ReLi3D: Relightable Multi-view 3D Reconstruction with Disentangled Illumination

Anonymous authors
Paper under double-blind review

Abstract

Reconstructing 3D assets from images has long required separate pipelines for geometry reconstruction, material estimation, and illumination recovery, each with distinct limitations and computational overhead. We present ReLi3D, the first unified end-to-end pipeline that simultaneously reconstructs complete 3D geometry, spatially-varying physically-based materials, and environment illumination from sparse multi-view images in under one second. Our key insight is that multi-view constraints can dramatically improve material and illumination disentanglement, a problem that remains fundamentally ill-posed for single-image methods. Key to our approach is the fusion of the multi-view input via a transformer cross-conditioning architecture, followed by a novel unified two-path prediction strategy. The first path predicts the object's structure and appearance, while the second path predicts the environment illumination from image background or object reflections. This, combined with a differentiable Monte Carlo multiple importance sampling renderer, creates an optimal illumination disentanglement training pipeline. In addition, with our mixed domain training protocol, which combines synthetic PBR datasets with real-world RGB captures, we establish generalizable results in geometry, material accuracy, and illumination quality. By unifying previously separate reconstruction tasks into a single feed-forward pass, we enable near-instantaneous generation of complete, relightable 3D assets.

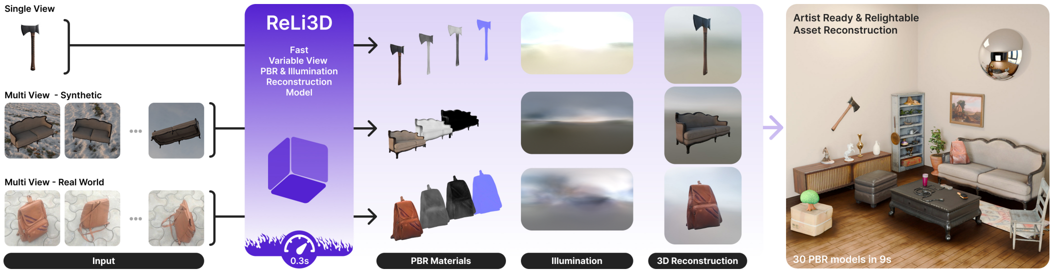

Figure 1: **Fast, illumination disentangled reconstructions.** ReLi3D reconstructs high-quality 3D meshes with physically based materials from sparse input images, while disentangling illumination effects; all in just 0.3s. It is robustly trained on cross-domain datasets and excels in both single- and multi-view cases, on synthetic data as well as on real-world examples.

1 Introduction

Reconstructing production-ready 3D assets from images remains a challenging task with immense potential for industrial design, interactive media, or robotics. Two lines of progress have emerged: (i) Generative models based on diffusion, which can achieve striking geometric fidelity, but with long inference times and hallucination, (ii) Large Reconstruction Models (LRMs) such as LRM (Hong et al., 2023), SF3D (Boss et al., 2024), and TripoSR (Tochilkin et al., 2024a) that perform direct feed-forward inference from images to 3D. While LRMs are fast and practical, a gap persists between research prototypes and what artists require from a 3D reconstruction, which is accurate reconstruction from multiple views and illumination disentanglement resulting in spatially varying Physically Based Rendering (PBR) materials that support relighting.

Unfortunately, many existing approaches optimize only for single-view reconstruction, which is inherently ill-posed. The same 2D appearance can arise from numerous combinations of surface reflectance and illumination. Regularization or learned priors help, but ambiguity remains, especially in unobserved areas, leading to incomplete spatially varying material predictions, unreliable normals, and therefore limited relighting fidelity.

In our perspective, geometric consistency across multiple views provides the missing constraints to separate material properties from lighting effects. When multiple observations see the same surface point under a common illumination, cross-view agreement narrows the feasible solution space and turns an ill-posed single-view problem into a much better constrained one. To operationalize this, we design an architecture where multi-view fusion is not an add-on for robustness, but the primary mechanism for material-lighting disentanglement.

In this paper, we present ReLi3D, a unified feed-forward system that turns a variable number of posed images into a textured mesh with spatially varying PBR materials and a coherent HDR environment in less than a second. In order to allow for **Multiview Illumination Disentanglement Reconstruction** we utilize a two-path approach achieved through the following novel contributions:

- **Cross-view Fusion** A shared cross-conditioning transformer ingests an arbitrary number of views and builds unified feature triplanes used by both paths, driving consistency across viewpoints.
- **Two-path Illumination Disentanglement.** A *geometry+appearance path* yields mesh and svBRDF (albedo/roughness/metallic/normal) from this unified triplane, while a *lighting path* fuses mask-aware tokens to predict an efficient RENI++ (Gardner et al., 2023) latent code representing a coherent HDR environment.
- **Disentangled Training via MC+MIS.** A differentiable physically-based Multiple Importance Sampling (MIS) Monte Carlo (MC) renderer ties both paths together, enforcing physically meaningful materials and illumination disentanglement.
- **Mixed-domain Training.** We train on a mixture of synthetic PBR-supervised data and real multi-view captures using image space self-supervision to bridge the gap and allow for real-world generalization.

