PBR3DGen: A VLM-Guided Mesh Generation with High-Quality PBR Texture

Xiaokang Wei[1,2*], Bowen Zhang[2*], Xianghui Yang[2], Yuxuan Wang[2,3], Chunchao Guo[2], Xi Zhao[4], Yan Luximon[1†]

[1]The Hong Kong Polytechnic University
[2]Tencent Hunyuan3D
[3]Nanyang Technological University
[4]Xi'an Jiaotong University
xiaokang.wei@connect.polyu.hk

Abstract

Generating high-quality physically based rendering (PBR) materials is important to achieve realistic rendering in the downstream tasks, yet it remains challenging due to the intertwined effects of materials and lighting. While existing methods have made breakthroughs by incorporating material decomposition in the 3D generation pipeline, they tend to bake highlights into albedo and ignore spatially varying properties of metallicity and roughness. In this work, we present **PBR3DGen**, a two-stage mesh generation method with high-quality PBR materials that integrates the novel multi-view PBR material estimation model and a 3D PBR mesh reconstruction model. Specifically, PBR3DGen leverages vision language models (VLM) to guide multi-view diffusion, precisely capturing the spatial distribution and inherent attributes of reflective-metalness material. Additionally, we incorporate view-dependent illumination-aware conditions as pixel-aware priors to enhance spatially varying material properties. Furthermore, our reconstruction model reconstructs high-quality mesh with PBR materials. Experimental results demonstrate that PBR3DGen significantly outperforms existing methods, achieving new state-of-the-art results for PBR estimation and mesh generation.

Website — https://pbr3dgen1218.github.io/

1 Introduction

Generating high-quality 3D mesh with physically based rendering (PBR) materials from images or text prompts has broad applications such as 3D graphics pipelines, movie production, gaming and AR/VR. 3D generation models have witnessed remarkable progress which can be attributed to the scalability of 3d generative models (Tang et al. 2023; Long et al. 2024; Shi et al. 2023a; Xu et al. 2024), and utilization of the large-scale training datasets (Deitke et al. 2023). However, most existing 3D mesh and texture generation models often lack PBR material properties so they lose the view-dependent photorealistic effect. Furthermore, the textures, which incorporate pre-baked shadows and lighting, restrict their applicability in downstream tasks.

PBR estimation from images is a recent development in the field of 3D assets with PBR generation, such as Clay (Zhang et al. 2024b) has demonstrated impressive capabilities in 3D mesh with PBR materials generation, but they rely on expensive geometric prior to ensure cross-view PBR consistency. 3DTopia-XL (Chen et al. 2024b) encodes detailed shape, Albedo, and material field into a Diffusion Transformer (DiT) framework. To achieve better performance, these methods have attempted to incorporate BSDF and illumination models into 3D generation model. More recently, Meta3DAssetGen (Siddiqui et al. 2024) exploit to integrate the differentiable BRDF optimization model into forward transformer-based architecture. SF3D (Boss et al. 2024) incorporates explicit lighting and a differentiable shading model for decomposing light from UV texture.

In the field of 3D generation, PBR material representation has recently emerged as a key advancement, significantly improving the rendering quality of generated objects. For example, Clay (Zhang et al. 2024b) has demonstrated impressive capabilities to generate PBR materials for 3D meshes, yet it requires geometric input to maintain cross-view consistency and concurrently suffers from the deterioration of detailed textures 3DTopia-XL (Chen et al. 2024b) develops a primitive presentation to encode detailed shape, Albedo, and material fields into a Diffusion Transformer (DiT) framework. On the other hand, recent methods have attempted to incorporate BSDF material functions and illumination models into 3D generation models. Meta3DAssetGen (Siddiqui et al. 2024) exploits the integration of the differentiable BRDF optimization model into a large reconstruction model (LRM) with a forward transformer-based architecture. SF3D (Boss et al. 2024) incorporates explicit lighting and a differentiable shading model for decomposing light from UV textures.

However, by simplifying material models with a single illumination for different objects and ignoring the spatially varying properties of metalness and roughness, these methods face three significant limitations that hinder their widespread adoption. Firstly, these methods are more likely to bake high light into the Albedo map due to the existing ambiguity between illumination and Albedo, especially for reflective objects. Secondly, roughness and metallicity are difficult to observe directly from RGB images. We also

*These authors contributed equally.
†Corresponding author.

Figure 1: We present **PBR3DGen**, a novel two-stage 3D assets generation framework with high-quality physically-based rendering materials. All objects in the scene are generated from PBR3DGen.

noticed that the BRDF distribution in current synthetic 3D datasets, such as Objaverse (Deitke et al. 2023), exhibits a strong long-tail effect. This causes models to overfit to frequent values while ignoring rare ones. As a result, the quality of PBR material generation is compromised, and spatially-varying attributes are poorly represented. Additionally, Meta3DAssetGen (Siddiqui et al. 2024) employs LRM to predict PBR materials, which is trained from scratch. Given the scarcity of high-quality 3D PBR data, the generalization capability of the LRM may not be as strong as that of diffusion models for predicting PBR. Meanwhile, the performance of sparse-view reconstruction models can decline if the quality of multiple views is poor, as these models generally depend exclusively on view-aware RGB images to predict 3D representations.

