DreamPBR: Text-driven High-Resolution SVBRDF Generation with Multimodal Guidance

Linxuan Xin[1], Zheng Zhang[2], Zhiyi Pan[1], Jinfu Wei[3], Duan Gao[2†], Wei Gao[1†]
[1]School of Electronic and Computer Engineering, Peking University, China
[2]Huawei Cloud Computing Technologies Co., Ltd., China
[3]Shenzhen International Graduate School, Tsinghua University, China

***Abstract*—Existing material creation methods are limited in diversity due to the scarcity of real-world data. To enhance controllability and diversity, we propose DreamPBR, a diffusion-based generative framework that creates spatially varying appearance properties guided by text and multimodal controls. By integrating large-scale vision-language models trained on billions of text-image pairs with material priors from hundreds of Physically Based Rendering (PBR) samples, we achieve high-quality PBR material generation. We employ a material Latent Diffusion Model (m-LDM) to map albedo maps to latent space, which is then decoded into full Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF) parameter maps via a rendering-aware PBR decoder. To achieve diverse control, we introduce a multimodal guidance module that includes image and 3D shape guidance. We demonstrate DreamPBR's effectiveness in material creation, showcasing its versatility and user-friendliness across various controllable generation and editing applications.**

***Index Terms*—Physically Based Rendering, SVBRDF, Multimodal Deep Generative Model, Deep Learning**

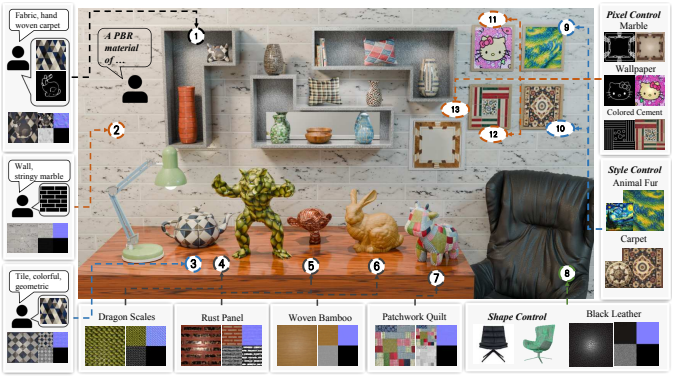

Fig. 1. DreamPBR, an innovative material generation framework, enables personalized creation with multimodal controls. We present various controls such as text descriptions (4, 5, 6, 7), binary images (2, 9, 10), RGB images (3, 11, 12, 13), segmented geometry (8), and their combination (1) in this figure. The high-quality and tileable textures from DreamPBR show high applicability in different objects.

I. Introduction

High-quality materials are crucial for photorealistic rendering. Although appearance modeling has advanced significantly, generating new materials remains challenging. Reconstruction-based methods estimate surface reflectance from photographs using either optimization-based inverse rendering [1]–[3] or deep neural inference [4], [5]. However, these rely on real-world images, limiting their capacity for creative material generation.

Generation-based works [2], [6] apply Generative Adversarial Networks (GANs) [7] to material generation, but their datasets comprise only hundreds to thousands of materials—far fewer than the billions in large-scale language-image models—limiting diversity. They also face challenges like unstable training, mode collapse, and poor scalability. Therefore, [8], [9] provide initial validation of the potential of diffusion models for 3D content generation. However, they mainly produce implicit representations or textured meshes, lacking the capacity to disentangle physically based materials and illumination.

To address these challenges, we present **DreamPBR**, a novel generative framework for creating high-resolution spatially varying bidirectional reflectance distribution functions (SVBRDFs) conditioned with text inputs and a variety of multimodal guidance. DreamPBR provides both flexibility and controllability, generating semantically accurate, detailed materials—ranging from structured, repetitive patterns to imaginative, dynamic designs (in Fig. 1).