Together, these pieces deliver the first feed-forward pipeline that jointly reconstructs geometry, spatially varying materials, and HDR illumination at interactive speed. Our experiments show improved reconstruction, relighting fidelity and material realism over recent (i) generative and (ii) reconstruction pipelines; we will release code and weights to foster adoption and reproducibility.

2 RELATED WORK

ReLi3D lies at the intersection of 3D reconstruction, inverse rendering, and appearance estimation. The most closely aligned approaches are image-to-3D reconstruction and generation methods, and we seek to clearly differentiate our feed-forward approach from optimization-based reconstruction methods.

Inverse Rendering Inverse rendering estimates shape, appearance, and environment lighting from image observations, an inherently ambiguous problem with many plausible material-lighting combinations explaining identical observations. Modern methods leverage differentiable rendering (Li et al., 2018a; Liu et al., 2019) with scene representations such as NeRF (Mildenhall et al., 2021) or Gaussian splats (Kerbl et al., 2023) to reconstruct scenes from dense RGB imagery (Zhang et al., 2021b; Boss et al., 2021; 2022; Engelhardt et al., 2024; Liang et al., 2024; Dihlmann et al., 2024). Although regularization losses in shape, materials, or environment (Barron & Malik, 2013; Li et al., 2018b; Gardner et al., 2017) help reduce ambiguity, these optimization-based approaches require dense multi-view imagery and lengthy inference times. None manages to reconstruct 3D objects from sparse views, let alone single images. In contrast, ReLi3D performs feed-forward inference from sparse views while jointly estimating spatially varying materials and HDR environments via RENI++ (Gardner et al., 2023).

Image-to-3D Generation Score Distillation Sampling methods (Poole et al., 2023; Shi et al., 2023; Wang et al., 2024b) optimize 3D representations using 2D diffusion priors but suffer from artifacts and impractically slow inference. Multi-view generation approaches (Liu et al., 2023; Long et al., 2024; Voleti et al., 2024; Tang et al., 2024) first generate consistent views and then apply reconstruction, but face view inconsistencies and inherit inverse rendering ambiguities.

Direct 3D diffusion methods model object distributions in triplane (Shue et al., 2023; Cheng et al., 2023; Yariv et al., 2024) or compressed latent spaces (Zhao et al., 2025; Xiang et al., 2024). SPAR3D (Huang et al., 2025) uniquely diffuses both geometry and PBR materials by first generating sparse point clouds and then regressing detailed structure and appearance, but requires expensive probabilistic sampling. The lack of large-scale PBR data typically precludes joint geometry-material modeling in diffusion frameworks. Our feed-forward approach achieves comparable quality without the computational overhead of generative sampling, enabling end-to-end joint structure and appearance prediction.

Image-to-3D Reconstruction Early regression approaches (Choy et al., 2016; Wang et al., 2018; Mescheder et al., 2019) were limited by small datasets like ShapeNet (Chang et al., 2015), restricting generalization. Large Reconstruction Models (LRMs) (Hong et al., 2023; Tochilkin et al., 2024b; Boss et al., 2024) now perform direct feed-forward inference at scale using transformer architectures and large datasets (Deitke et al., 2022; Reizenstein et al., 2021).

Although fast and practical, existing methods such as SF3D (Boss et al., 2024) predict only single roughness/metallic values per object rather than spatially varying materials, and lack environment estimation. Most critically, these approaches optimize for single-view reconstruction, leaving material-lighting disentanglement fundamentally ill-posed, and the same appearance can arise from countless material-illumination combinations.

The parallel work LIRM (Li et al., 2025) addresses similar goals through progressive optimization but lacks illumination prediction and relies purely on synthetic supervision, limiting real-world applicability. ReLi3D uniquely leverages multi-view constraints as the primary mechanism for material-lighting disentanglement, enabling robust spatially varying PBR reconstruction with environment estimation through mixed-domain training that bridges synthetic and real-world data.

3 Preliminaries

Reconstructing 3D objects with realistic materials and lighting from images requires understanding how light interacts with surfaces and how to efficiently represent 3D information. This section introduces the key concepts underlying our approach: physically based material representations, environment illumination modeling, and neural 3D representations that enable feed-forward reconstruction.

3.1 Physically Based Material Representation

An object's visual appearance results from how its surface reflects and refracts light, formally described by the bidirectional reflectance distribution function (BRDF) $f_r(\omega_{\text{in}}, \omega_{\text{out}})$. This function models the fraction of light reflected into direction ω_{out} given incoming light from direction ω_{in}. When material properties vary across the surface, we have a spatially varying BRDF (svBRDF).