In response to these challenges, we propose an efficient approach for high-quality 3D mesh with PBR materials generation from a single image or text prompt that disentangles the highlight and reconstructs spatially-varying metallic and roughness to enable relighting (see Fig.1). Our method introduces PBR3DGen, a two-stage 3D mesh generation method with high-fidelity PBR materials that integrates the novel PBR multiview diffusion models and PBR-based sparse-view reconstruction models to achieve high-quality 3D mesh with PBR materials generation. In the first stage, we leverage vision language models (VLM) like GPT-4V and view-dependent illumination-aware conditions to guide our multi-view PBR estimation model. Specifically, we employ VLM to precisely capture the spatial distribution and inherent attributes of reflective-metallic materials. This detailed information is then seamlessly integrated into a PBR multi-view diffusion model. This integration plays a pivotal role in drastically mitigating the ambiguity often encountered between specular highlights and Albedo within the rendered imagery. Furthermore, it addresses the issue of prediction inaccuracies in metallic and roughness properties, which were previously exacerbated by the severe long-tail distribution of training data. Consequently, our approach significantly enhances the consistency of part-aware material representation. In addition, we inject view-dependent illumination-aware conditions to enhance the spatially varying material properties. In the second stage, unlike most sparse-view reconstruction models that reconstruct 3D assets using only rgb color, we propose our PBR-based large reconstruction model, which employs a dual-head VAE encoder to separately encode the Albedo and Metallic-Roughness maps. Subsequently, we reconstruct the 3D mesh and PBR materials from the input PBR multi-view images. To conclude, our contributions are: **1)** We propose PBR3DGen, a novel two-stage generation framework for 3D assets with high-quality PBR materials from image or text inputs. **2)** We explore leveraging vision-language models to guide PBR estimation within a multi-view diffusion model, enabling more accurate estimation of PBR materials and significantly reducing ambiguities between Albedo and illumination, especially specular highlights. **3)** We introduce a view-dependent illumination-aware condition as a local pixel-wise prior, resulting in a more accurate capture of spatially varying reflectance effects. **4)** Our method exhibits superior 3D mesh and PBR material generation quality compared to current methods.

2 Related Works

2.1 Multi-view Diffusion

Cross-view consistency is crucial in a reconstruction-based 3D generation. MVDiffusion(Tang et al. 2023) first gener-

ates consistent multi-view images from text prompts, given pixel-to-pixel correspondences. SyncDreamer(Liu et al. 2023a), MVDream(Shi et al. 2023b), and Wonder3D(Long et al. 2024) leverage attention mechanisms to facilitate information transfer between multi-view images, enhancing multi-view consistency. Zero123++(Shi et al. 2023a) stitches multi-view images together while denoising them simultaneously, improving geometric consistency and texture quality. Era3D(Li et al. 2024) introduces row-wise multi-view attention, reducing the computational overhead of multi-view generation. Although the quality of current multi-view image generation has made significant progress, multi-view images embed lighting and lack physical properties. In contrast to these previous methods, our approach not only maintains multi-view consistency but also provides PBR estimation. This enables our reconstruction results to support relighting and physically-based rendering.

2.2 Multi-view 3D Reconstruction

3D reconstruction has been a well-researched area in the field of computer vision for a long time. Although traditional methods such as Structure from Motion (SfM)(Agarwal et al. 2011; Pollefeys et al. 2004; Schonberger and Frahm 2016) and Multi-View Stereo (MVS)(Furukawa and Ponce 2009; Pollefeys et al. 2008; Schönberger et al. 2016) can perform camera calibration and 3D reconstruction, they lack robustness when dealing with inconsistent multi-view images. Recently, deep learning-based 3D reconstruction methods have become mainstream. LRM(Hong et al. 2023) first proposed utilizing a transformer backbone to simultaneously reconstruct geometry and texture through a single forward pass. LRM can learn how to reconstruct 3D geometry and texture from a single image using large-scale 3D datasets. Instant3D(Li et al. 2023) increases the number of input views, further enhancing the geometric detail and texture quality of the reconstruction. Subsequent works(Zhang et al. 2024a; Wei et al. 2024; Zhang et al. 2025; Xu et al. 2024; Li et al. 2022) have made further improvements in reconstruction quality and computational efficiency. Compared to existing large reconstruction models, we not only reconstruct the 3D geometry, but we also reconstruct physics-based material properties, making the 3D assets generated by our method more realistic.