The key idea of our method is to integrate pretrained 2D text-to-image diffusion models [10] with material priors to generate high-fidelity and diverse materials. Although 2D LDMs excel at generating natural images, they struggle with spatially varying, physically based materials. To address this, we propose a novel **material LDM (m-LDM)** using a two-stage strategy. In the first stage, we fine-tune the LDM from text-to-image to text-to-albedo, effectively distilling from a large source domain (natural images) to a smaller target domain (albedo maps). In the second stage, we employ a PBR decoder to reconstruct full SVBRDFs from the latent space, using a decoder-only architecture to ensure strong correlations among all parameter maps. This approach maintains generative diversity, as the denoising U-Net remains unchanged. Additionally, a highlight-aware decoder refines all maps, enhancing regularization and overall fidelity.

To ensure diverse control and user-friendliness, we introduce a **multimodal guidance module** designed to serve as the conditioning mechanism for our material LDM. Specifi-

† The co-corresponding authors are Wei Gao (gaowei262@pku.edu.cn) and Duan Gao (gaoduan0306@gmail.com). This work was supported in part by the National Science and Technology Major Project of China (2024ZD01NL00101).

cally, this guidance module includes three key components: (1) *Pixel Control* allows pixel-aligned guidance from inputs like sketches or inpainting masks, (2) *Style Control* extracts style features from reference images and employs them to guide the generation process, and (3) *Shape Control* enables automatic material generation for a given 3D object with segmentations with an optional 2D exemplar image for reference. Importantly, our framework supports the seamless concurrent use of multiple guidances.

We train our DreamPBR on a publicly available SVBRDF dataset, comprising over 700 high-resolution SVBRDFs. Moreover, we achieve seamless tileable material generation by applying circular padding to all convolutional operations. As a result, DreamPBR demonstrates its ability to generate high-fidelity, diverse, and customizable materials, effectively bridging the gap between text-driven creativity and physically based realism.

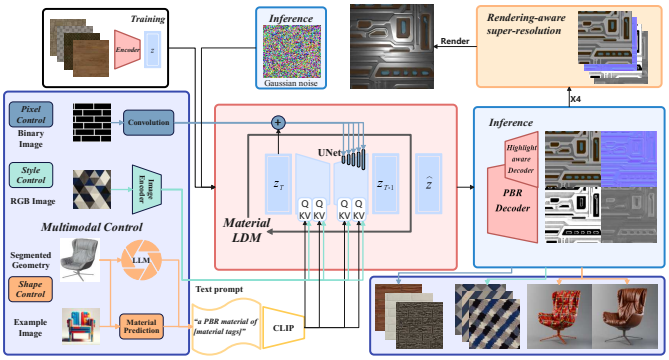

Fig. 2. Overview of DreamPBR: The denoising UNet in our Material LDM is trained with only albedo textures (upper left) and a PBR Decoder with Highlight Aware Decoder is used to transform albedo textures to other physically based textures (middle right). In the blue box on the left, we present three individual control modules: Pixel Control, Style Control, and Shape Control, whose results under controls are shown on the lower right. Besides, an additional Rendering-aware-super-resolution module is given for higher-quality textures (upper right).

II. Related Work

A. Material estimation

Material estimation approaches aim to acquire material data from real-world measurements under varying viewpoints and lighting conditions. Traditional methods often rely on multiple images or video sequences captured by a handheld camera, using regularization techniques or leveraging material priors to estimate appearance properties and simplify the problem, especially given the limitations of lightweight setups.

In recent years, deep learning-based methods have shown significant progress in recovering SVBRDFs from single image [4], [5], [11]–[14]. These methods employ deep convolutional neural network to predict plausible SVBRDFs from in-the-wild input images in a feed-forward manner. A single-image-based solution was extended to multiple images through latent space max-pooling by [15]. A deep inverse rendering pipeline that enables appearance estimation from an arbitrary number of input images was introduced by [1]. In the domain of procedural material modeling, material parameters with fixed node graphs are optimized to match input images by some methods [3], [16], [17]. A new pipeline that eliminates the need for predefined node graphs was introduced by [18]. Most recently, a diffusion-based model for estimating material properties from a single photograph was proposed by [19].

The methods mentioned above rely on captured photographs to reconstruct material and cannot produce non-real-world materials. In contrast, our approach can generate diverse and creative SVBRDFs using natural language inputs.

