DOI: 10.1111/cgf.70343

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

HiMat: DiT-based Ultra-High Resolution SVBRDF Generation

Zixiong Wang[1], Jian Yang[1], Yiwei Hu[2], Miloš Hašan[2,3], Beibei Wang[†4]

[1]College of Computer Science, Nankai University, [2]Adobe Research, [3]NVIDIA, [4]Nanjing University

Figure 1: *We present HiMat, a diffusion-based framework generating ultra-high-resolution (*4096×4096*) SVBRDF materials from text prompts. Our approach achieves this resolution while preserving high-frequency details crucial for meso-structure components such as normal and height maps. We showcase a variety of materials within a single scene, highlighting the preserved fine-scale details and texture fidelity.*

Abstract

Creating ultra-high-resolution spatially varying bidirectional reflectance functions (SVBRDFs) is critical for photorealistic 3D content creation, to faithfully represent fine-scale surface details required for close-up rendering. However, achieving 4K generation faces two key challenges: (1) the need to synthesize multiple reflectance maps at full resolution, which multiplies the pixel budget and imposes prohibitive memory and computational cost, and (2) the requirement to maintain strong pixel-level alignment across maps at 4K, which is particularly difficult when adapting pretrained models designed for the RGB image domain. We introduce HiMat, a diffusion-based framework tailored for efficient and diverse 4K SVBRDF generation. To address the first challenge, HiMat generates in a high-compression latent space via a DC-AE and employs a pretrained diffusion transformer with linear attention to improve per-map efficiency. To address the second challenge, we propose CrossStitch, a lightweight convolutional module that enforces cross-map consistency without incurring the cost of global attention. Our experiments show that HiMat achieves high-fidelity 4K SVBRDF generation with superior efficiency, structural consistency, and diversity compared to prior methods. Beyond materials, our framework also generalizes to related applications such as intrinsic decomposition.

CCS Concepts

• ***Computing methodologies*** → ***Reflectance modeling; Neural networks;***

† Corresponding author

1. Introduction

Modeling the reflectance properties of spatially-varying bidirectional reflectance functions (SVBRDFs) is a fundamental task in photorealistic rendering. To reduce production effort and the need for heavy manual intervention, automatic generation has emerged as a promising alternative. For such methods to achieve high-quality results, two requirements are particularly critical: ultra-high resolution (i. e. , 4K) and material diversity. High resolution enables fine-grained detail in close-up views. At the same time, diversity ensures broad coverage of real-world materials and supports scalable asset creation for applications such as games, visual effects, and architectural and product visualization.

Achieving both goals, however, remains challenging. First, 4K generation imposes extreme demands on GPU memory and computational throughput, limiting practical scalability. Furthermore, the limited scale of existing accessible SVBRDF datasets [DAD*18, MXZ*23, VD24] prevents generative models from capturing the vast appearance space of real-world materials, resulting in limited diversity and poor generalization.

While recent generative approaches have made progress, they still fall short of meeting these demands. Most existing techniques are based on generative adversarial networks (GANs) [GSH*20, ZHD*22, ZHD*23] or diffusion-based methods [HGZ*23, VSPS24, XGZM24, XZW*25]. These typically operate at low resolutions (e.g., 512×512) and rely on limited synthetic datasets, which restrict both quality and diversity. MatGen [VMR*24] pioneers 4K material generation with a cascaded denoising pipeline [DCH*24], but suffers from low efficiency and error accumulation. To improve diversity, MaterialPicker [MDH*25] fine-tunes pretrained video diffusion models to generate material maps rather than RGB frames, using softmax attention [VSP*17]. While such fine-tuning achieves diversity from the pretrained prior, the quadratic attention cost limits the input resolution to 256×256, making the approach impractical for 4K generation in the near future.

In this paper, we propose *HiMat*, a novel framework for efficient and diverse 4K SVBRDF generation. We identify two key challenges in 4K SVBRDF generation: (1) each of multiple reflectance maps must be generated at full 4K resolution, multiplying the pixel budget and causing prohibitive memory and compute cost, and (2) these maps must remain strictly pixel-aligned at 4K, a requirement that is particularly difficult when adapting pretrained image models designed for 3-channel RGB inputs.

To tackle the first issue, our core idea is to reduce the effective pixel budget and improve per-map processing efficiency. We achieve this by performing generation in a high-compression latent space and replacing quadratic attention with a more efficient alternative. Specifically, we leverage a deep compression autoencoder (DC-AE) [CCC*25] to compress 4K inputs while preserving key reflectance properties, and a linear diffusion transformer [XCC*25] to accelerate per-map generation at ultra-high resolution. For the second issue, our key insight is that SVBRDF maps, being inherently pixel-aligned and limited in number, allow consistency to be enforced without resorting to costly global attention. Based on this, we design *CrossStitch*, a lightweight convolution-based module enabling alignment across maps. Convolution is hardware-friendly and benefits from mature optimizations such as Winograd [LG16], making CrossStitch both efficient and scalable. This allows us to adapt pretrained image diffusion models to the SVBRDF domain, effectively exploiting strong priors while preserving inter-map consistency. Notably, CrossStitch is non-destructive and remains compatible with a wide range of latent diffusion architectures (e.g., U-Net).

Our results demonstrate that HiMat enables high-fidelity 4K SVBRDF generation with enhanced visual quality and reduced computational cost (see Fig. 1). Thanks to its lightweight design, HiMat produces 4096×4096 SVBRDFs in 90 seconds on consumer-grade hardware such as an NVIDIA RTX 4090D (in 20 steps). Further, our framework can be generalized to other tasks, including intrinsic decomposition. Our main contributions are:

- A memory-efficient and computationally scalable diffusion model for native ultra-high-resolution SVBRDF generation,
- A lightweight CrossStitch module that captures localized inter-map dependencies, enabling structural consistency across SVBRDF maps, and can be non-destructively added to latent diffusion models,

2. Related Work

Material creation and editing have a long history in graphics. Early efforts explored the use of text [MCS23] and images [DAD*19] to enable more intuitive authoring and control. Below, we focus on generative model-based approaches that aim to automate material generation through deep learning techniques. For a comprehensive overview of artistic authoring and editing techniques across appearance, we refer the readers to surveys [SPN*16, KHM*24].

2.1. Material Generation

Learning-based material generation has attracted increasing attention for simplifying material creation. Early works leveraged GANs [GSH*20, ZHD*22] trained on synthetic datasets. To improve realism, PhotoMat [ZHD*23] utilizes real materials captured with flash photographs. Based on the generator of PhotoMat, DiffMat [YYSF24] introduces an auxiliary diffusion network to enhance the latent representation of flash photographs, leading to improved reconstruction quality. Meanwhile, transformer-based approaches have been explored for procedural material generation [GHS*22, HGH*23], aiming to enhance scalability and diversity. However, due to the inherent limitations of GANs, these methods are difficult to scale to 4K SVBRDF generation.

Recently, diffusion-based methods have emerged as a promising direction for improving SVBRDF generation. Text2Mat [HGZ*23], DreamPBR [XZW*25], and ReflectanceFusion [XGZM24] adopt a dual-phase design for richer semantics: the first stage generates a latent representation of a natural image using pretrained image models, and the second stage recovers SVBRDF parameters from the retrained decoder. While this strategy enhances diversity, retraining the decoder on a limited SVBRDF dataset limits overall reconstruction quality. Moreover, DreamPBR employs a post-hoc super-resolution module, which often leads to over-smoothed outputs and missing some fine-grain details. In contrast, some methods aim to train models to