2.3 Diffusion-based PBR Material Generation

Material generation and estimation from RGB images is inherently difficult because of its under-constrained nature, including the ambiguity between illumination and materials. Diffusion models reveal the impressive capability of learning the distribution of the target domain, which has become prominent in texture content generation. Recent studies such as TEXTure (Richardson et al. 2023), Text2Tex (Chen et al. 2023a), TexFusion (Cao et al. 2023) extend the diffusion model to texture synthesize from multi-view images. SyncMVD (Liu et al. 2023b) improves the consistency of multi-view texture by sharing the denoised content among different views in each denoising step to ensure texture consistency and avoid seams and fragmentation. However, these methods tend to generate RGB textures with highlights and shadows due to suffering from disentangling the materials and illumination properties. Recently, Paint3D (Zeng et al. 2024a) proposed a coarse-to-fine strategy to delight the generated texture in UV space, but they still lack Physically-based materials when dealing with inconsistent multi-view images. Fantasia3D (Chen et al. 2023b) can generate more realistic textures by incorporating physics-based materials. FlashTex (Deng et al. 2024) introduces the lighting condition in ControlNet and optimizes texture based on Score Distillation Sampling loss, which can disentangle lighting from surface material/reflectance. However, these optimization-based methods require extensive training time. RGBX (Zeng et al. 2024b) utilizes diffusion models to estimate the Physically-based intrinsics of RGB images and demonstrates a significant improvement on the generalization. Additionally, IntrinsicAnything (Chen et al. 2024a) and IntrinsicReal (Wei et al. 2025) only utilizes diffusion models to estimate albedo and specular shading from a single RGB image, which lacks roughness and metallic properties in 2D diffusion prior stage. Meanwhile, these methods still suffer from disentangling the spatially varying materials from highlight and reflective object images.

3 Methods

3.1 Overview

In this section, we introduce our high-quality 3D mesh with PBR materials generation pipeline, illustrated in Fig.2. Firstly, we establish a novel multi-view PBR texture diffusion model from a single image or text prompt, which unleashes Vision Language model (VLM) and illumination-aware conditions to guide PBR material generation. Secondly, we further design dual-head reconstruction model by extending a sparse-view large reconstruction model to handle multi-view PBR material inputs, significantly boosting the 3D mesh and texture reconstruction quality.

3.2 Multi-view PBR Estimation

Without prior geometric information, generating multi-view PBR materials directly from a single RGB image using a diffusion model is a challenging problem. Due to the inherent ambiguities between material properties and lighting conditions, and the observation that the local distribution of material characteristics does not always align with geometric consistency. Multi-view PBR estimation from an RGB image cannot achieve both high-quality PBR materials and geometrically consistent multi-view images simultaneously. Therefore, We first resolve multi-view consistency by applying an off-the-shelf Multi-view Diffusion Model, such as Zero123++(Shi et al. 2023a). Meanwhile, we propose a PBR generation module from the multi-view RGB image input by introducing VLM guidance and view-dependent illumination-aware condition, which can significantly handle the ambiguity between PBR and illumination well.

Multi-view PBR Material Estimation with SD Since PBR material estimation can be seen as domain transfer from an RGB image, The diffusion model can be utilized to transfer from the RGB domain distribution to the PBR domain distribution. And we observed that Roughness and