B. Material generation

GAN-based methods show advantages in generating high-resolution and visually compelling materials. [2] proposed an unconditional MaterialGAN for synthesizing SVBRDFs from random noise. The learned latent space facilitates efficient material estimation in inverse renderings. [6] developed a StyleGAN2-based model, conditioned by spatial structure and material category, for tileable material synthesis. However, their diversity is constrained by the training instability of GANs and the limited range of training datasets. In procedural material generation, [20] first introduced a transformer-based autoregressive model. Later work by [21] proposed a multi-model node graph generation architecture for creating high-quality procedural materials, guided by both text and image inputs. While procedural representations are compact and resolution-independent, they are limited to stationary patterns and cannot create arbitrary styles.

The closely related work to ours is ControlMat [22], a diffusion-based material generative model, capable of generating tileable materials using text and a single photograph as input. Despite augmenting the data in the material domain, this dataset is relatively small compared to the billions of text-image pairs used in text-to-image diffusion model training. Furthermore, this work only supports guidance of text and a single photograph, limiting the scenario scope.

In contrast, our method significantly enhances material generation diversity through the efficient integration of pretrained diffusion models with material priors. We also provide a variety of user-friendly controls for guiding the generation process, expanding the scope and flexibility of applications.

III. Method

A. Overview

DreamPBR is an LDM-based generative framework capable of producing diverse, high-quality SVBRDF maps under text and multimodal guidance, as illustrated in Fig. 2. The goal of our method is to generate spatially varying materials which are represented by the Cook-Torrance microfacet BRDF model with GGX normal distribution function [23]. Specifically, we use metallic-based PBR workflow and represent surface reflectance properties as albedo map $\mathcal{P}$, normal map $\mathcal{N}$, roughness map $\mathcal{R}$, and metallic map $\mathcal{M}$. Here, $\mathcal{P}, \mathcal{N} \in \mathbb{R}^{H \times W \times 3}$, $\mathcal{M}, \mathcal{R} \in \mathbb{R}^{H \times W \times 1}$, $H = W = 512$.

The core generative module of our framework is the m-LDM, where the textual description T guides the denoising

process to encode high-dimensional surface reflectance properties into a compact latent representation z. This representation effectively compresses complex material data and guides the SVBRDF decoder in reconstructing detailed SVBRDF maps $S = \{\mathcal{P}, \mathcal{N}, \mathcal{R}, \mathcal{M}\}$. Our critical observation is that while pre-trained text-to-image diffusion models can capture a wide range of natural images that fulfill the diversity needs of material generation, their flexibility often leads to less plausible materials due to the absence of material priors. Instead of training material LDM from scratch with limited material data, we opted to fine-tune a pre-trained text-to-image diffusion model with target material data. This strategy effectively tailors the model from a broad image domain to a specific material domain, ensuring both diversity and authenticity of output.

Our multimodal guidance module integrates three control modules to enhance material generation. The Pixel Control module provides spatial guidance using pixel-aligned inputs like sketches or masks. The Style Control module extracts feature from image prompts to adapt the material LDM via cross-attention. The Shape Control module automatically generates SVBRDF maps for segmented 3D shapes, leveraging large language models for text prompt creation and optionally incorporating 2D photo exemplars for part-specific material generation.

B. Physically Based material diffusion

Our material LDM uses CLIP's text encoder $\tau(\cdot)$ [24] to extract text features $\tau(T)$ from user prompts T, and transforms these features into a latent representation z for the SVBRDF maps S. The latent space is characterized by a Variational Autoencoder (VAE) architecture $\mathcal{E}$ [25]. Specifically, an albedo map $\mathcal{P}$ is compressed into a latent space $z = \mathcal{E}(\mathcal{P})$.