In practice, we parameterize materials using Disney's principled BRDF (Burley & Studios, 2012) with metallic-roughness representation: RGB albedo (base color) ρ, scalar roughness r (controlling surface smoothness), and scalar metallic parameter m. Additionally, normal bump maps encode high-frequency surface perturbations for fine geometric detail. For reconstruction scenarios without predefined UV mappings, we define the local tangent space with the surface normal as up-direction and align the tangent with the world coordinate system (Vainer et al., 2024).

3.2 Environment Illumination

Realistic rendering requires modeling the incoming illumination from all directions, typically represented as an environment map $L_{\text{env}}(\omega)$ that depends only on direction ω. Traditional representations using spherical harmonics or spherical Gaussians are limited in capturing high-frequency lighting

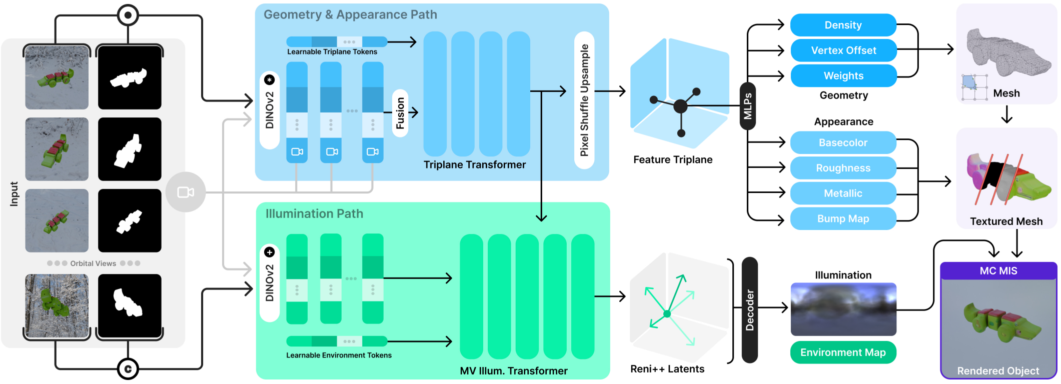

Figure 2: **ReLi3D Overview.** Multi-view input images are fused by a shared cross-conditioning transformer into two parallel paths: a Geometry & Appearance Path (blue) using a Triplane Transformer to predict mesh geometry and PBR materials, and an Illumination Path (green) using a Multi-View Illumination Transformer to estimate HDR environments. Both paths are unified through a differentiable Monte Carlo Multiple Importance Sampling rendering to learn to wproduce complete relightable 3D assets.

details like sharp shadows or bright light sources. RENI++ (Gardner et al., 2023) provides a more condensed expressive representation by learning a compact latent space for realistic illumination patterns. Environment maps are decoded from latent codes $\mathbf{z} \in \mathbb{R}^{49\times3}$ as:

$$L_{\text{env}}(\omega) = \exp(f_\theta(\mathbf{z}, \gamma(\omega))) \tag{1}$$

where f_θ is the pre-trained decoder and $\gamma(\omega)$ provides positional encoding. This enables a low dimensional representation perfectly suited for fast feed-forward reconstruction.

3.3 Large Reconstruction Models and Triplane Representations

Recent advances in feed-forward 3D reconstruction leverage large transformer models trained on extensive 3D datasets. Methods like LRM (Hong et al., 2023) and TripoSR (Tochilkin et al., 2024a) demonstrate that direct image-to-3D reconstruction is feasible without per-object optimization.

These approaches typically use triplane representations to efficiently encode 3D information. A triplane $\mathbf{T} \in \mathbb{R}^{3\times C\times H\times W}$ consists of three orthogonal 2D feature planes. For any 3D point $\mathbf{p} = (x, y, z)$, features are extracted by projecting onto each plane:

$$\mathbf{f}(\mathbf{p}) = \text{concat}(\mathbf{T}_{xy}(x, y), \mathbf{T}_{yz}(y, z), \mathbf{T}_{zx}(z, x)) \tag{2}$$

These concatenated features are then decoded through MLPs to predict geometric and appearance properties. SF3D (Boss et al., 2024) exemplifies this paradigm, it encodes input images with DINOv2 (Oquab et al., 2023), processes them through a transformer with camera conditioning, and outputs triplane features. These are decoded into geometry via DMTet (Shen et al., 2021) and textured using fast UV unwrapping. However, SF3D is limited to single-view input, global material properties, and lacks environment estimation. Limitations our approach addresses through multi-view fusion and spatially varying material prediction.