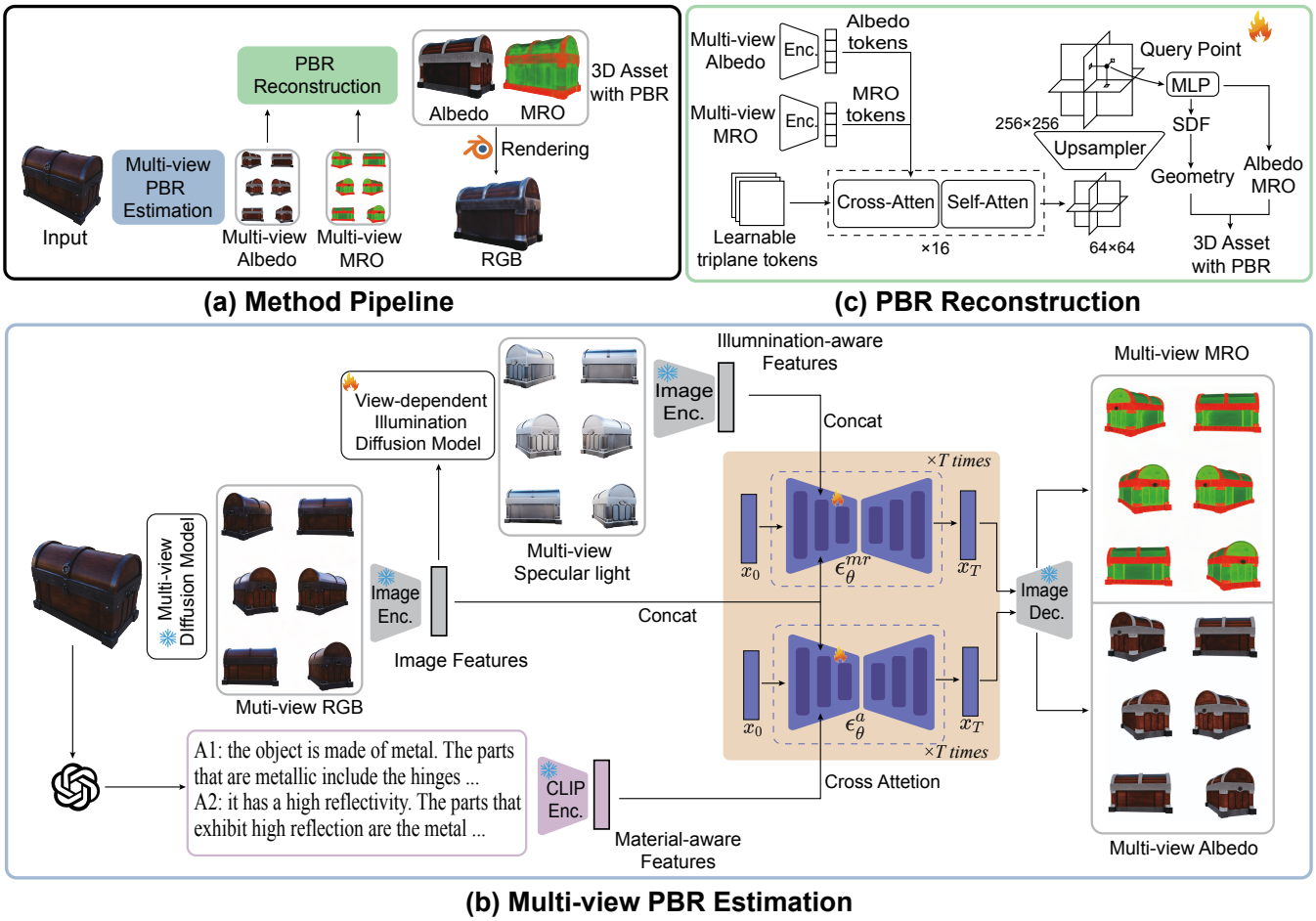

Figure 2: Overview of PBR3DGen. Our method consists of two stages: Multi-view PBR materials estimation and 3D mesh with PBR materials reconstruction. Given an RGB image as input, we first generate multi-view Albedo images and multi-view MRO images using Multi-view PBR estimation model, and then reconstruct 3D assets with Dual-head PBR-based LRM.

Metallic often interact to produce specular reflection effects on objects. In our methods, we combined two maps into a single MRO (Metallic/Roughness/Zero channel) for generation. Therefore, we can model the conditional distribution of corresponding Albedo and Metallic-Roughness by utilizing the RGB image as the conditioning signal, as in IntrinsicAnything (Chen et al. 2024a) and RGBX (Zeng et al. 2024b). Specifically, we first use the pre-trained VAE image encoder $\mathcal{E}$ to extract the conditional signal feature from input grid image $\boldsymbol{I}$. Then, the diffusion process adds noise to the encoded latent $z = \mathcal{E}(x)$ producing a noisy latent z_t where the noise level increases over timesteps $\boldsymbol{x} \in T$. We learn network ϵ_θ that predicts the noise added to the noisy latent z_t given image conditioning $\mathcal{E}(\boldsymbol{I})$. We minimize the following loss:

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(I),\epsilon,t}\left[|\epsilon - \epsilon_\theta\left(z_t, t, \mathcal{E}(\boldsymbol{I})\right)|_2^2\right] \quad (1)$$

where z_t is the noisy latent feature of the input z with t uniformly sampled from $\{1, \ldots, T\}$, and estimating ϵ from a Gaussian distribution, denotes $\epsilon \sim \mathcal{N}(0, 1)$. Here, we separately optimize double U-Net Network ϵ_θ^a and ϵ_θ^{mr} corresponding Albedo estimation and MRO estimation.

VLM-guided PBR Material generation We observed that the Roughness and Metallic distributions for objects tend to be strongly part-aware consistent, and there are significant ambiguities between Albedo and specular light during the Albedo generation process, especially for specular and metallic objects. This is due to the high variance in the diffusion inference procedure. As argued above, the PBR material generation model requires a strong prior to alleviate these challenging ambiguities. Inspired by (Fang et al. 2024), VLM can recognize object materials and types due to its extensive prior knowledge of objects. We design a hierarchical VLM-guided material policy to obtain reflective-metallic material attributes through unleashing GPT-4V with a strong material knowledge, which helps us capture the reflective-metallic properties for the input RGB image. Specifically, we first gather global information about metallic and reflective properties, and then we further investigate which parts exhibit these relevant attributes. The

pipeline of the designed hierarchical strategy is depicted in Fig.2(b), with detail provided in the Fig.11 of Appendix.