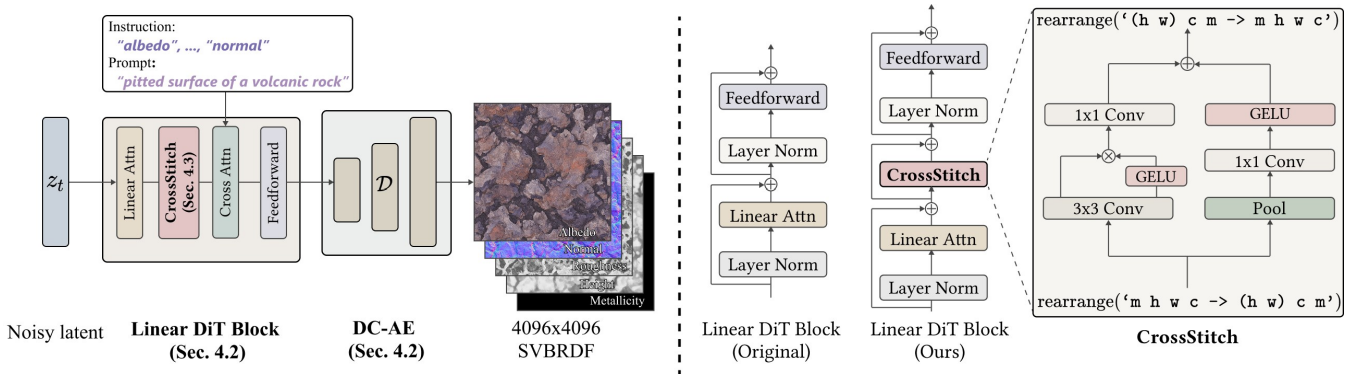

Figure 2: ***Overview. Left****: Given text instructions, our framework generates 4K SVBRDF maps through a latent denoising pipeline based on linear DiT (Sec. 4.2), with outputs reconstructed by a deep compression autoencoder (DC-AE) (Sec. 4.2).* CrossStitch *layers (Sec. 4.3) are integrated into the linear DiT block after each linear attention layer. The combination of linear DiT and DC-AE enables efficient ultra-high-resolution generation, while the CrossStitch design ensures consistency across maps.* ***Right****: Architecture of our modified DiT block (cross-attention omitted for clarity). A lightweight convolutional* CrossStitch *module enables localized feature exchange across maps, ensuring pixel alignment.*

generate SVBRDF parameters directly. Among these, MatFuse and MatGen [VSPS24, VMR*24] are trained from scratch in the style of latent diffusion models (LDMs) [RBL*22], with multiple encoders conditioned on text, image, or sketch inputs. However, due to limited training data, these models often struggle with diversity and generalization. MaterialPicker [MDH*25], on the other hand, fine-tunes a softmax-attention video diffusion model to generate SVBRDFs while retaining pretrained priors. Still, it is currently limited to a 256×256 resolution and requires a post hoc super-resolution module to upscale the outputs.

2.2. Diffusion Models Architectures

Diffusion models are generative frameworks for producing diverse, high-quality images [YZS*23], with architectures evolving from convolutional U-Nets augmented by self-attention [DN21, RDN*22, RBL*22]. Recently, Diffusion Transformers (DiT-s) [PX23] have emerged, leveraging pure attention-based architectures [VSP*17]. Stable Diffusion 3 [EKB*24], Flux [Lab24], Lumina-Image [QZX*25], and many other recent models extend this paradigm by improving quality, efficiency, and cross-modal alignment.

However, the quadratic complexity of softmax attention limits the scalability of DiT models, particularly for high-resolution image and video generation [Lab24, BPH*24, YTZ*25]. This has motivated the development of more efficient alternatives, such as linear attention [XCC*25] and state-space models [PDD*24, HBG*24]. In this work, we adopt a linear attention-based Diffusion Transformer, Sana [XCC*25, XCZ*25] for 4K SVBRDF generation, achieving high visual fidelity with significantly reduced memory and computational costs.

2.3. Ultra-High Resolution Diffusion Models

Generating 4K RGB images is crucial for many applications but remains computationally demanding. Existing approaches can be broadly divided into two categories: training-free cascade pipelines and model-based architectural adaptations.

Training-free cascade methods [DCH*24, KHZP25] progressively upscale diffusion outputs without retraining, using techniques such as residual connections and low-frequency injection. However, their multi-stage design introduces error accumulation and incurs high latency (e.g., DiffuseHigh [KHZP25] takes 258s per 4K image on an H100 GPU), making them impractical for generating multiple correlated SVBRDF maps.

In contrast, model-based strategies adapt diffusion architectures for 4K generation. PixArt-Σ [CGX*24] and Sana [XCC*25] redesign transformers for improved efficiency, while Diffusion4K [ZHL*25] combines high-compression VAEs with wavelet supervision but applies uniform weights across all frequency bands. URAE [YLTW25] further provides practical guidelines for ultra-resolution training. Unlike standard image generation, SVBRDF generation requires the simultaneous generation of multiple consistent maps. To this end, we adopt a model-based fine-tuning strategy and introduce a learnable CrossStitch module to enforce cross-map alignment during generation effectively.

3. Preliminaries

Latent Diffusion Models. Latent diffusion models (LDMs) [RBL*22] perform denoising in a compressed latent space, enabling efficient high-resolution synthesis. Specifically, a variational autoencoder (VAE) [KW14] encodes an input RGB image $x \in \mathbb{R}^{H \times W \times 3}$ into a latent representation $z_0 = \mathcal{E}(x) \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$, where $\mathcal{E}$ denotes the encoder. Here, H and W are the spatial dimensions of the input image, C is the number of latent channels, and $\hat{H} = \frac{H}{F}$ and $\hat{W} = \frac{W}{F}$ represent the height and width of the latent representation, respectively, downsampled by a factor of F.

The forward process corrupts latent z_0 by adding Gaussian noise:

$$z_t = \alpha_t \cdot z_0 + \sigma_t \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \tag{1}$$

with α_t, σ_t controlling the noise schedule.

In LDM [RBL*22], the denoising network Θ predicts the noise. Most recent variants (e.g. SD3 [EKB*24]) use velocity prediction via flow matching [LCBH*23]:

$$v_\Theta(z_t, t, c_{\text{text}}) = \epsilon - z_0. \tag{2}$$

The network Θ is typically a U-Net or DiT with self-attention for global context and cross-attention for conditioning.

SVBRDF. An SVBRDF typically represents the parameters of the Cook-Torrance microfacet model [CT82] with a GGX distribution function [WMLT07]. In our setting, an SVBRDF of size $H \times W$ consists of albedo $a \in \mathbb{R}^{H\times W\times 3}$, normal $n \in \mathbb{R}^{H\times W\times 3}$, roughness $r \in \mathbb{R}^{H\times W}$, metallicity $m \in \mathbb{R}^{H\times W}$, and height $h \in \mathbb{R}^{H\times W}$. This yields a set of maps:

$$I = \{a, n, r, m, h\}, \quad I \in \mathbb{R}^{M\times H\times W\times 3}. \tag{3}$$

For efficient processing, we concatenate the scalar maps r, m, and h into a single three-channel image, resulting in $M = 3$ maps.

4. Method

4.1. Overview

Our goal is to train a diffusion-based SVBRDF generator that produces 4K materials from text prompts. High-quality generation requires efficiency, diversity, and consistency across reflectance maps. The key challenges are twofold: (1) each map must be generated at full 4K resolution, and the presence of multiple maps multiplies the pixel budget, leading to prohibitive memory and computational cost, and (2) the maps are physically interdependent and must remain pixel-aligned at 4K, a requirement that is particularly difficult when adapting pretrained image models designed initially for 3-channel RGB inputs rather than multi-channel SVBRDFs.

To address these challenges, we propose *HiMat*, a framework for efficient, diverse, and consistent 4K SVBRDF generation (see Fig. 2). To overcome the first challenge, HiMat employs a high-compression autoencoder and a linear-attention diffusion transformer (Sec. 4.2) to reduce the effective pixel budget for 4K SVBRDF, substantially lowering memory consumption and computational cost. To address the second issue, HiMat introduces the CrossStitch module (Sec. 4.3), a lightweight convolution-based design that efficiently enforces pixel-level alignment across maps. In addition, to improve data quality and enhance material diversity, we incorporate prompt-based dataset augmentation with enriched textual descriptions (Sec. 4.4).

4.2. A Lightweight Latent Diffusion Model for 4K SVBRDFs

To mitigate the prohibitive pixel budgets of generating multiple 4K maps, HiMat integrates two complementary designs: a Deep Compression AutoEncoder (DC-AE) to reduce the effective pixel count, and a Linear-Attention Diffusion Transformer to accelerate per-map processing.

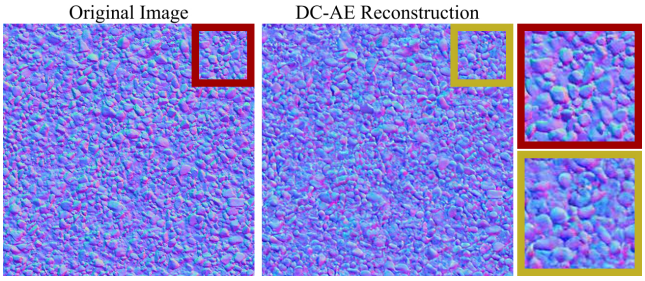

Figure 3: *Normal map reconstruction quality with DC-AE. Color bias in the reconstructed normals indicates distribution mismatch with ground truth, motivating our fine-tuning of the decoder for SVBRDF maps.*

High-Compression Autoencoder. High-resolution latent diffusion requires compact yet expressive representations. However, standard VAEs with F=8 compression [RBL*22] incur excessive overhead at 4K, often leading to OOM errors.