4 Method

Our core insight is that multi-view constraints provide the missing information to disentangle material properties from lighting effects, a problem that remains fundamentally ill-posed for single-view methods. We achieve this through a unified two-path architecture that jointly predicts object structure with spatially varying materials and environment illumination from arbitrary numbers of input views. Figure 2 illustrates our complete pipeline.

4.1 Multi-view Illumination Disentanglement Architecture

Our approach centers on a novel two-path prediction strategy enabled by multi-view fusion. The **geometry+appearance path** predicts mesh structure and spatially varying BRDF parameters from unified triplane features, while the **illumination path** estimates HDR environment maps via our multi-view RENI++ extension. Both paths are driven by a shared cross-conditioning transformer that fuses arbitrary numbers of input views, creating consistent feature representations that enable robust material-lighting disentanglement.

4.1.1 Cross-view Feature Fusion

Let the input be a set of N masked images with cameras $\{(\mathbf{I}_i, \mathbf{M}_i, \mathbf{C}_i)\}_{i=1}^{N}$. We first form per-view tokens with DINOv2 and camera modulation:

$$\mathbf{T}_i^{\text{img}} = \text{DINOv2}(\mathbf{I}_i \odot \mathbf{M}_i), \quad \mathbf{e}_i = f_{\text{cam}}(\mathbf{C}_i), \quad \mathbf{T}_i^{\text{cond}} = \big[\,\mathbf{T}_i^{\text{img}} \odot \mathbf{e}_i \;;\; \mathbf{e}_i\,\big]. \tag{3}$$

We designate one view as the hero view h and its tokens are concatenated to the learned triplane token bank $\mathbf{T}^{\text{tri}}$ and drive the query stream of the transformer:

$$\mathbf{Q}_0 = \big[\,\mathbf{T}^{\text{tri}} \;;\; \mathbf{T}_h^{\text{img}}\,\big]. \tag{4}$$

The hero view serves as the query stream for cross-conditioning and is selected uniformly at random during training and evaluation, ensuring robust performance independent of viewpoint choice.

To make cross-view context compact yet expressive, we employ latent mixing. A bank of learnable latent tokens $\mathbf{L}_0 \in \mathbb{R}^{L \times D}$ is mixed with the projected cross-view tokens (all non-hero views) to form a memory $\mathbf{M}$ that the query stream will attend to:

$$\mathbf{H}_i = P_\ell\big(\text{LayerNorm}(\mathbf{T}_i^{\text{cond}})\big),\; i \in \mathcal{V}_{\text{cross}}, \tag{5}$$

$$\mathbf{L}_1 = \text{SelfAttn}(\text{LayerNorm}(\mathbf{L}_0)) \tag{6}$$

$$\mathbf{M} = \text{Interleave}\big(\mathbf{L}_1,\; \text{TokenConcat}(\{\mathbf{H}_i\}_{i \in \mathcal{V}_{\text{cross}}})\big). \tag{7}$$

Here P_ℓ projects tokens to the latent dimensionality D, and Interleave denotes the two-stream interleaved transformer, which alternates blocks that (i) update $\mathbf{Q}$ with cross-attention to $\mathbf{M}$ and (ii) refine $\mathbf{M}$ via self-/cross-attention. The main transformer thus computes:

$$\mathbf{T}^{\text{out}} = \text{TwoStream}(\mathbf{Q}_0,\, \mathbf{M})\,, \tag{8}$$

which yields triplane-conditioned features that are consistent across an arbitrary number of input views while preserving a dedicated hero view pathway for stable geometry/appearance alignment. In implementation, we use pixel-shuffle upsampling to obtain higher-resolution triplanes from raw predictions.

4.1.2 Spatially Varying Material Prediction

Our **geometry+appearance path** operates on the unified triplane representation to predict spatially varying material properties and mesh structure. The transformer output tokens $\mathbf{T}^{\text{out}}$ are directly interpreted as triplane pixels, forming our unified 3D representation $\mathbf{T} \in \mathbb{R}^{3 \times 40 \times 384 \times 384}$. For any 3D point $\mathbf{p}$, we extract features via triplane projection as established in Equation (2).

Crucially, we predict all material and geometric properties from this single shared triplane embedding using task-specific MLP heads:

$$\{\sigma, \rho, r, m, \mathbf{n}_{\text{bump}}\}(\mathbf{p}) = \{\text{MLP}_{\text{density}}, \text{MLP}_{\text{albedo}}, \text{MLP}_{\text{rough}}, \text{MLP}_{\text{metal}}, \text{MLP}_{\text{normal}}\}(\mathbf{f}(\mathbf{p})) \tag{9}$$

where σ is density, ρ is albedo, r is roughness, m is metallic, and $\mathbf{n}_{\text{bump}}$ represents normal perturbations. This unified approach eliminates the need for separate material tokens and enables complex multi-material object support.