To effectively utilize reflective-metallic material information in the PBR diffusion model process, we inject material caption features into the U-Net. Specifically, we first use the CLIP text encoder (Radford et al. 2021) to extract language features from the material captions. Then, to inject the conditioning signal into the Albedo U-Net network ϵ_θ and the Metallic-Roughness U-Net network ϵ_σ, we capture the embedding relationship using cross-attention mechanism between the material caption language features and the noised latent of the U-Net. We reparameterize the loss functions of ϵ_θ^a and ϵ_θ^{mr} following the corresponding conditioning signals in Eq.2 and Eq.3:

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(I),\epsilon,t}[|\epsilon - \epsilon_\theta^{mr}(z_t^{mr}, t, f_s(\mathcal{E}(\boldsymbol{I})), C_T(\boldsymbol{I}))|_2^2] \tag{2}$$

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(I),\epsilon,t}[|\epsilon - \epsilon_\theta^{a}(z_t^{a}, t, C_T(\boldsymbol{I}))|_2^2] \tag{3}$$

where C_T denotes CLIP text feature encoder, z_t^a is the latent feature encoded from the ground truth Albedo at timestep t.

View-dependent illumination-aware condition Although the designed VLM-guided material diffusion model can effectively handle part-wise properties by injecting material description information, Metallic-Roughness estimation still struggles to capture spatially varying properties, due to the challenges posed by the severe long-tail distribution effect in the training data. Therefore, it is essential to introduce additional physics-based prior information.

Physically Based Rendering (PBR) materials often utilize the spatially-varying bi-directional reflectance distribution functions (svBRDFs) to approximate the surface reflectance property with a set of decomposed intrinsic terms, which can encapsulate the interaction of light with the object surface. Following the popular specular svBRDFs model, like the Cook-Torrance model (Cook and Torrance 1982), we can observe that the Roughness and Metallic properties are closely integrated with the specular shading term. The detailed rendering equation is provided in the Appendix (Sec.A.1). Specifically, the rendering equation is rewritten as Eq. 4:

$$L_o(\hat{\mathbf{x}}, \boldsymbol{\omega}_o) = L_{diff}(\hat{\mathbf{x}}, k_d, L_i) + L_{spec}(\hat{\mathbf{x}}, \boldsymbol{\omega}_o, F_0(k_m), k_r, L_i) \tag{4}$$

where the rendering results comprise diffuse shading L_{diff} and specular shading L_{spec}, k_d and k_r represent Albedo and Roughness, k_m is metallic term related to the Fresnel term.

Previous methods (Boss et al. 2024; Zhang et al. 2024b; Siddiqui et al. 2024) for generating Roughness and Metallic maps tend to homogenize the distribution of BRDF in synthetic 3D data. For example, in the widely used Objaverse dataset (Deitke et al. 2023), we observe that the Metalness and Roughness of objects often have fixed values. This results in a lack of spatial variability in the BRDF distribution obtained through diffusion model training, which negatively impacts the accuracy of the Metalness and Roughness attributes. To overcome the severe impact of the long-tail distribution, and motivated by Eq.4, we propose introducing view-dependent illumination as a condition to guide the sampling process of Roughness and Metallic. We further reformulate the loss function shown in Eq.1 into the following form in Eq. 5:

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(I),\epsilon,t}[|\epsilon - \epsilon_\theta(z_t^{mr}, t, f_s(\mathcal{E}(\boldsymbol{I})))|_2^2] \tag{5}$$

where z_t^{mr} is the latent feature encoded from the ground truth Metallic-Roughness gt at time step t, f_s is a specular illumination diffusion model to generate spatially-varying illumination based on single RGB grid image $\boldsymbol{I}$. This specular illumination map acts as a local cue, enhancing the Metallic and Roughness spatial distribution closely linked to the actual distribution of the object.

3.3 Dual-head PBR Reconstruction Model

Our PBR reconstruction model takes multi-view Albedo multi-view images and MRO images as input, and it outputs 3D assets with PBR material. Since our reconstruction model has two types of inputs, using a single image encoder to extract features from both types would cause the model to be unable to distinguish between them. Therefore, we use dual-head PBR encoders: one encoder is for encoding the Albedo images, and the other encoder is for encoding the MRO images. Then two kinds of image tokens and learnable triplane tokens are passed through a transformer backbone, outputting a low-resolution triplane. The low-resolution triplane has a resolution of 64 with 1024 channels. We use an upsampler similar to (Yang et al. 2024) to upsample the low-resolution triplane, ultimately obtaining a high-resolution triplane with a resolution of 256 and 120 channels. For each query point, we project it onto a triplane to get its triplane feature, and we use an MLP to predict its signed distance, Albedo and MRO.