To address this, we adopt DC-AE [CCC*25], which achieves up to F=32 compression via residual encoding and resolution-aware adaptation. This reduces 4K inputs to 128×128 latent features, enabling tractable diffusion without compromising reconstruction fidelity.

However, since DC-AE is originally trained on natural image datasets, it may not fully capture the unique characteristics of SVBRDF maps, particularly for normal maps that require accurate preservation of geometric details and orientation information (see Fig. 3). To better align DC-AE with the SVBRDF domain, we follow latent diffusion models [RBL*22] and fine-tune the decoder using a combination of pixel-wise loss $\mathcal{L}_{\text{rec}}$ and perceptual loss $\mathcal{L}_{\text{LPIPS}}$ [ZIE*18]:

$$\mathcal{L}_{\text{vae}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}. \tag{4}$$

We omit adversarial loss due to its instability at ultra-high resolutions [CCC*25]. This adaptation yields a compact latent space while faithfully preserving SVBRDF-specific details essential for high-fidelity 4K generation.

Efficient Diffusion Transformers. DiT [PX23] adopts stacked softmax attention blocks [VSP*17] for global context modelling. Given an input sequence $s \in \mathbb{R}^{N\times C}$, where N denotes the sequence length and C the channel dimension, the softmax attention is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{5}$$

where Q, K, V are the query, key, and value matrices obtained via learned linear projections.

While effective, softmax attention incurs $\mathcal{O}(N^2)$ complexity, making it impractical for ultra-high resolution image and video generation [BPH*24, YTZ*25], as evidenced by the 256×256 resolution cap in MaterialPicker [MDH*25].

To improve scalability, HiMat adopts a linear attention-based

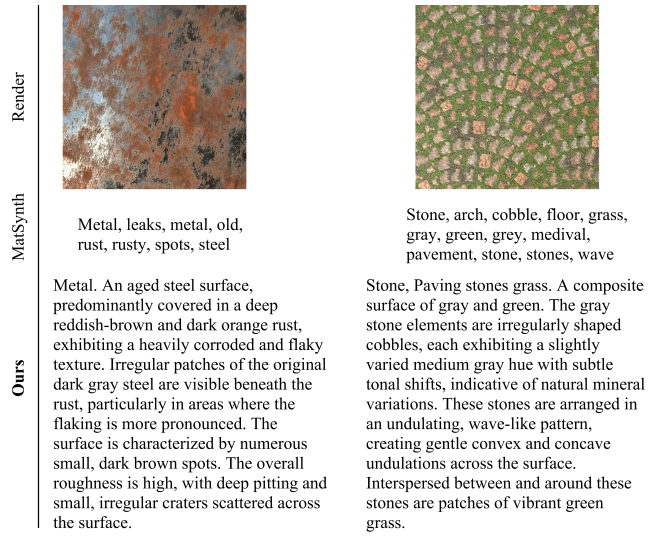

Figure 4: *Textual description comparison. MatSynth [VD24] provides short keyword labels, whereas our method generates rich, perceptually faithful descriptions that better capture material appearance and structure.*

DiT [XCC*25, XCZ*25], which reduces complexity to $\mathcal{O}(N)$:

$$\text{LinearAttention}(Q,K,V) = \frac{\text{ReLU}(Q)\left(\text{ReLU}(K)^\top V\right)}{\text{ReLU}(Q)\left(\text{ReLU}(K)^\top \mathbf{1}\right)}, \quad (6)$$

where $\mathbf{1}$ is an all-ones vector. This design preserves global receptive fields while significantly lowering memory footprint and runtime cost, enabling efficient scaling to 4K.

Together, DC-AE and linear-attention DiT directly alleviate the pixel burden, making high-resolution SVBRDF generation both tractable and efficient.

4.3. CrossStitch: Enforcing Cross-Map Consistency in 4K SVBRDFs

While the previous designs address the prohibitive cost of processing individual 4K maps, a second challenge is enforcing consistency across multiple reflectance maps. Since SVBRDF channels are physically coupled, even minor misalignments at 4K resolution can introduce shading discontinuities and implausible reflectance.

A natural idea is to adapt video-generation strategies, where softmax-attention layers enforce coherence across frames [MDH*25]. However, SVBRDF maps differ fundamentally from video: they are few in number, inherently pixel-aligned, and free of temporal drift. Consequently, softmax attention is both inefficient (quadratic cost) and unnecessary. Linear attention, although reducing complexity to linear in very long sequences, offers no advantage in this setting, as the small number of maps leaves the complexity effectively quadratic.

Our key insight is that SVBRDF maps only require local neighborhood communication rather than global attention. To this end, we design a *CrossStitch* module, a lightweight convolution-based layer that sparsely exchanges information across maps while preserving spatial alignment. Convolution is hardware-friendly and benefits from mature optimisations such as Winograd [LG16], making CrossStitch scalable to ultra-high resolutions.

Formally, given latent features $f \in \mathbb{R}^{M\times\hat{H}\times\hat{W}\times\hat{C}}$ after self-attention, where M is the number of SVBRDF maps and $\hat{C}$ the channel dimension, CrossStitch rearranges f to align the map dimension as channels, applies a 1D convolution across maps, and restores the original layout (using einops [Rog22]):

$$\begin{aligned} f &\leftarrow \text{rearrange}(f, \texttt{'m h w c -> (h w) c m'}) \\ f &\leftarrow \text{CrossStitch}(f) \\ f &\leftarrow \text{rearrange}(f, \texttt{'(h w) c m -> m h w c'}). \end{aligned} \quad (7)$$

In practice, *CrossStitch* is a dual-branch 1D convolutional module. One branch applies a depthwise-separable convolution (spatial 3×3 followed by pointwise 1×1) for efficient local feature mixing, while the other aggregates information across maps via average pooling, followed by a 1×1 convolution and GELU activation to capture shared semantic context. To integrate this module into the pretrained diffusion network, we insert it after the self-attention layer. All convolutional layers are zero-initialized and connected via residual connections to preserve pretrained representations. Although numerous convolutional variants exist [LLY*21], our simple yet effective design achieves both local integration and global alignment without incurring the overhead of self-attention.

CrossStitch is non-destructive, maintaining the structural integrity of each map while enforcing semantic and spatial coherence across them. This enables us to efficiently adapt pretrained image generation models to the material domain and leverage their learned priors for 4K SVBRDF generation.

4.4. Enhancing Material Diversity with Richer Text Prompts

While efficiency and consistency address computational bottlenecks, the diversity of generated materials remains constrained by the sparsity of supervision in existing datasets. Public SVBRDF datasets such as MatSynth [VD24] and the dataset proposed by Deschaintre et al. [DAD*18] provide only tag-style labels, whose descriptive power is insufficient to guide generative models effectively.

To overcome this limitation, we introduced prompt enrichment during both training and inference. During training, we expand the original tag annotations using large language models (LLMs) with carefully designed templates (detailed in the supplementary materials), inspired by DTDMat [CWH*24]. These templates explicitly capture intrinsic material attributes, such as color, texture, roughness, and surface imperfections, yielding more descriptive and diverse prompts that align better with generative objectives (Fig. 4).

At inference, we leverage a lightweight local LLM (Gemma-2 [TRP*24]) as the text encoder and reuse the same templates to generate system-level prompts that serve as semantic guides. This design draws inspiration from Lumina-Image-2.0 [QZX*25], which shows that system prompts can significantly improve generation quality without requiring architectural changes. By enriching supervision at both stages, our approach expands the diversity