Geometry is extracted using Flexicubes (Shen et al., 2023) for superior mesh quality, and the resulting mesh is textured with spatially varying PBR parameters via fast UV unwrapping.

4.1.3 Multi-view Environment Estimation

We introduce a novel multi-view illumination inference approach that fundamentally differs from existing methods. While prior work typically predicts environment maps using simple MLPs from triplane features or single-view observations, we present the first method to leverage multi-view reasoning with adaptive background masking for robust environment estimation.

Our **illumination path** operates in parallel to the geometry reconstruction, enabling dual-mode operation where our method can robustly recover HDR environments from either direct background observations or indirect material reflectance cues across multiple viewpoints. We utilize RENI++ as an efficient illumination representation, however this approach could be easily extended to other lighting representations.

We encode mask–image pairs $(\mathbf{M}_i,\, \mathbf{I}_i)$ via a trainable DINOv2-small with two extra input channels to obtain mask-aware tokens

$$\mathbf{T}_i^{\text{mask}} = f_{\text{mask}}\big([\mathbf{M}_i,\, \mathbf{I}_i]\big), \quad i = 1 \dots N. \tag{10}$$

These tokens are concatenated with the object-transformer outputs to form the environment context

$$\mathbf{T}^{\text{env-ctx}} = \text{concat}\big(\{\mathbf{T}_i^{\text{mask}}\}_{i=1}^{N},\, \mathbf{T}^{\text{out}}\big). \tag{11}$$

A dedicated 1D transformer maps learned environment tokens to a RENI++ latent *and* a global rotation (6D) via cross-attention:

$$[\, \mathbf{z}_{\text{env}},\, \mathbf{r}_{\text{6D}}\,] = \text{EnvTransformer}\big(\mathbf{T}^{\text{env-bank}},\, \mathbf{T}^{\text{env-ctx}}\big), \quad \mathbf{z}_{\text{env}} \in \mathbb{R}^{K\times d},\; \mathbf{r}_{\text{6D}} \in \mathbb{R}^{6}, \tag{12}$$

where $K \times d$ matches the RENI++ latent grid dimensionality. The final HDR environment is decoded as established in Equation (1).

Critically, our training employs stochastic background masking, randomly occluding background pixels in a subset of views during training. This forces the network to solve two complementary tasks: when background pixels are visible, it can read lighting directly from the environment; when they are masked, it must infer lighting from indirect cues in object reflections and shading. This dual mode training enables robust illumination inference in real-world scenes where backgrounds are often partially cropped, saturated, or noisy.

4.2 Disentangled Training via MC+MIS

Our differentiable physically based Monte Carlo (MC) renderer with Multiple Importance Sampling (MIS) ties both reconstruction paths together, enforcing physically meaningful material-illumination disentanglement while enabling mixed-domain training. We found that utilizing VNDF sampling (Heitz, 2018) with spherical caps (Dupuy & Benyoub, 2023) and antithetic sampling (Zhang et al., 2021a) helps stabilize the training. This MC+MIS approach enables the following capabilities:

- **Physical disentanglement**: The renderer enforces that predicted materials f_r and illumination L_{env} must jointly explain observed images through physically based light transport.
- **Mixed supervision**: When PBR ground truth exists, we additionally use direct material supervision; otherwise, the renderer ensures material and lighting consistency purely through image reconstruction.
- **Domain bridging**: This allows seamless training across synthetic PBR data, synthetic RGB-only renders, and most importantly real-world captures, dramatically improving generalization and robustness.

The result is the first system capable of learning spatially varying material reconstruction from mixed-domain data without supervision collapse, enabling robust performance on real-world inputs while maintaining physical plausibility.

5 Experiments

We evaluate ReLi3D across three core dimensions that validate our central thesis: multi-view constraints enable superior material and lighting disentanglement for fast, production ready 3D asset