The core component of the diffusion model is the denoising U-Net module [26] which is conditioned on timestep t. Following Denoising Diffusion Probabilistic Models (DDPM) [27], our model employs a deterministic forward diffusion process $q(z_t|z_{t-1})$ to transform latent vectors z towards an isotropic Gaussian distribution. The U-Net network is specifically trained to reverse the diffusion process $q(z_{t-1}|z_t)$, iteratively denoising the Gaussian noise back into latent vectors. Adopting the strategy proposed by [10], we incorporates the text feature $\tau(T) \in \mathbb{R}^{M \times d_\tau}$ into the intermediate layer of U-Net through a cross-attention mechanism $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$, where $Q = W_Q^i \cdot \varphi_i(z_t), K = W_K^i \cdot \tau(T), V = W_V^i \tau(T)$, $\varphi_i(z_t)$ represents an intermediate representation of the U-Net ϵ_θ, and W_V^i, W_Q^i, W_K^i are learnable projection matrices.

Our material LDM is fine-tuned on text-material pairs via:

$$\mathcal{L}_{ldm} = \mathbb{E}_{\mathcal{E}(\mathcal{P}),T,\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(T))\|_2^2\right], \quad (1)$$

Here, ϵ represents the ground truth noise added to the latent representation z_t during the training process.

Seamless tileable texture synthesis is critical in material generation, requiring consistent spatial patterns and artifact-free tiling. To achieve this, we employ circular padding in all convolutional layers, ensuring boundary continuity and preserving internal patterns more effectively than zero padding. This approach allows our model to directly produce tileable material maps without extra post-processing and also enables the transformation of non-tileable textures into tileable versions through image-to-image generation, maintaining visual similarity with the original.

C. Render-aware SVBRDF decoder

The SVBRDF decoder, denoted as $\mathcal{D} = \{\mathcal{D}_P, \mathcal{D}_S\}$, decodes the unified latent representation z into SVBRDFs $S := \{\mathcal{P}, \mathcal{N}, \mathcal{R}, \mathcal{M}\} = \mathcal{D}(z)$. Specifically, we utilize separate decoder networks: $\mathcal{D}_P(z)$ for the albedo map $\mathcal{P}$, and $\mathcal{D}_S(z)$ for other property maps $\{\mathcal{N}, \mathcal{R}, \mathcal{M}\}$. These decoder networks follow the decoder architecture in VAE proposed by [25], and are initialized with the weights from a pre-trained VAE decoder.

Training of PBR decoder: The training loss function for our PBR decoder $\mathcal{D}_S$ comprises the following terms:

$$\mathcal{L}_{\text{PBR}} = \mathcal{L}_{map} + \mathcal{L}_{perp} + \mathcal{L}_{gan} + \mathcal{L}_{reg} + \mathcal{L}_{render}, \quad (2)$$

$$\mathcal{L}_{render}(x, y) = \|\log(x + 0.01), \log(y + 0.01)\|_1, \quad (3)$$

where $\mathcal{L}_{map}$ is L_1 loss on the material property maps, $\mathcal{L}_{perp}$ is perceptual loss based on LPIPS [28], $\mathcal{L}_{gan}$ is the generative adversarial loss, $\mathcal{L}_{reg}$ is the Kullback-Leibler divergence penalty, and $\mathcal{L}_{render}$ is $L1$ log rendering loss applied to the rendered images. The constant 0.01 in $\mathcal{L}_{render}$ is added to prevent numerical instability.

For the rendering loss, we adopt the sampling scheme proposed by [29] to render nine images per material map, ensuring desirable SVBRDF reconstructions by minimizing errors in crucial material parameters.

To reduce highlight artifacts in albedo maps, we introduce a highlight-aware albedo decoder $\mathcal{D}_P$ specifically finetuned on synthetic shaded-to-albedo data. While the standard VAE decoder can generate plausible RGB images, it often produces strong highlights for shiny materials like leather or metal. Our decoder addresses this issue by training on various shaded images rendered from our SVBRDF dataset under random lighting and viewpoint configurations, which are then mapped into latent space and decoded back to albedo using the VAE loss [25]. This process ensures robust regularization, effectively minimizing highlight artifacts and improving albedo map quality.