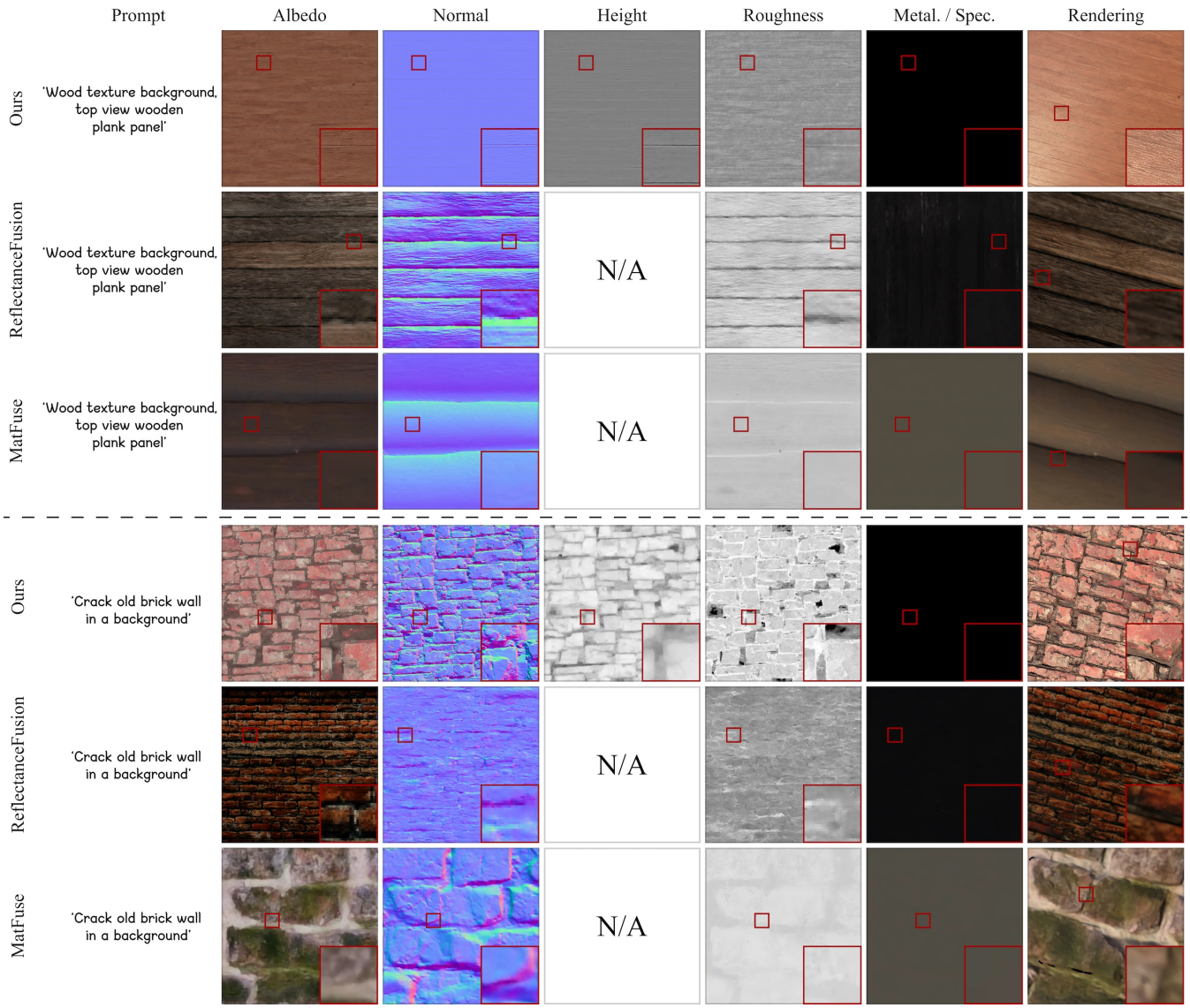

Figure 5: *Visual comparison between HiMat, ReflectanceFusion [XGZM24], and MatFuse [SP23]. ReflectanceFusion exhibits baked-in lighting artifacts and is limited to a resolution of* 256×256*. MatFuse suffers from reduced realism and diversity due to training exclusively on synthetic data at* 512×512 *resolution. In contrast, HiMat delivers high-quality 4K materials with fine detail. A slightly tilted camera view is employed in the rendering to visualize the details better.*

of generated materials and improves generalization to unseen categories, complementing the efficiency and consistency modules described earlier.

5. Results

5.1. Implementation Details

We trained our model on the combined datasets of MatSynth [VD24] and the dataset proposed by Deschaintre et al. [DAD*18], for a total of 6,198 unique physically-based rendering (PBR) materials. The text prompts are augmented with Gemini 2.5 [CBS*25]. To address the issue of uneven distribution for material types (e.g., Wood and Metal), we resample different categories to the same large number. For training, we initialize our model from the pre-trained Sana-1024px checkpoint [XCC*25] with 1.6B parameters. To effectively capture high-frequency material details while maintaining training stability, we employ a progressive-resolution strategy, gradually increasing the training resolution from 1024×1024 to 2048×2048 and finally to 4096×4096. To support tileable material generation, we additionally adopt noise rolling [VMR*24].

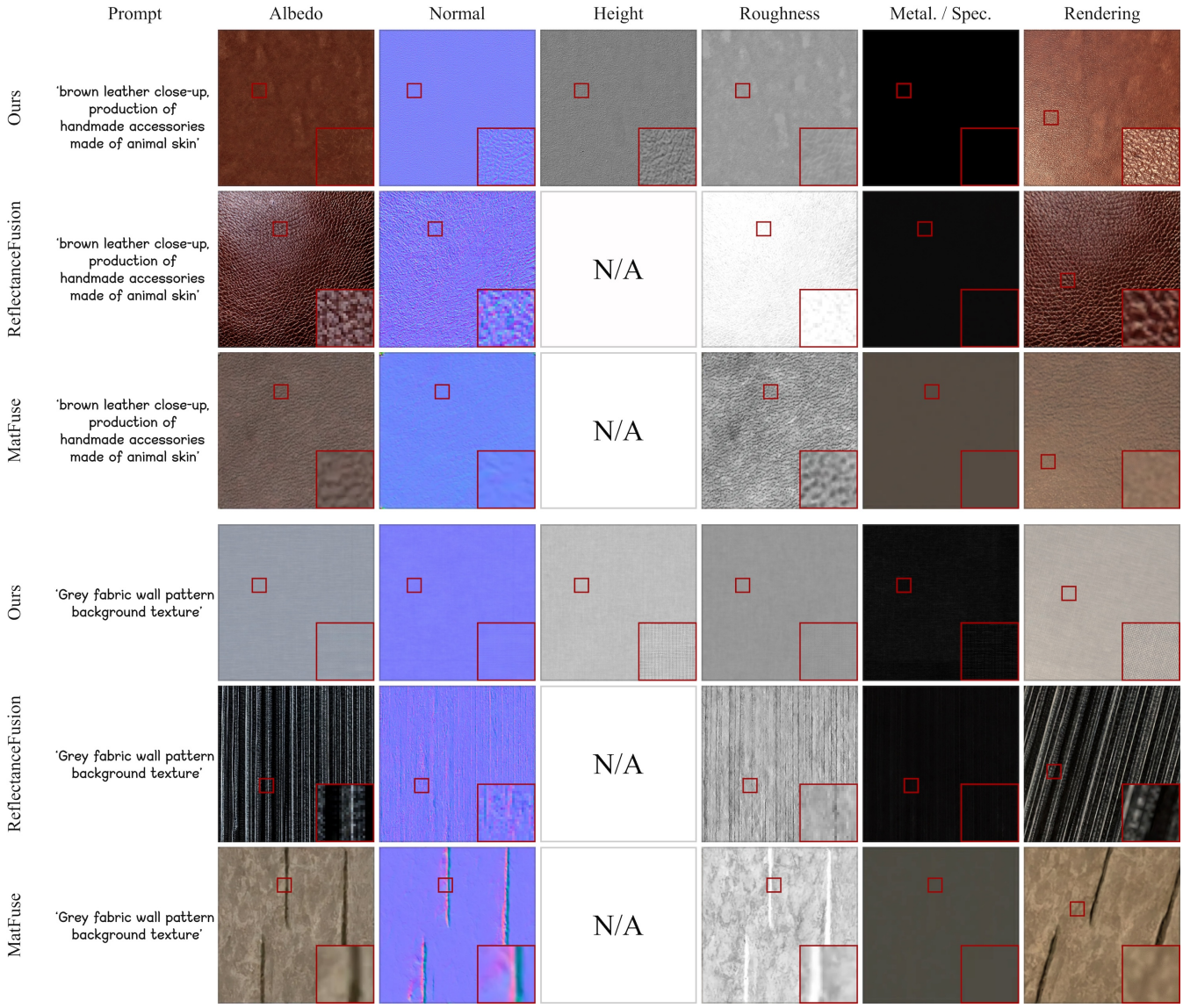

Figure 6: *Further visual comparison between HiMat, ReflectanceFusion [XGZM24] and MatFuse [VSPS24]. Note the baked-in lighting artifact at the top of the albedo map in the second row of the ReflectanceFusion result.*

5.2. Metrics

Since ground-truth data is unavailable for the generation task, we follow the evaluation protocol of MatFuse [VSPS24], which assesses SVBRDF quality through rendered images. Specifically, we render the generated SVBRDFs under environment lighting at 4096×4096 and evaluate them across three aspects: semantic alignment measured by CLIPScore [HHF*21], perceptual image quality measured by Q-Align [WZZ*24] (quality / aesthetic), respectively, and aesthetics score [SBV*22], and high-frequency detail preservation measured by the Gray Level Co-occurrence Matrix (GLCM) [ZHL*25]. Higher values indicate better performance for all metrics. The complete set of text prompts is provided in the supplemental materials.