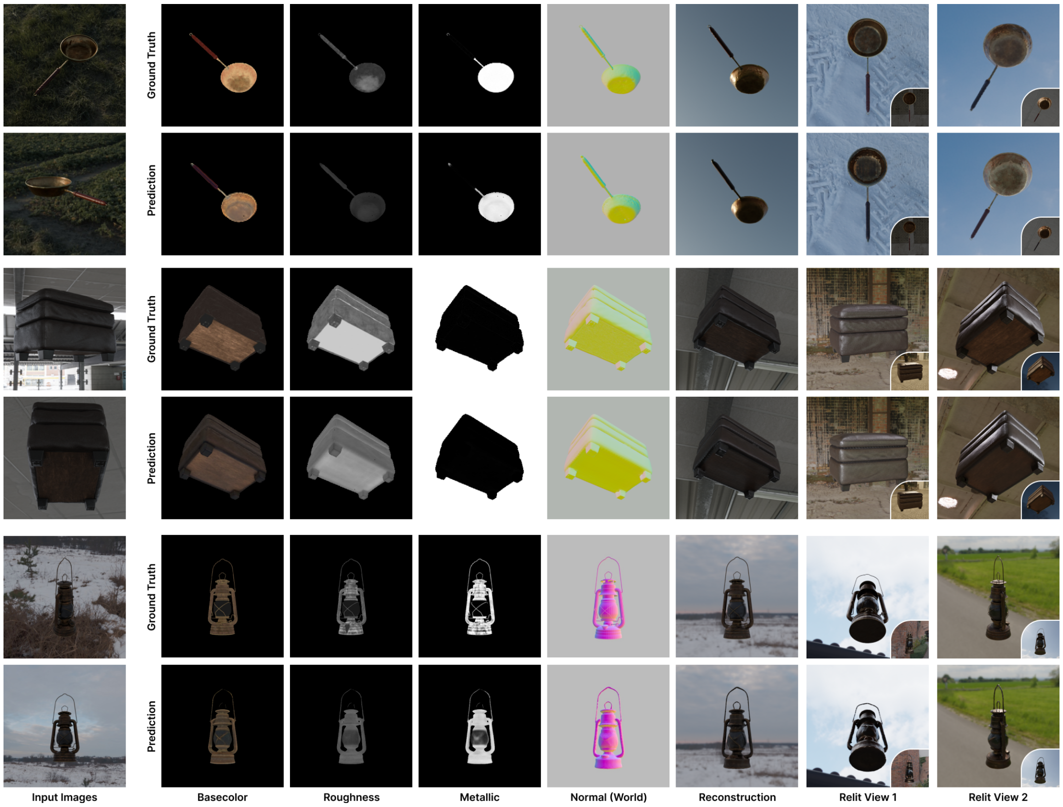

Figure 3: **PBR & Relighting Results.** We show that our spatially varying PBR prediction is faithful to the ground truth and therefore produces highly detailed and realistic relightings.

creation. Our experiments demonstrate that while we achieve competitive geometry reconstruction at interactive speeds, our primary contribution lies in illumination disentanglement, delivering spatially varying PBR materials and coherent HDR environments that enable high-fidelity relighting.

5.1 Implementation and Evaluation Setup

We train on 174k objects total: 42k synthetic PBR (full material supervision), 70k synthetic RGB-only, and 62k real-world captures from UCO3D (Liu et al., 2024). For evaluation, we test on out-of-distribution datasets including Google Scanned Objects (GSO) (Downs et al., 2022), Polyhaven (Haven, 2024) objects rendered with HDRI-Skies (IHDRI, 2024), Stanford ORB (Kuang et al., 2024), and challenging real-world UCO3D captures with motion blur and imperfect masks. We compare against recent feed-forward and generative methods: SF3D (Boss et al., 2024), SPAR3D (Huang et al., 2025), 3DTopia-XL (Chen et al., 2024), and Hunyuan3D (Zhao et al., 2025). All experiments run on a single H100 GPU, including mesh extraction and texture baking. To ensure fair comparison, we apply rigid ICP alignment to ground truth meshes before evaluating image metrics, as baselines often produce meshes in arbitrary canonical spaces. ReLi3D predictions are naturally aligned, highlighting a useful feature for practical applications. For more details, please refer to the appendix Appendix B.

5.2 Material-Lighting Disentanglement: Our Core Contribution

While overall 3D reconstruction is important, we are particularly interested in the quality of material estimation and illumination disentanglement.

Spatially Varying Material Prediction. For PBR results in Figure 3 and Table 1, we demonstrate that ReLi3D predicts fully spatially varying PBR materials that improve significantly with additional views (e.g., where the base of the bed is corrected in Figure 5). Our method ranks first across all material metrics: albedo reconstruction achieves 25.00 dB PSNR (vs SF3D's 18.42 dB), roughness

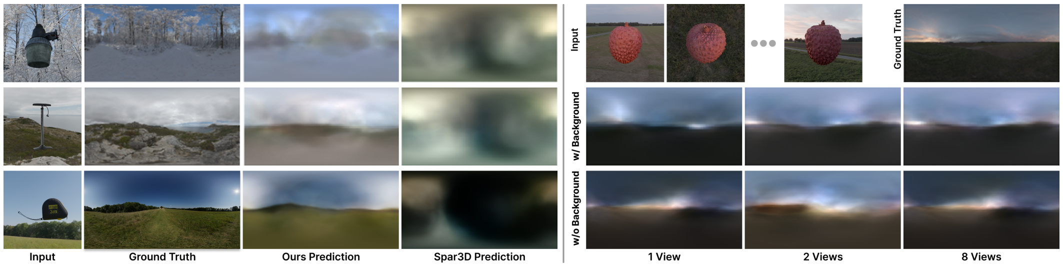

Figure 4: **Illumination Comparison.** (Left) Single view, illumination prediction results compared to ground truth and SPAR3D, which also predicts RENI++ latents, indicating our severely improved method. (Right) Influence of increasing numbers of views and background information. Notice how well we can predict the illumination in the top rows with background information locate light sources correctly, whereas the bottom row is more spread out as it is inferred from diffuse surface reflections only.

reaches 22.69 dB PSNR, and metallic prediction achieves 32.73 dB. Multi-view input further enhances these results, demonstrating that cross-view constraints successfully resolve material-lighting ambiguities.