Since NeRF(Mildenhall et al. 2021) and 3D Gaussian(Kerbl et al. 2023) have difficulty generating high-quality meshes, and the training of DMTet(Shen et al. 2021) and Flexicubes(Shen et al. 2023) is unstable, we adopt NeuS(Wang et al. 2021) as our 3D representation and obtain the geometry through marching cubes. With geometry, Albedo, and MRO, we obtain 3D asset with PBR as our final output. During training, we apply image loss for both Albedo images and MRO images:

$$\begin{aligned} L = & \sum_i ||I_{i,A} - I_{i,A}^{gt}||_2^2 \\ & + \sum_i ||I_{i,MRO} - I_{i,MRO}^{gt}||_2^2 \\ & + \lambda_{lpips} \sum_i L_{lpips}(I_{i,A}, I_{i,A}^{gt}) \\ & + \lambda_{lpips} \sum_i L_{lpips}(I_{i,MRO}, I_{i,MRO}^{gt}) \\ & + \lambda_{mask} \sum_i ||M_i - M_i^{gt}||_2^2 \end{aligned} \tag{6}$$

where $I_{i,A}$, $I_{i,MRO}$, M_i, $I_{i,A}^{gt}$, $I_{i,MRO}^{gt}$ and M_i^{gt} denote rendered Albedo images, rendered MRO images, rendered mask images, ground truth Albedo images, ground truth MRO images, ground truth mask images of the i-th view. We randomly select 2 target views and set $\lambda_{lpips} = 2$, $\lambda_{mask} = 0.2$ during the training stage.

Figure 3: Qualitative comparison of the generated 3D assets with other methods.

Methods	Albedo		Roughness		Metallic		RGB		Geometry			
	PSNR↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	MSE ↓	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑
SF3D	15.98	0.03	15.00	0.04	16.04	0.03	16.37	0.03	0.33	0.61	0.79	0.92
3DTopia-XL	13.69	0.05	11.84	0.07	13.14	0.06	13.95	0.04	0.55	0.33	0.55	0.83
Ours	**17.77**	**0.02**	**16.52**	**0.03**	**16.16**	**0.03**	**17.99**	**0.02**	**0.22**	**0.72**	**0.88**	**0.97**

Table 1: Quantitative comparison on Objaverse (Deitke et al. 2023) dataset.

4 Experiments

In this section, we evaluate PBR3DGen, detailing the experiment settings and datasets in Sec. 4.1, and compare it with SOTAs in 3D generation and albedo estimation based on diffusion models across various datasets in Sec. 4.2. Finally, we demonstrate the effectiveness of different components in improving PBR estimation and mesh generation in Sec. 4.3. More results are shown in the Appendix.

4.1 Experimental Setup

Dataset. We first generate multi-view PBR material training data from the Objaverse dataset (Deitke et al. 2023). Specifically, we select 48k objects by filtering out those with low-quality PBR textures. For each object, we render 21-view multi-domain images (RGB/Albedo/MRO/Specular light) using Blender. To evaluate PBR reconstruction and geometry accuracy, we randomly sample 300 objects from Objaverse (Deitke et al. 2023) and Google Scanned Objects (GSO) (Downs et al. 2022). More details in the Appendix.

Implementation. For PBR estimation model, we use a diffusion model fine-tuned from the Stable Diffusion V2.1 (Rombach et al. 2022) based on the framework of Intructpix2pix (Brooks, Holynski, and Efros 2023) as a starting point and continue to train for 50k iterations using an Adam (Kingma 2014) optimizer at a learning rate of 1×10^{-4}. For the PBR reconstruction model, we pretrain the model with only the Albedo input for 200k iterations, and then we finetune the model for 150k iterations with both Albedo and MRO image input, the MRO image encoder is initialized with the weights of the Albedo image encoder. The learning rate is 3×10^{-5} in both stages.

Metrics. We evaluate both 2D appearance and 3D geometry. For image quality, we compute PSNR and MSE between rendered and ground-truth views in both the PBR and mesh generation stages. For geometry, we align the generated mesh with the ground-truth mesh and report Chamfer Distance(CD) and F-Score(FS).

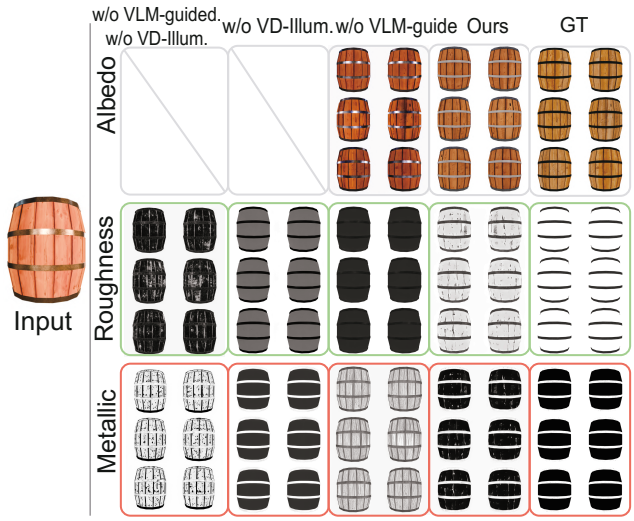

Figure 4: Qualitative ablation on PBR estimation.

Methods	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑
LGM	0.409	0.442	0.658	0.881
OpenLRM	0.214	0.605	0.840	**0.997**
TripoSR	0.356	0.511	0.727	0.920
InstantMesh	0.216	0.670	0.862	0.977
SF3D	0.274	0.554	0.786	0.956
3DTopia-XL	0.239	0.635	0.832	0.962
Ours	**0.175**	**0.739**	**0.903**	0.984

Table 2: Quantitative comparison on GSO dataset.