5.3. Text Conditioned Generation

We compare our method against MatFuse [VSPS24] and ReflectanceFusion [XGZM24] on 500 text prompts (adopted from ReflectanceFusion, detailed in the supplementary material). ReflectanceFusion is limited to 256×256 and MatFuse to 512×512, whereas our approach natively supports 4K. As shown in Fig. 5 and 6, both baselines exhibit low-resolution artifacts that fail under close-up inspection. MatFuse produces less plausible and less diverse materials due to its exclusive reliance on synthetic training data, missing the strong priors learned from real images that both ReflectanceFusion and our approach exploit. ReflectanceFusion, by contrast, achieves relatively higher quality but suffers from baked-lighting artifacts since it relies on a base text-to-image model that

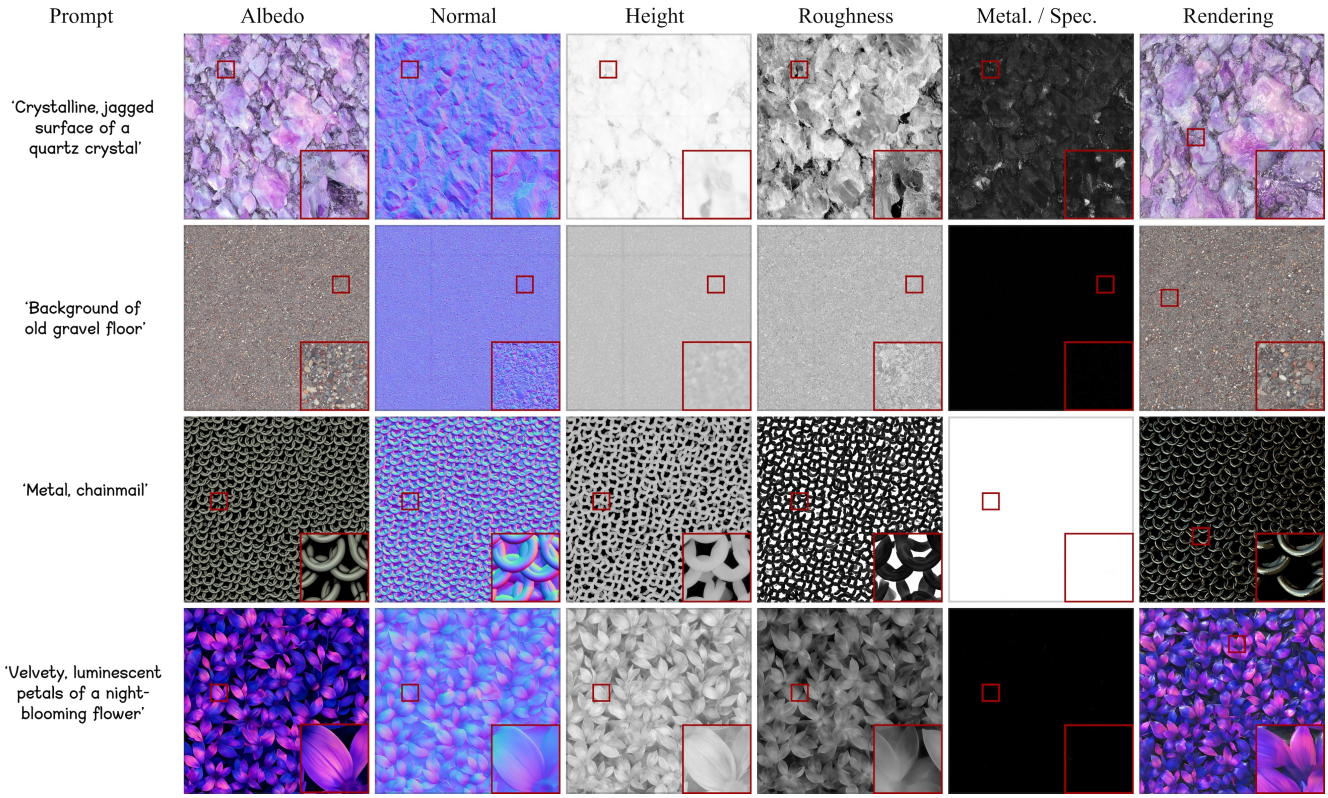

Figure 7: *Additional visual results from HiMat. These examples further demonstrate the diversity, realism, and fine structural details achieved by our method.*

generates RGB images with highlights and shadows, which are difficult to remove during its second-stage recovery (see the first column of Fig. 5). In comparison, HiMat generates diverse, realistic, and detailed 4K SVBRDFs that preserve fidelity even under close-up views.

In Tab. 1, we present quantitative comparisons across multiple metrics. MatFuse obtains the lowest CLIPScore and Q-Align scores, reflecting limited plausibility and diversity. ReflectanceFusion achieves a higher CLIPScore (30.22) owing to its base text-to-image backbone; however, its reliance on a 256×256 resolution yields weak structural fidelity, as evidenced by its low GLCM score. In contrast, HiMat achieves the best performance across nearly all metrics, demonstrating not only greater material diversity but also superior preservation of fine structural detail. Additional qualitative results are shown in Fig. 7.

Table 1: *Quantitative comparison for SVBRDF generation with MatFuse [VSPS24] and ReflectanceFusion [XGZM24].*

	MatFuse	ReflectanceFusion	Ours
CLIPScore ↑	25.91	30.22	**30.27**
Q-Align (quality) ↑	2.17	2.34	**3.23**
Q-Align (aesthetic) ↑	1.63	1.98	**2.33**
Aesthetics ↑	3.87	4.13	**4.63**
GLCM Score ↑	0.63	0.31	**0.96**

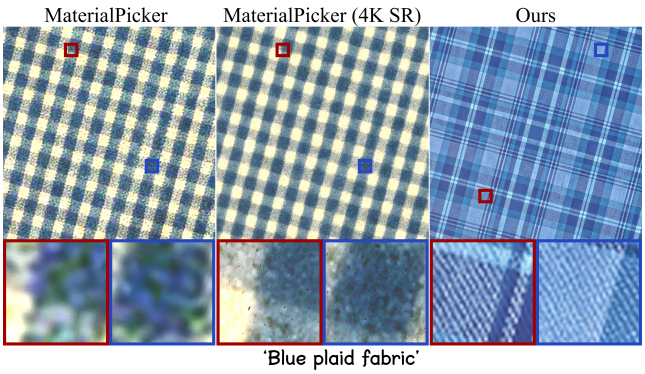

Figure 8: *Visual comparison between our native 4K SVBRDF generation and the outputs of MaterialPicker, including its upscaled 4K version using SUPIR [YGL*24]. Recovering fine fabric microstructures from low-resolution MaterialPicker outputs is challenging, underscoring the advantages of directly generating high-resolution results.*

MaterialPicker vs. Ours. We compare our method against MaterialPicker [MDH*25], which leverages video diffusion priors by fine-tuning a pretrained model for material generation. Since its output is limited to 256×256, we apply SUPIR [YGL*24] to upscale each map to 4096×4096. As shown in Fig. 8, our native 4K generation preserves fine details significantly better, whereas recovering fabric microstructures from the low-resolution outputs of MaterialPicker remains challenging. This highlights the advantages of directly generating high-resolution results.

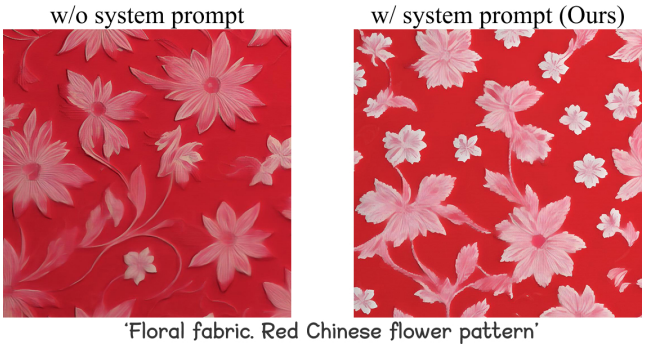

Figure 9: *Comparison of generations with and without system-level prompts. Incorporating our designed templates yields finer structure and detail.*

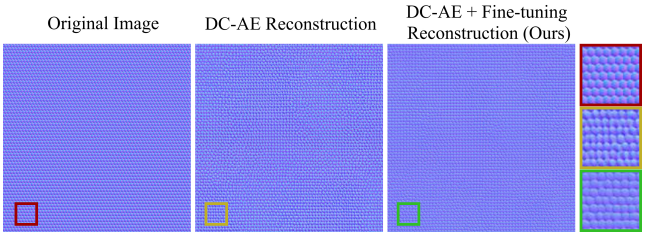

Figure 10: *Effect of DC-AE fine-tuning. Fine-tuned models produce normal maps with improved spatial structure and orientations. Zoom in for better visualization.*

5.4. Ablation studies and Analysis

Textual Prompt Enrichment. To evaluate the effectiveness of textual prompt enrichment, we render materials from the original dataset and compare the original tag-based descriptions with our enhanced long-form prompts using CLIP-Score [HHF*21]. While MatSynth achieves a CLIP-Score of 27.74, our enriched prompts yield a score of 29.26, indicating improved textual-image alignment through simple yet targeted expansions. Furthermore, we assess the impact of system-level prompts during inference by comparing generations with and without our designed template. As shown in Fig. 9, incorporating system-level prompts yields richer texture detail without additional efforts.