Relighting Performance. The ultimate test of material-lighting disentanglement is relighting under novel environments. For quantitative relighting evaluation, we rendered each reconstruction in a novel out-of-distribution HDR environment. Even when competing methods receive ground-truth environment maps as input, ReLi3D ranks first across all relighting metrics in Table 1. Visually, Figure 3 shows that our material estimation is so accurate that the relit reconstructions closely resemble the ground truth, confirming that our material decomposition generalizes well to novel lighting conditions.

Environment Estimation. Figure 4 compares our predicted HDR environment maps with ground truth. Even a single view suffices to recover the correct sky color and sun direction. We also show how background information helps recover correct light sources, and utilizing multiple views helps recover correct light directions, even in dark environments. In contrast, SPAR3D often predicts over-smoothed, low-contrast maps with no clear light sources.

		Polyhaven + Blender Shinny														
		Relighting			Image			Basecolor			Roughness			Metallic		
Method	**Time (s)**	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LSSIMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
SF3D	**0.26**	15.79	0.843	0.150	18.03	0.875	0.120	18.42	0.831	0.220	19.60	0.876	0.127	28.37	0.888	0.116
SPAR3D	0.36	15.23	0.836	0.154	17.02	0.862	0.132	17.70	0.822	0.251	19.53	0.874	0.121	30.52	0.895	0.088
3DTopia-XL	31.38	14.20	0.869	0.140	14.60	0.853	0.168	19.52	0.818	0.330	15.16	0.847	0.191	27.60	0.861	0.071
Hunyuan3D	69.40	14.81	0.845	0.151	17.41	0.875	0.118	21.25	0.837	0.265	—	—	—	—	—	—
ReLi3D (Ours)	0.28	**19.77**	**0.906**	**0.088**	**20.09**	**0.897**	**0.094**	**25.00**	**0.866**	**0.151**	**22.69**	**0.893**	**0.085**	**32.73**	**0.913**	**0.050**
Hunyuan3D (2 Views)	41.25	14.94	0.846	0.148	17.33	0.875	0.115	21.29	0.837	0.271	—	—	—	—	—	—
Hunyuan3D (4 Views)	43.06	14.89	0.845	0.149	17.29	0.876	0.116	21.34	0.838	0.270	—	—	—	—	—	—
ReLi3D (Ours) (2 Views)	**0.28**	20.40	0.909	0.082	21.11	0.905	0.082	25.90	0.874	0.120	23.75	0.901	0.075	33.06	0.917	0.046
ReLi3D (Ours) (4 Views)	0.29	20.94	0.912	0.078	21.48	0.909	0.078	26.45	0.878	0.112	24.08	0.904	0.072	33.18	0.918	0.045
ReLi3D (Ours) (8 Views)	0.31	21.17	0.913	0.076	21.63	0.910	0.076	26.65	0.880	0.111	24.30	0.906	0.071	**33.30**	**0.919**	**0.044**
ReLi3D (Ours) (16 Views)	0.32	**21.21**	**0.914**	**0.075**	**21.73**	**0.911**	**0.075**	**26.78**	**0.881**	**0.109**	**24.50**	**0.907**	**0.069**	33.21	**0.919**	**0.044**

Table 1: **Relighting & Image & PBR Metrics Comparison.** (Left) Relighting performance. (Middle) Image reconstruction performance. (Right) PBR material reconstruction performance. While most methods produce only global PBR parameters, ours produce spatially varying material maps which increase in quality with more views.

5.3 Overall Reconstruction Quality

While geometry reconstruction is not our primary focus, ReLi3D achieves competitive results at unprecedented speed. Our model achieves quantitative and qualitative state-of-the-art single-view reconstruction results on out of distribution synthetic (GSO, Stanford ORB) and real-world (UCO3D) data in Table 2. In the multi-view setting, ReLi3D permorms well on geometric and outperforms on all image metrics while running in avg. 0.31s. Supplying just four views improves CD by 27% and pushes the F-score@0.5 to 0.993, showcasing the effectiveness of our multi-view cross-conditioning at virtually unchanged cost. Performance saturation beyond 4–8 views stems from coverage saturation: once surface coverage is sufficient, additional random views often provide redundant information rather than new constraints, leading to marginal gains.