To address the heavy need for high-resolution material maps, we employ a material super-resolution module that enhances $H \times W$ SVBRDF maps to a resolution four times their original size. This module comprises four Real-ESRGAN-based [30] networks SR_P, SR_N, SR_R, SR_M, each specialized for a different SVBRDF map. We fine-tune these networks on synthetic material data and incorporate a rendering loss (similar to Eq. 3) to ensure that added details improve shading fidelity rather than introducing artifacts.

D. Multi-model control

We propose three control modules for DreamPBR: Pixel Control, Style Control, and Shape Control. These modules are designed to be decoupled, allowing for flexible combinations of multiple controls.

1) Pixel Control: Spatial property guidance is widely used in material creation by artists. We introduce a Pixel Control module using the ControlNet architecture [31] and spatial control maps, supporting both sketch-based generation and image-to-image inpainting with binary masks. Starting from a pre-trained ControlNet and fine-tuning it on our SVBRDF dataset significantly enhances controllability and material quality. We generate sketch guidance by extracting outlines from albedo maps via Pidinet [32], ensuring spatially consistent SVBRDF creation.

2) Style Control: The Style Control module extracts style features from an image prompt using CLIP's image encoder and a decoupled cross-attention adaptation module [33]. Combined with a text prompt, it enables multimodal generation. By capturing the appearance and structural characteristics of the input image, Style Control produces realistic, coherent material maps that align with specific exemplar images, which is a frequent requirement in the material design industry.

3) Shape Control: The Shape Control module takes a segmented 3D model and an optional photo exemplar as input and uses LLMs like ChatGPT [34] to enrich text prompts for each segmented part, enabling automatic SVBRDF generation for complex 3D shapes. Integration with Pixel and Style Control modules further refines quality and detail. We incorporate the TMT pipeline [35], which translates color and segmentation information between the exemplar image and the 3D shape, and employs a material classifier. Unlike [35], we do not rely on predefined material sets. Instead, we utilize predicted material labels and the exemplar image to produce diverse creative materials through the DreamPBR pipeline, resulting in unique 3D assets.

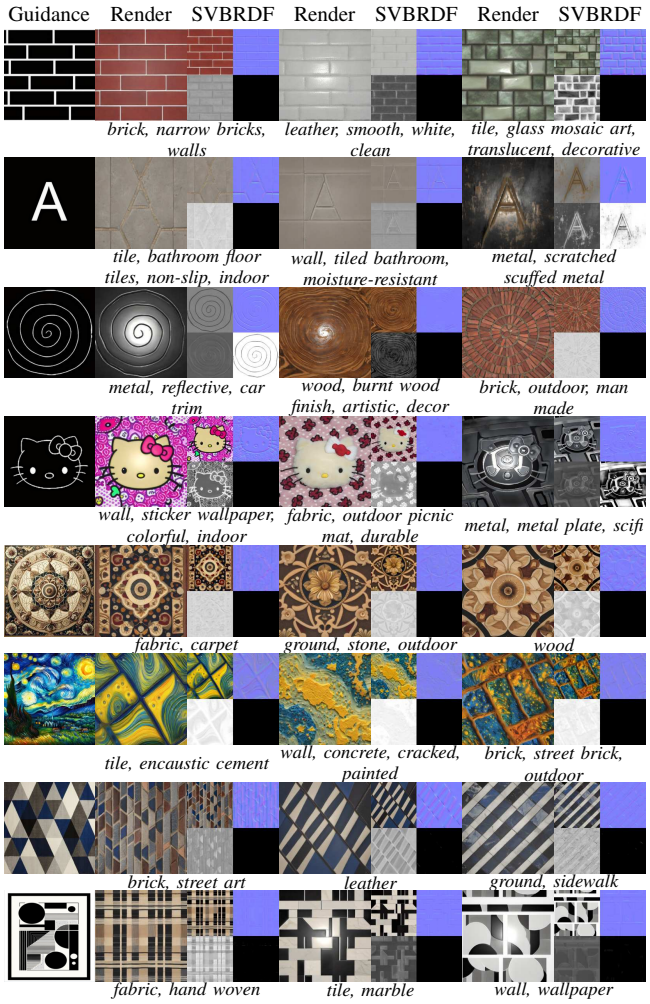

Fig. 3. Results of Pixel Control and Style Control. The first column represents the Guidance, followed by the corresponding material generation results and rendering effects. The text beneath each material represents the prompt used to generate the material, with the uniform prefix "a PBR material of" omitted.