VAE Fine-Tuning. To assess the benefits of fine-tuning the DC-AE decoder, we compare models before and after fine-tuning. As reported in Tab. 2, our fine-tuned variant consistently improves reconstruction quality across multiple metrics, including Fréchet Inception Distance (rFID), Peak Signal-to-Noise Ratio (PSNR), Root Mean Square Error (RMSE), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). Visual comparisons in Fig.10 further confirm fine-tuning enhanced structural fidelity on 4K MatSynth [VD24] results.

Table 2: *Quantitative evaluation of decoder fine-tuning for DC-AE on the MatSynth [VD24] dataset at 4K resolution.*

Model	rFID ↓	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DC-AE	1.50	0.10	28.71	0.75	0.16
DC-AE + fine-tuning	**1.29**	**0.08**	**30.28**	**0.79**	**0.13**

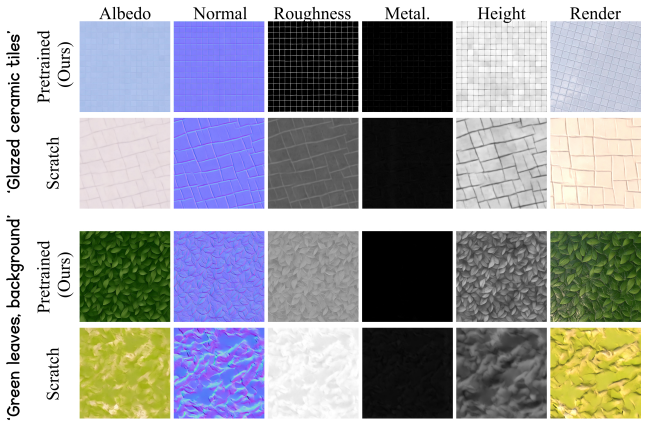

Figure 11: *Effect of pretrained priors. Models trained from scratch exhibit limited diversity, whereas pretrained initialization enables higher-quality and more diverse material generation.*

Pretraining vs. Training from Scratch. To validate the impact of pretrained prior, we compare two 1024×1024 models with identical architectures: one trained from scratch and the other initialized from our pretrained weights. As shown in Fig. 11, the scratch model collapses to categories in the trainset and fails to generate diverse types such as 'leaves'. In contrast, the pretrained model achieves higher quality and better diversity, highlighting the importance of strong priors.

CrossStitch vs. Attention. We evaluate our CrossStitch-based architecture against standard softmax attention and linear attention baselines. We first analyze performance across resolutions, then compare training dynamics by training all three variants at 1024×1024 under identical settings. In Tab. 3, we replace CrossStitch with either standard or linear attention and report parameter count, forward FLOPs, peak memory usage, and inference time per step, all measured on an RTX 4090D GPU. Our CrossStitch model consistently reduces computational overhead: at 1024×1024 and 2048×2048, it achieves up to 22% fewer FLOPs and 25% less memory than linear attention, while also running faster. At 4096×4096, both standard and linear attention incur prohibitive costs, with linear attention leading to out-of-memory failures.

Table 3: *Comparison of computational cost during inference. All results are measured on a consumer-level RTX 4090D GPU. We report both absolute results and relative ratios (normalized by CrossStitch variant, denoted as 1.00) for variants with attention and linear attention.*

	Params (B)	Forward FLOPs (T)↓			Memory (GB)↓			Time (s/step)↓		
		1024×1024	2048×2048	4096×4096	1024×1024	2048×2048	4096×4096	1024×1024	2048×2048	4096×4096
Attention	2.01 / 1.14	11.20 / 1.21	43.68 / 1.22	173.58 / 1.22	10.89 / 1.05	12.76 / 1.01	20.44 / 1.03	0.39 / 1.30	1.23 / 1.29	5.13 / 1.28
Lin. ttn.	2.01 / 1.14	11.22 / 1.21	43.75 / 1.22	173.88 / 1.22	11.59 / 1.12	15.80 / 1.25	OOM	0.43 / 1.43	1.30 / 1.37	-
Ours (CrossStitch)	**1.76 / 1.00**	**9.25 / 1.00**	**35.87 / 1.00**	**142.36 / 1.00**	**10.39 / 1.00**	**12.63 / 1.00**	**19.93 / 1.00**	**0.30 / 1.00**	**0.95 / 1.00**	**4.01 / 1.00**

Table 4: *Quantitative comparison of intrinsic decomposition performance on the Hypersim [RRR*21] test dataset. Our retrained results demonstrate that CrossStitch delivers consistent improvements under identical data conditions. Note that our method does not match the original RGB↔X results, as the original model was trained on a large-scale dataset.*

		RGB↔X (retrained) w/ CS	RGB↔X (retrained) w/o CS
PSNR↑	Albedo	**13.16**	12.40
	Normal	**15.07**	13.83
	Irradiance	**16.42**	15.94
	Mean	**14.89**	14.05
LPIPS↓	Albedo	**0.43**	0.47
	Normal	**0.40**	0.48
	Irradiance	**0.37**	0.41
	Mean	**0.40**	0.45
DreamSim ↓	Albedo	**0.21**	0.24
	Normal	**0.13**	0.17
	Irradiance	**0.18**	0.21
	Mean	**0.17**	0.21

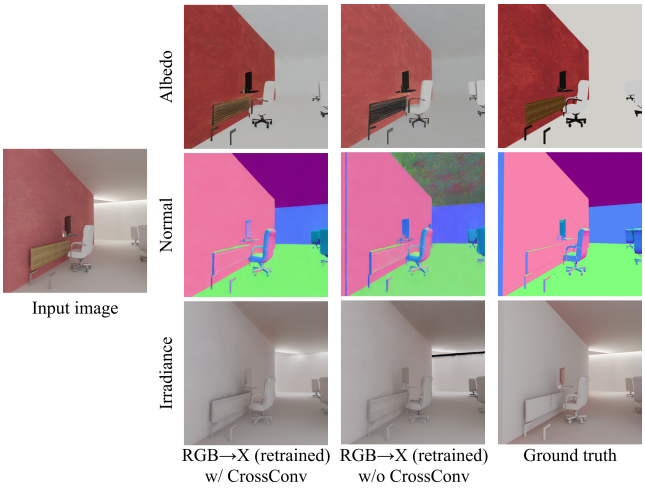

Figure 12: *Effect of CrossStitch on intrinsic decomposition. By retraining RGB↔X [ZDG*24] on Hypersim [RRR*21], we demonstrate that CrossStitch effectively improves structural consistency across decomposed outputs, even when applied to a U-Net denoising backbone, leading to better overall reconstruction.*

CrossStitch for Intrinsic Decomposition. To evaluate the adaptability of CrossStitch beyond material generation, we apply it to the intrinsic decomposition task on the Hypersim dataset [RRR*21], which contains over 74K indoor images with ground-truth albedo, normal, and irradiance maps. We retrained RGB↔X [ZDG*24] based on Stable Diffusion 2.1 [RBL*22], with and without our CrossStitch module. As shown in Tab. 4 and Fig. 12, incorporating CrossStitch consistently improves performance in both PSNR, LPIPS, and DreamSim [FTS*23] by enabling structural consistency across predicted maps. Despite using only the publicly available training split (unlike RGB↔X, which uses larger private datasets), our model demonstrates strong generalization, validating CrossStitch's plug-and-play applicability across architectures (e.g., U-Net) and tasks beyond DiT-based material generation.

Effects of Different Branches in CrossStitch. To analyze the effects of each component within the CrossStitch module, we conduct an ablation study by selectively turning its two branches on or off: (i) global branch: the average-pooling branch for global context aggregation, and (ii) local branch: the convolution branch for localized inter-map interactions. To quantitatively evaluate consistency, we set up the model to generate multiple output maps that are all intended to replicate the same albedo map, each associated with a different switcher name (e.g., albedo_1, albedo_2). Since all outputs originate from the same underlying signal, the ideal outputs should be identical, allowing us to directly assess consistency using image similarity metrics such as PSNR, LPIPS, and DreamSim [FTS*23]. From Tab. 5 and Fig. 13, we observe that using either branch alone already enforces a reasonable level of consistency, while enabling both branches jointly more effectively suppresses color bias and yields the most consistent results across maps, demonstrating the full potential of the CrossStitch module for multi-map alignment.