4.2 Comparison with Other Methods

Baselines. We compare our results with SOTA methods at both stages of our pipeline. For the MV-PBR image generation stage, we compare with IntrinsicAnything (Chen et al. 2024a), a diffusion-based intrinsic image generation method. For 3D asset generation, we focus on 3D mesh with PBR generation methods to maintain a consistent evaluation protocol. Specifically, we compare against SF3D (Boss et al. 2024), 3DTopia-XL (Chen et al. 2024b), Instantmesh (Xu et al. 2024), OpenLRM (He and Wang 2023), LGM (Tang et al. 2025), TripoSR (Tochilkin et al. 2024).

Results. From the 2D PBR estimation results in Table 3, our method yields notably higher PSNR and lower MSE than IntrinsicAnything (Chen et al. 2024a). IntrinsicAnything often produces black albedos on metallic objects, whereas our method preserves proper reflectance and shading cues, leading to more accurate PBR reconstruction. As shown in Table 1 and Table 2, our method outperforms all baselines in both novel view PBR asset quality and geometric accuracy.

	Albedo	
	PSNR↑	MSE ↓
IntrinsicAny (Chen et al. 2024a)	16.775	0.031
Ours	**18.186**	**0.023**
w/o VLM-guide	17.155	0.030

Table 3: Quantitative results for IntrinsicAnything and ours.

Methods	Roughness		Metallic	
	PSNR↑	MSE ↓	PSNR↑	MSE ↓
w/o VD-Illum w/o VLM-guide	18.649	0.020	15.293	0.046
w/o VD-Illum	19.803	0.019	15.871	0.038
w/o VLM-guide	18.023	0.023	15.890	0.038
Ours	**21.095**	**0.013**	**17.718**	**0.028**

Table 4: Ablation studies for MV-PBR estimation model.

4.3 Ablation Study

We conduct ablation studies to analyze the contribution of each component in our framework on the Objaverse PBR dataset. The key innovation of our method is its ability to effectively estimate multi-view spatially varying PBR.

(1) w/o VD-Illum. This term indicates training without the view-dependent illumination condition for the multi-view metallic-roughness diffusion model. As shown in Table 4 and Fig 4, VD-Illum. module as local pixel-aware priors prove to be effective for boosting the metallic-roughness distribution according to spatially varying attribution.

(2) w/o VLM-guide. This setting removes the VLM-guided condition from both the multi-view albedo and metallic-roughness diffusion models to evaluate its impact. As shown in Table 4, excluding the VLM module causes a significant drop in metallic-roughness accuracy and degrades albedo estimation. Qualitative results in Fig. 4 further reveal baked highlights and incorrect albedo without VLM guidance. This module improves the predicted roughness and metallic values by adding global and local material priors, reducing bias from uniform training data.

(3) w/o VD-Illum.& VLM-guide. This ablation is to evaluate the multi-view PBR diffusion model without both of the above conditions. From the first column of Fig 4, it is evident that without the VLM-guided condition and the VD-illum condition, the predicted distributions for roughness and metallic are more disorganized.

5 Conclusion

We present PBR3DGen, a two-stage 3D mesh with high-fidelity PBR materials generation framework by introducing a novel PBR multi-view diffusion model. Specifically, we explore unleashing VLM strong material prior and view-dependent illumination-aware conditions to guide PBR multi-view generation, which significantly alleviates the ambiguity often encountered between highlights and albedo within the rendered imagery and handles spatially-varying properties. Due to accurate PBR multi-view images, we significantly improved 3D mesh with PBR materials reconstruction quality based on PBR reconstruction model.

Acknowledgments

This work was supported by the Laboratory for Artificial Intelligence in Design (Project 3.1), Innovation and Technology Fund, Hong Kong SAR and by P0050655 from Non-PAIR Research Centres of The Hong Kong Polytechnic University.

References

Agarwal, S.; Furukawa, Y.; Snavely, N.; Simon, I.; Curless, B.; Seitz, S. M.; and Szeliski, R. 2011. Building rome in a day. *Communications of the ACM*, 54(10): 105–112.

Boss, M.; Huang, Z.; Vasishta, A.; and Jampani, V. 2024. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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*.

Chen, D. Z.; Siddiqui, Y.; Lee, H.-Y.; Tulyakov, S.; and Nießner, M. 2023a. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023b. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Chen, X.; Peng, S.; Yang, D.; Liu, Y.; Pan, B.; Lv, C.; and Zhou, X. 2024a. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *European Conference on Computer Vision*, 450–467. Springer.

Chen, Z.; Tang, J.; Dong, Y.; Cao, Z.; Hong, F.; Lan, Y.; Wang, T.; Xie, H.; Wu, T.; Saito, S.; et al. 2024b. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*.