IV. Experiments

A. Implementation Details

Our dataset includes 711 PBR materials, each with four $2k$ texture maps (albedo, normal, metallic, roughness) and accompanying textual labels. The materials are sourced from PolyHaven [1] and freePBR [2]. We manually categorize them into ten types: Brick, Fabric, Ground, Leather, Metal, Organic, Plastic, Tile, Wall, and Wood. We fine-tune the material-LDM with prompts in the format "a PBR material of [type], [name], [tags]," where [type] is the category categorized before, and [name] and [tags] are from the textual labels sourced from the dataset. Tags are randomly retained at a rate of 30%-100% during training.

We augment the $2k$ textures by flipping, rotating, multi-scale cropping, and resizing them to $H \times W$, adjusting normal maps accordingly. For the highlight-aware albedo decoder, we render these augmented textures under random viewpoints and lighting [36], using the outputs as inputs. For the pixel control module, we use Pidinet [32] to extract sketch representations from albedo maps.

DreamPBR is trained on four Nvidia RTX 3090 GPUs. We fine-tune m-LDM from stable-diffusion-v1-5. The PBR decoder produces albedo, normal, metallic, and roughness maps, and a highlight-aware albedo decoder refines the albedo. Our rendering-aware super-resolution module, adapted from Real-ESRGAN, generates high-res SVBRDFs guided by rendering loss. To improve image control, we train ControlNet on paired data extracted by Pidinet for sketch guidance.

B. Generation Results

DreamPBR enables the generation of both realistic and imaginative PBR materials from textual descriptions alone. We leverage an LLM to create material descriptions for each category, use them to sample a wide range of textures via DreamPBR, and then refine these outputs with our super-resolution module. Our experiments show that from the sam-

[1]https://polyhaven.com/
[2]https://freepbr.com/

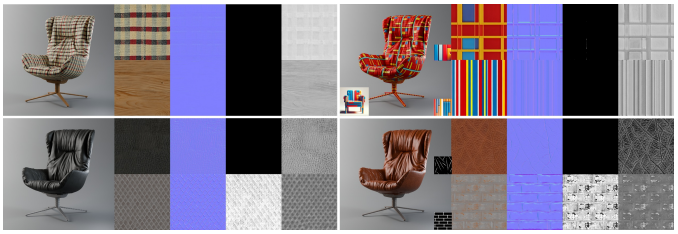

Fig. 4. Results of Shape Control. We use an LLM to describe segmented 3D shapes (e.g., chair legs and back) for texture generation via DreamPBR. Alongside text-based descriptions (left), user-specific controls like RGB images (top right) and binary masks (bottom right) enable personalized designs.

pled set of 1000 textures, they align closely with their textual descriptions, achieving a mean CLIP Score of 30.198.

Beyond basic text-driven synthesis, DreamPBR also supports pixel and style control to produce diverse materials with specific visual guidance. As shown in Fig. 3, users can guide generation through image references. Moreover, users can combine style control and pixel control to fine-tune appearances, enabling more freedom and accuracy in producing the desired textures.

Moving further, DreamPBR extends beyond planar textures to 3D objects by applying material generation to segmented shapes such as chairs. By interacting with an LLM for region-specific descriptions, or directly using cropped exemplars with pixel and style controls, DreamPBR can produce detailed, tileable textures that wrap naturally around object surfaces. Fig. 4 demonstrates the versatility and adaptability of our method in shape-aware scenarios.

Our approach effectively generates tileable textures (both direct and guided) and their stitched outcomes, demonstrating the effectiveness of our circular padding method. These results maintain diversity and seamlessly tile across different scales, allowing flexible application in various scenarios.