Effect of Denoising Steps. We analyze the impact of the number of denoising steps during 4K-resolution generation. Specifically, we conduct experiments using 5, 10, 20, 30, 50, and 100 steps. As shown in Fig. 14, increasing the number of steps improves the preservation of fine details, but it also incurs a higher computational cost. By default, we adopt 20 steps to balance quality and efficiency, while allowing users to adjust this parameter according to application-specific requirements.

5.5. Discussion and limitations

Our method has several limitations. First, we observe that noise-rolling occasionally introduces horizontal and vertical streaks in

Table 5: *Quantitative comparisons of different branch configurations within the CrossStitch module. Lower LPIPS and DreamSim scores indicate stronger perceptual consistency, highlighting the module's potential to improve semantic alignment in multi-map generation tasks.*

	Pair	PSNR↑	LPIPS↓	DreamSim↓
Global branch	1↔2	33.13	0.07	0.0019
	1↔3	33.48	0.07	0.0019
	2↔3	32.32	0.07	0.0026
	Mean	32.98	0.07	0.0021
Local branch	1↔2	32.01	0.06	0.0017
	1↔3	31.45	0.07	0.0038
	2↔3	31.79	0.06	0.0020
	Mean	31.75	0.06	0.0025
Both	1↔2	34.85	0.06	0.0014
	1↔3	34.58	0.07	0.0016
	2↔3	34.59	0.07	0.0012
	Mean	**34.67**	**0.06**	**0.0014**

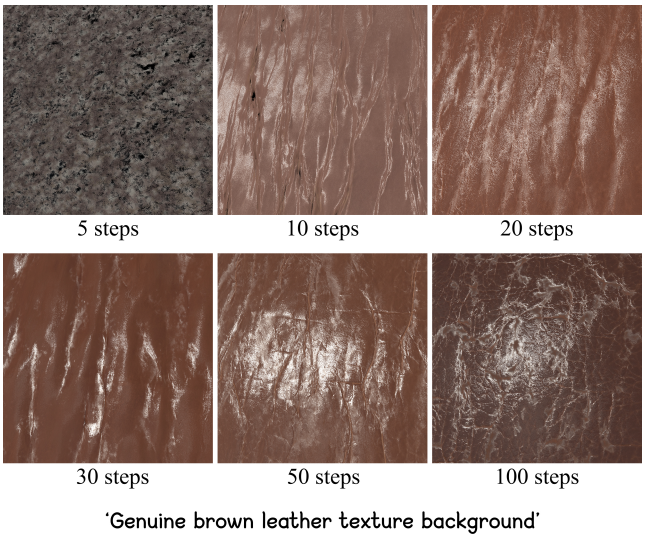

Figure 14: *Visual comparison across different denoising steps. Increasing the number of sampling steps yields greater texture fidelity, including finer details.*

4K generation (see supplementary material for examples). Additional denoising stages can mitigate this artifact, though at the cost of increased inference time. Second, even with decoder fine-tuning, DC-AE could be further improved in preserving low-frequency structures [CZH*25]. Finally, although our current implementation is text-conditioned, the framework readily generalizes to multimodal inputs (e.g., image prompts or ControlMat-style controls [VMR*24]), offering greater flexibility and diversity, as with other diffusion-based methods. We leave this exploration for future work.

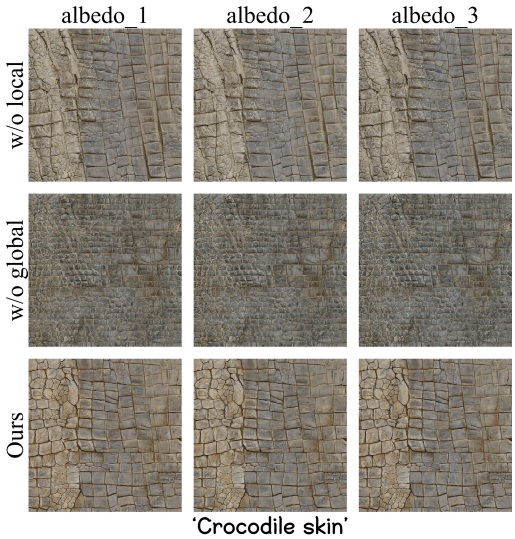

Figure 13: *Visual results showing the effect of CrossStitch's two branches. When both branches are enabled, the model produces more consistent results with reduced color-bias.*

6. Conclusion

We propose *HiMat*, a lightweight diffusion framework tailored for 4K SVBRDF generation. By combining a deep compression autoencoder with a linear-attention DiT, HiMat reduces the prohibitive pixel budget and scales generation efficiently to ultra-high resolution. To address the challenge of strict pixel alignment across reflectance maps, we introduce *CrossStitch*, a convolutional module that enforces inter-map consistency while remaining non-destructive and compatible with standard diffusion backbones. Extensive experiments demonstrate that HiMat produces diverse, high-fidelity 4K SVBRDFs with practical runtime on consumer GPUs, establishing a new baseline for scalable material generation. Beyond materials, HiMat also generalizes naturally to related tasks such as intrinsic decomposition, highlighting its potential as a versatile foundation for efficient DiT-based pipelines in digital content creation.

Acknowledgments

We thank the reviewers for the valuable comments. This work has been partially supported by the National Natural Science Foundation of China under grant No. 62572230.

References

[BPH*24] BROOKS T., PEEBLES B., HOLMES C., DEPUE W., GUO Y., JING L., SCHNURR D., TAYLOR J., LUHMAN T., LUHMAN E., NG C., WANG R., RAMESH A.: Video generation models as world simulators. URL: https://openai.com/research/video-generation-models-as-world-simulators. 3, 4

[CBS*25] COMANICI G., BIEBER E., SCHAEKERMANN M., PASUPAT I., SACHDEVA N., DHILLON I., BLISTEIN M., RAM O., ZHANG D., ROSEN E., ET AL.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025). 6

[CCC*25] CHEN J., CAI H., CHEN J., XIE E., YANG S., TANG H., LI M., HAN S.: Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations* (2025). 2, 4

[CGX*24] CHEN J., GE C., XIE E., WU Y., YAO L., REN X., WANG Z., LUO P., LU H., LI Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision* (2024), Springer, pp. 74–91. 3

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG) 1*, 1 (1982), 7–24. 4

[CWH*24] CHEN M., WANG Y., HU D., ZHU P., GUO J., GUO Y.: Dtdmat: A comprehensive svbrdf dataset with detailed text descriptions. In *The 19th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry* (2024), pp. 1–15. 5

[CZH*25] CHEN J., ZOU D., HE W., CHEN J., XIE E., HAN S., CAI H.: Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. In *IEEE International Conference on Computer Vision (ICCV)* (2025). 11

[DAD*18] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG) 37*, 4 (2018), 1–15. 2, 5, 6

[DAD*19] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Flexible svbrdf capture with a multi-image deep network. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) 38*, 4 (July 2019). URL: http://www-sop.inria.fr/reves/Basilic/2019/DADDB19. 2

[DCH*24] DU R., CHANG D., HOSPEDALES T., SONG Y.-Z., MA Z.: Demofusion: Democratising high-resolution image generation with no $$$. In *CVPR* (2024). 2, 3

[DN21] DHARIWAL P., NICHOL A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems 34* (2021), 8780–8794. 3

[EKB*24] ESSER P., KULAL S., BLATTMANN A., ENTEZARI R., MÜLLER J., SAINI H., LEVI Y., LORENZ D., SAUER A., BOESEL F., ET AL.: Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning* (2024). 3, 4

[FTS*23] FU S., TAMIR N., SUNDARAM S., CHAI L., ZHANG R., DEKEL T., ISOLA P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems* (2023), vol. 36, pp. 50742–50768. 10

[GHS*22] GUERRERO P., HASAN M., SUNKAVALLI K., MECH R., BOUBEKEUR T., MITRA N.: Matformer: A generative model for procedural materials. *ACM Trans. Graph. 41*, 4 (2022). doi:10.1145/3528223.3530173. 2

[GSH*20] GUO Y., SMITH C., HAŠAN M., SUNKAVALLI K., ZHAO S.: Materialgan: Reflectance capture using a generative svbrdf model. *ACM Trans. Graph. 39*, 6 (2020), 254:1–254:13. 2

[HBG*24] HU V. T., BAUMANN S. A., GUI M., GREBENKOVA O., MA P., SCHUSTERBAUER J., OMMER B.: Zigma: A dit-style zigzag mamba diffusion model. In *ECCV* (2024). 3