Cook, R. L.; and Torrance, K. E. 1982. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1).

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*.

Deng, K.; Omernick, T.; Weiss, A.; Ramanan, D.; Zhu, J.-Y.; Zhou, T.; and Agrawala, M. 2024. FlashTex: Fast Relightable Mesh Texturing with LightControlNet. *arXiv preprint arXiv:2402.13251*.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE.

Fang, Y.; Sun, Z.; Wu, T.; Wang, J.; Liu, Z.; Wetzstein, G.; and Lin, D. 2024. Make-it-real: Unleashing large multimodal model for painting 3d objects with realistic materials. *arXiv preprint arXiv:2404.16829*, 3.

Furukawa, Y.; and Ponce, J. 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8): 1362–1376.

He, Z.; and Wang, T. 2023. Openlrm: Open-source large reconstruction models.

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*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, P.; Liu, Y.; Long, X.; Zhang, F.; Lin, C.; Li, M.; Qi, X.; Zhang, S.; Luo, W.; Tan, P.; et al. 2024. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. *arXiv preprint arXiv:2405.11616*.

Li, S.; Li, C.; Zhu, W.; Yu, B.; Zhao, Y.; Wan, C.; You, H.; Shi, H.; and Lin, Y. 2023. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 1–13.

Li, Y.; He, X.; Jiang, Y.; Liu, H.; Tao, Y.; and Hai, L. 2022. MeshFormer: High-resolution Mesh Segmentation with Graph Transformer. In *Computer Graphics Forum*. Wiley Online Library.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023a. 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. 2023b. 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. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Pollefeys, M.; Nistér, D.; Frahm, J.-M.; Akbarzadeh, A.; Mordohai, P.; Clipp, B.; Engels, C.; Gallup, D.; Kim, S.-J.; Merrell, P.; et al. 2008. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78: 143–167.

Pollefeys, M.; Van Gool, L.; Vergauwen, M.; Verbiest, F.; Cornelis, K.; Tops, J.; and Koch, R. 2004. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59: 207–232.

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*. PMLR.

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*.

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*.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 501–518. Springer.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Shen, T.; Munkberg, J.; Hasselgren, J.; Yin, K.; Wang, Z.; Chen, W.; Gojcic, Z.; Fidler, S.; Sharp, N.; and Gao, J. 2023. Flexible Isosurface Extraction for Gradient-Based Mesh Optimization. *ACM Trans. Graph.*, 42(4): 37–1.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023a. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

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

Siddiqui, Y.; Monnier, T.; Kokkinos, F.; Kariya, M.; Kleiman, Y.; Garreau, E.; Gafni, O.; Neverova, N.; Vedaldi, A.; Shapovalov, R.; et al. 2024. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445*.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*. Springer.

Tang, S.; Zhang, F.; Chen, J.; Wang, P.; and Furukawa, Y. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. *arXiv*.

Tochilkin, D.; Pankratz, D.; Liu, Z.; Huang, Z.; Letts, A.; Li, Y.; Liang, D.; Laforte, C.; Jampani, V.; and Cao, Y.-P. 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wei, X.; Yan, Z.; Xiong, Z.; Hao, Y.; Qin, Y.; and Han, X. 2025. IntrinsicReal: Adapting IntrinsicAnything from Synthetic to Real Objects. *arXiv preprint arXiv:2509.00777*.

Wei, X.; Zhang, K.; Bi, S.; Tan, H.; Luan, F.; Deschaintre, V.; Sunkavalli, K.; Su, H.; and Xu, Z. 2024. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Yang, X.; Shi, H.; Zhang, B.; Yang, F.; Wang, J.; Zhao, H.; Liu, X.; Wang, X.; Lin, Q.; Yu, J.; et al. 2024. Hunyuan3D-1.0: A Unified Framework for Text-to-3D and Image-to-3D Generation. *arXiv preprint arXiv:2411.02293*.

Zeng, X.; Chen, X.; Qi, Z.; Liu, W.; Zhao, Z.; Wang, Z.; Fu, B.; Liu, Y.; and Yu, G. 2024a. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4252–4262.

Zeng, Z.; Deschaintre, V.; Georgiev, I.; Hold-Geoffroy, Y.; Hu, Y.; Luan, F.; Yan, L.-Q.; and Hašan, M. 2024b. Rgb↔x: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*.

Zhang, C.; Song, H.; Wei, Y.; Chen, Y.; Lu, J.; and Tang, Y. 2024a. Geolrm: Geometry-aware large reconstruction model for high-quality 3d gaussian generation. *arXiv preprint arXiv:2406.15333*.

Zhang, K.; Bi, S.; Tan, H.; Xiangli, Y.; Zhao, N.; Sunkavalli, K.; and Xu, Z. 2025. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, 1–19. Springer.

Zhang, L.; Wang, Z.; Zhang, Q.; Qiu, Q.; Pang, A.; Jiang, H.; Yang, W.; Xu, L.; and Yu, J. 2024b. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 43(4): 1–20.