C. Comparative Experiments

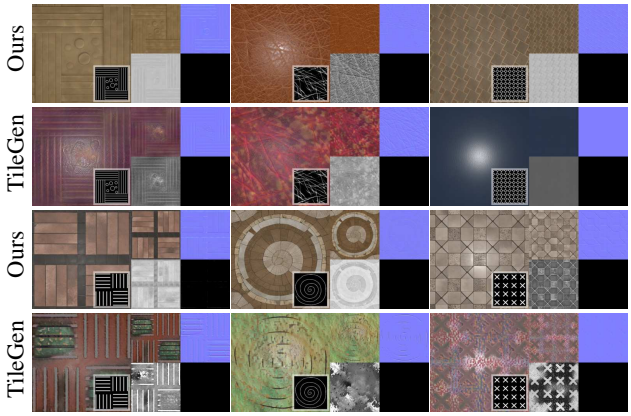

Fig. 5. A qualitative comparison of pixel guidance between PixelControl and TileGen [6] is presented for leather (top two rows) and tile (bottom two rows) textures. TileGen frequently fails to align patterns, resulting in artifacts in some renderings. In contrast, PixelControl consistently produces more accurate and artifact-free results.

We compare PixelControl with TileGen [6] in a sketch-guided generation. As shown in Fig. 5, using the same sketch, DreamPBR produces fewer artifacts and provides more precise control than TileGen. This highlights our method's advantage in sketches-driven generation and its ability to outperform prior work in material synthesis.

Leveraging Stable Diffusion, DreamPBR demonstrates strong competitiveness in material generation. DreamPBR's outputs for various materials indicate that it not only aligns with real-world distribution like GAN-based methods but also excels at generating imaginative "magic" textures.

D. Ablation Study

DreamPBR training incorporates multiple modules and additional loss functions to enhance realism and minimize the output search space. To assess the effectiveness of the previously mentioned rendering loss ($\mathcal{L}_{\text{render}}$), we trained two PBR decoders—one with and one without the loss—and evaluated them on 500 samples. As shown in Fig.6 and Table.I, the rendering-aware decoder produces more realistic images with consistent textures. Additionally, it achieves greater consistency in rendering both images and SVBRDF components, especially normal maps, and attains the lowest LPIPS and RMSE scores. These results confirm that integrating rendering feedback into the decoding process significantly improves performance.

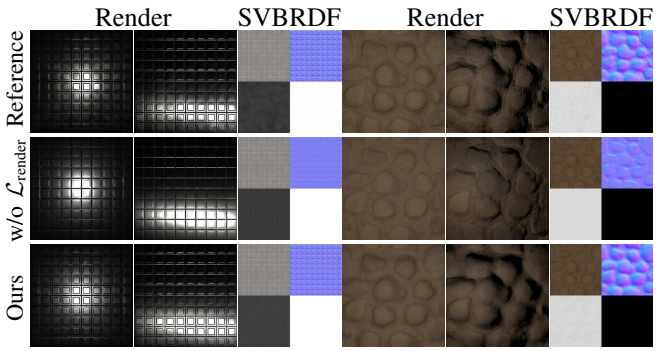

Fig. 6. Qualitative comparison on PBR decoder with rendering loss.

TABLE I
QUANTITATIVE RESULTS OF PBR DECODERS WITH RENDERING LOSS

Method	LPIPS	RMSE			
	Render	*Albedo*	*Metallic*	*Normal*	*Roughness*
w/o $\mathcal{L}_{\text{render}}$	0.107	0.0361	0.0126	0.0542	0.0406
Ours (w/ $\mathcal{L}_{\text{render}}$)	0.101	0.0357	0.0086	0.0531	0.0365

Due to space constraints, the ablation studies for the super-resolution module, highlight-aware decoder, and Pixel Control module are presented in the supplemental material.

V. CONCLUSIONS

In this paper, we present DreamPBR, a diffusion-based generative framework for creating high-resolution, physically based material textures. DreamPBR offers flexibility and

controllability by allowing users to generate detailed and semantically accurate materials through text descriptions and various multimodal inputs, including images and geometries. It supports diverse material styles, ensuring high-quality, tileable textures for various applications. By integrating multiple guidance types, DreamPBR enables personalized and intuitive creation, effectively bridging creative text generation with realistic, physically based materials.

