. Z.; Li, H.; Lee, H.-Y.; Tulyakov, S.; and Nießner, M. 2023a. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. *arXiv preprint arXiv:2311.17261*.

Chen, D. Z.; Siddiqui, Y.; Lee, H.-Y.; Tulyakov, S.; and Nießner, M. 2023b. Text2Tex: Text-driven Texture Synthesis via Diffusion Models. *arXiv preprint arXiv:2303.11396*.

Chen, Z.; Yin, K.; and Fidler, S. 2022. Auv-net: Learning aligned uv maps for texture transfer and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1465–1474.

civitai. 2024. civitai — The Home of Open-Source Generative AI. https://civitai.com/.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Gao, C.; Jiang, B.; Li, X.; Zhang, Y.; and Yu, Q. 2024. GenesisTex: Adapting Image Denoising Diffusion to Texture Space. *arXiv preprint arXiv:2403.17782*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Huang, T.; Zeng, Y.; Dong, B.; Xu, H.; Xu, S.; Lau, R. W.; and Zuo, W. 2023. Textfield3d: Towards enhancing open-vocabulary 3d generation with noisy text fields. *arXiv preprint arXiv:2309.17175*.

Huggingface. 2024. Huggingface. https://huggingface.co/.

Jun, H.; and Nichol, A. 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15954–15964.

Kim, J.; Koo, J.; Yeo, K.; and Sung, M. 2024. SyncTweedies: A General Generative Framework Based on Synchronized Diffusions. *arXiv preprint arXiv:2403.14370*.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2024. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9298–9309.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023b. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Liu, Y.; Xie, M.; Liu, H.; and Wong, T.-T. 2023c. Text-Guided Texturing by Synchronized Multi-View Diffusion. *arXiv preprint arXiv:2311.12891*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2023. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*.

Meshy. 2024. Meshy — 3D AI Generator. https://www.meshy.ai/. Accessed: 2024-07-20.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12663–12673.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(5), 4296–4304.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nichol, A.; Jun, H.; Dhariwal, P.; Mishkin, P.; and Chen, M. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*.

Oechsle, M.; Mescheder, L.; Niemeyer, M.; Strauss, T.; and Geiger, A. 2019. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4531–4540.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; and Cohen-Or, D. 2023. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Siddiqui, Y.; Thies, J.; Ma, F.; Shan, Q.; Nießner, M.; and Dai, A. 2022. Texturify: Generating textures on 3d shape surfaces. In *European Conference on Computer Vision*, 72–88. Springer.

Tsalicoglou, C.; Manhardt, F.; Tonioni, A.; Niemeyer, M.; and Tombari, F. 2023. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439*.

Turk, G.; and Levoy, M. 1994. Zippered polygon meshes from range images. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, 311–318.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12619–12629.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Xu, Y.; Tan, H.; Luan, F.; Bi, S.; Wang, P.; Li, J.; Shi, Z.; Sunkavalli, K.; Wetzstein, G.; Xu, Z.; et al. 2023. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*.

Yang, S.; Zhou, Y.; Liu, Z.; ; and Loy, C. C. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *ACM SIGGRAPH Asia Conference Proceedings*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2024. FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. *arXiv preprint arXiv:2403.12962*.

Young, J. 2016. xatlas. In *github.com/jpcy/xatlas*.

Youwang, K.; Oh, T.-H.; and Pons-Moll, G. 2023. Paint-it: Text-to-Texture Synthesis via Deep Convolutional Texture Map Optimization and Physically-Based Rendering. *arXiv preprint arXiv:2312.11360*.

Yu, R.; Dong, Y.; Peers, P.; and Tong, X. 2021. Learning texture generators for 3d shape collections from internet photo sets. In *British Machine Vision Conference*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.