[HGH*23] HU Y., GUERRERO P., HASAN M., RUSHMEIER H., DESCHAINTRE V.: Generating Procedural Materials from Text or Image Prompts. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023). 2

[HGZ*23] HE Z., GUO J., ZHANG Y., TU Q., CHEN M., GUO Y., WANG P., DAI W.: Text2Mat: Generating Materials from Text. In *Pacific Graphics Short Papers and Posters* (2023), The Eurographics Association. 2

[HHF*21] HESSEL J., HOLTZMAN A., FORBES M., BRAS R. L., CHOI Y.: Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021). 7, 9

[KHM*24] KAVOOSIGHAFI B., HAJISHARIF S., MIANDJI E., BARAVDISH G., CAO W., UNGER J.: Deep svbrdf acquisition and modelling: A survey. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15199. 2

[KHZP25] KIM Y., HWANG G., ZHANG J., PARK E.: Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence* (2025), vol. 39, pp. 4338–4346. 3

[KW14] KINGMA D. P., WELLING M.: Auto-Encoding Variational Bayes. In *The 2nd International Conference on Learning Representations* (2014). 3

[Lab24] LABS B. F.: Flux. https://github.com/black-forest-labs/flux, 2024. 3

[LCBH*23] LIPMAN Y., CHEN R. T. Q., BEN-HAMU H., NICKEL M., LE M.: Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations* (2023). 4

[LG16] LAVIN A., GRAY S.: Fast algorithms for convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4013–4021. 2, 5

[LLY*21] LI Z., LIU F., YANG W., PENG S., ZHOU J.: A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems 33*, 12 (2021), 6999–7019. 5

[MCS23] MEMERY S., CEDRON O., SUBR K.: Generating parametric brdfs from natural language descriptions. In *Computer graphics forum* (2023), vol. 42, Wiley Online Library, p. e14980. 2

[MDH*25] MA X., DESCHAINTRE V., HAŠAN M., LUAN F., ZHOU K., WU H., HU Y.: Materialpicker: Multi-modal material generation with diffusion transformers. *ACM Trans. Graph.* (July 2025). 2, 3, 4, 5, 9

[MXZ*23] MA X., XU X., ZHANG L., ZHOU K., WU H.: Opensvbrdf: a database of measured spatially-varying reflectance. *ACM Transactions on Graphics (TOG) 42*, 6 (2023), 1–14. 2

[PDD*24] PHUNG H., DAO Q., DAO T., PHAN H., METAXAS D., TRAN A.: Dimsum: Diffusion mamba - a scalable and unified spatial-frequency method for image generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems* (2024). 3

[PX23] PEEBLES W., XIE S.: Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 4195–4205. 3, 4

[QZX*25] QIN Q., ZHUO L., XIN Y., DU R., LI Z., FU B., LU Y., LI X., LIU D., ZHU X., BEDDOW W., MILLON E., VICTOR PEREZ W. W., QIAO Y., ZHANG B., LIU X., LI H., XU C., GAO P.: Lumina-image 2.0: A unified and efficient image generative framework, 2025. 3, 5

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695. 3, 4, 10

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents, 2022. URL: https://arxiv.org/abs/2204.06125, arXiv:2204.06125. 3

[Rog22] ROGOZHNIKOV A.: Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations* (2022). 5

[RRR*21] ROBERTS M., RAMAPURAM J., RANJAN A., KUMAR A., BAUTISTA M. A., PACZAN N., WEBB R., SUSSKIND J. M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 10912–10922. 10

[SBV*22] SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M., ET AL.: Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems 35* (2022), 25278–25294. 7

[SP23] SARTOR S., PEERS P.: Matfusion: a generative diffusion model for svbrdf capture. In *ACM SIGGRAPH Asia Conference Proceedings* (December 2023). URL: https://doi.org/10.1145/3610548.3618194. 6

[SPN*16] SCHMIDT T.-W., PELLACINI F., NOWROUZEZAHRAI D., JAROSZ W., DACHSBACHER C.: State of the art in artistic editing of appearance, lighting and material. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 216–233. 2

[TRP*24] TEAM G., RIVIERE M., PATHAK S., SESSA P. G., HARDIN C., BHUPATIRAJU S., HUSSENOT L., MESNARD T., SHAHRIARI B., RAMÉ A., ET AL.: Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024). 5

[VD24] VECCHIO G., DESCHAINTRE V.: Matsynth: A modern pbr materials dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024). 2, 5, 6, 9

[VMR*24] VECCHIO G., MARTIN R., ROULLIER A., KAISER A., ROUFFET R., DESCHAINTRE V., BOUBEKEUR T.: Controlmat: A controlled generative approach to material capture. *ACM Trans. Graph. 43*, 5 (sep 2024). doi:10.1145/3688830. 2, 3, 6, 11

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.: Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 6000–6010. 2, 3, 4

[VSPS24] VECCHIO G., SORTINO R., PALAZZO S., SPAMPINATO C.: Matfuse: Controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 4429–4438. 2, 3, 7, 8

[WMLT07] WALTER B., MARSCHNER S. R., LI H., TORRANCE K. E.: Microfacet models for refraction through rough surfaces. *Rendering techniques 2007* (2007), 18th. 4

[WZZ*24] WU H., ZHANG Z., ZHANG W., CHEN C., LIAO L., LI C., GAO Y., WANG A., ZHANG E., SUN W., YAN Q., MIN X., ZHAI G., LIN W.: Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning* (2024), ICML'24, JMLR.org. 7

[XCC*25] XIE E., CHEN J., CHEN J., CAI H., TANG H., LIN Y., ZHANG Z., LI M., ZHU L., LU Y., HAN S.: SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations* (2025). 2, 3, 5, 6

[XCZ*25] XIE E., CHEN J., ZHAO Y., YU J., ZHU L., LIN Y., ZHANG Z., LI M., CHEN J., CAI H., LIU B., ZHOU D., HAN S.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *International Conference on Machine Learning* (January 2025). 3, 5

[XGZM24] XUE B., GUARNERA C., ZHAO S., MONTAZERI Z.: Reflectancefusion: Diffusion-based text to svbrdf generation. In *Eurographics Symposium on Rendering* (2024), Eurographics Association. 2, 6, 7, 8

[XZW*25] XIN L., ZHANG Z., WEI J., GAO W., GAO D.: Dreampbr: Text-driven generation of high-resolution svbrdf with multi-modal guidance. *IEEE International Conference on Multimedia & Expo(ICME)* (2025). 2

[YGL*24] YU F., GU J., LI Z., HU J., KONG X., WANG X., HE J., QIAO Y., DONG C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 25669–25680. 8, 9

[YLTW25] YU R., LIU S., TAN Z., WANG X.: Ultra-resolution adaptation with ease. *International Conference on Machine Learning* (2025). 3

[YTZ*25] YANG Z., TENG J., ZHENG W., DING M., HUANG S., XU J., YANG Y., HONG W., ZHANG X., FENG G., YIN D., YUXUAN.ZHANG, WANG W., CHENG Y., XU B., GU X., DONG Y., TANG J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations* (2025). 3, 4

[YYSF24] YUAN L., YAN D., SAITO S., FUJISHIRO I.: Diffmat: Latent diffusion models for image-guided material generation. *Visual Informatics 8*, 1 (2024), 6–14. 2

[YZS*23] YANG L., ZHANG Z., SONG Y., HONG S., XU R., ZHAO Y., ZHANG W., CUI B., YANG M.-H.: Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys 56*, 4 (2023), 1–39. 3

[ZDG*24] ZENG Z., DESCHAINTRE V., GEORGIEV I., HOLD-GEOFFROY Y., HU Y., LUAN F., YAN L.-Q., HAŠAN M.: RGB↔X: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers* (New York, NY, USA, 2024), SIGGRAPH '24, Association for Computing Machinery. doi:10.1145/3641519.3657445. 10

[ZHD*22] ZHOU X., HASAN M., DESCHAINTRE V., GUERRERO P., SUNKAVALLI K., KALANTARI N. K.: Tilegen: Tileable, controllable material generation and capture. In *SIGGRAPH Asia 2022 Conference Papers* (New York, NY, USA, 2022), SA '22, Association for Computing Machinery. 2

[ZHD*23] ZHOU X., HAŠAN M., DESCHAINTRE V., GUERRERO P., HOLD-GEOFFROY Y., SUNKAVALLI K., KALANTARI N. K.: Photomat: A material generator learned from single flash photos. In *SIGGRAPH 2023 Conference Papers* (2023). 2

[ZHL*25] ZHANG J., HUANG Q., LIU J., GUO X., HUANG D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025). 3, 7

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018). 4