REFERENCES

[1] Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong, "Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–15, 2019.

[2] Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao, "Materialgan: Reflectance capture using a generative svbrdf model," *ACM Trans. Graph.*, vol. 39, no. 6, nov 2020.

[3] Yiwei Hu, Julie Dorsey, and Holly Rushmeier, "A novel framework for inverse procedural texture modeling," *ACM Trans. Graph.*, vol. 38, no. 6, nov 2019.

[4] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau, "Single-image svbrdf capture with a rendering-aware deep network," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, pp. 1–15, 2018.

[5] Jie Guo, Shuichang Lai, Qinghao Tu, Chengzhi Tao, Changqing Zou, and Yanwen Guo, "Ultra-high resolution svbrdf recovery from a single image," *ACM Trans. Graph.*, vol. 42, no. 3, jun 2023.

[6] Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari, "Tilegen: Tileable, controllable material generation and capture," in *SIGGRAPH Asia 2022 conference papers*, 2022, pp. 1–9.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[8] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv*, 2022.

[9] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[11] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong, "Modeling surface appearance from a single photograph using self-augmented convolutional neural networks," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–11, 2017.

[12] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker, "Materials for masses: Svbrdf acquisition with a single mobile phone image," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, Berlin, Heidelberg, 2018, p. 74–90, Springer-Verlag.

[13] Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan, "Highlight-aware two-stream network for single-image svbrdf acquisition," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021.

[14] Philipp Henzler, Valentin Deschaintre, Niloy J. Mitra, and Tobias Ritschel, "Generative modelling of brdf textures from flash images," *ACM Trans. Graph.*, vol. 40, no. 6, dec 2021.

[15] Valentin Deschaintre, Miika Aittala, Fr'edo Durand, George Drettakis, and Adrien Bousseau, "Flexible svbrdf capture with a multi-image deep network," *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)*, vol. 38, no. 4, July 2019.

[16] Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik, "Match: Differentiable material graphs for procedural material capture," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, Dec. 2020.

[17] Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre, "Node graph optimization using differentiable proxies," in *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022, SIGGRAPH '22, Association for Computing Machinery.

[18] Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier, "An inverse procedural modeling pipeline for svbrdf maps," *ACM Trans. Graph.*, vol. 41, no. 2, jan 2022.

[19] Sam Sartor and Pieter Peers, "Matfusion: a generative diffusion model for svbrdf capture," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10.

[20] Paul Guerrero, Milos Hasan, Kalyan Sunkavalli, Radomir Mech, Tamy Boubekeur, and Niloy Mitra, "Matformer: A generative model for procedural materials," *ACM Trans. Graph.*, vol. 41, no. 4, 2022.

[21] Yiwei Hu, Paul Guerrero, Milos Hasan, Holly Rushmeier, and Valentin Deschaintre, "Generating procedural materials from text or image prompts," in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*. July 2023, SIGGRAPH '23, ACM.

[22] Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur, "Controlmat: Controlled generative approach to material capture," *arXiv preprint arXiv:2309.01700*, 2023.

[23] Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance, "Microfacet models for refraction through rough surfaces," in *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, Goslar, DEU, 2007, EGSR'07, p. 195–206, Eurographics Association.

[24] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," 2020.

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[29] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau, "Single-image svbrdf capture with a rendering-aware deep network," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, pp. 1–15, 2018.

[30] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

[32] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu, "Pixel difference networks for efficient edge detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5117–5127.

[33] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[34] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[35] Ruizhen Hu, Xiangyu Su, Xiangkai Chen, Oliver van Kaick, and Hui Huang, "Photo-to-shape material transfer for diverse structures," *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, vol. 39, no. 6, pp. 113:1–113:14, 2022.

[36] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila, "Modular primitives for high-performance differentiable rendering," *ACM Transactions on Graphics*, vol. 39, no. 6, 2020.