Figure 6 offers an end-to-end comparison across all datasets and methods. Competing techniques frequently fail or output planar artifacts, while our multi-view fusion reconstructs complete assets, including hidden backsides, with better ground truth lighting and shadowing. For real-world captures, ReLi3D remains robust, and our method improves with multi-view input while others do not (e.g., the face of the teddy bear in Figure 6).

We acknowledge that specialized high-resolution diffusion methods may achieve superior geometric detail through longer optimization. However, our contribution lies in the speed-quality trade-off for material-aware reconstruction: we deliver complete, relightable assets in under a second while running 100× faster than generative approaches like Hunyuan3D.

		Gso + Standford Orb							Uco3D						
		3D				Image			3D				Image		
Method	**Time (s)**	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓
SF3D	**0.28**	0.132	0.543	0.810	0.974	17.64	0.856	0.131	0.248	0.297	0.564	0.867	12.79	0.748	0.288
SPAR3D	0.39	0.152	0.507	0.766	0.959	16.34	0.837	0.151	0.232	0.368	0.634	0.871	12.39	0.723	0.285
TripoSG	8.54	0.232	0.357	0.619	0.881	14.47	0.832	0.211	0.274	0.297	0.520	0.842	11.85	0.752	0.330
3DTopia-XL	45.03	0.217	0.341	0.636	0.907	14.40	0.831	0.183	0.250	0.262	0.512	0.888	12.00	0.727	0.304
Trellis	69.09	0.149	0.551	0.780	0.958	16.56	0.871	0.132	**0.182**	**0.433**	**0.705**	**0.936**	13.27	0.760	0.309
Hunyuan3D	39.69	0.133	**0.557**	**0.819**	0.970	16.68	0.851	0.139	0.214	0.356	0.610	0.913	13.75	0.752	0.273
ReLi3D (Ours)	0.30	**0.105**	0.322	0.671	**0.985**	**19.57**	**0.902**	**0.103**	0.209	0.243	0.309	0.935	**15.28**	**0.839**	**0.214**
Hunyuan3D (2 Views)	43.94	0.114	0.604	0.869	0.986	17.36	0.855	0.132	0.219	0.329	0.583	0.923	13.54	0.747	0.275
Hunyuan3D (4 Views)	48.06	0.110	0.636	0.875	0.986	17.40	0.856	0.130	0.222	0.341	0.600	0.904	13.58	0.749	0.277
ReLi3D (Ours) (2 Views)	0.31	0.088	0.752	0.914	0.991	20.72	0.885	0.090	0.190	0.343	0.611	0.952	15.45	**0.841**	0.217
ReLi3D (Ours) (4 Views)	**0.28**	0.081	0.787	0.926	0.993	21.43	0.894	0.080	0.188	0.346	0.622	0.953	15.60	0.839	0.212
ReLi3D (Ours) (8 Views)	0.29	**0.076**	0.815	**0.937**	**0.994**	22.14	0.899	0.072	0.186	0.355	0.625	0.954	15.48	0.838	0.219
ReLi3D (Ours) (16 Views)	0.36	**0.076**	**0.817**	0.936	0.993	**22.29**	**0.901**	**0.070**	**0.184**	**0.363**	**0.631**	**0.955**	**15.73**	0.839	**0.210**

Table 2: **3D and Image Metrics.** ReLi3D clearly achieves SOTA in single and sparse multi-view reconstruction while also achieving great speeds. It is worth noting that that TripoSG and Hunyuan3D also produce signficantly higher vertex counts (100k+ vs 4.5k for ours).

5.4 Cross-Domain Training Efficiency

Our mixed-domain training protocol enables robust real-world performance Figure 7 despite training on only 174k objects 10-50× less data than recent large-scale methods. The key insight is that multi-view constraints provide stronger supervision signals than massive single-view datasets, enabling efficient learning of material-lighting disentanglement.

We evaluate on real-world Stanford ORB dataset (Kuang et al., 2024) to demonstrate generalization (Table 3). ReLi3D outperforms all baselines across 3D reconstruction, image quality, and material prediction metrics. Multi-view input further improves performance.

	Stanford ORB								
	3D				Image			Basecolor	
Method	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
SF3D	0.152	0.512	0.769	0.954	17.75	0.891	0.111	18.52	0.865
SPAR3D	0.165	0.488	0.751	0.940	17.10	0.886	0.113	17.80	0.857
Trellis	0.152	0.561	0.782	0.948	17.13	0.888	0.112	—	—
Hunyuan3D	0.141	0.571	0.801	0.960	16.96	0.877	0.110	21.37	0.872
ReLi3D (Ours)	**0.116**	**0.608**	**0.856**	**0.980**	**18.68**	**0.907**	**0.098**	**24.21**	**0.891**
Hunyuan3D (2 Views)	0.134	0.588	0.809	0.967	16.91	0.876	0.108	21.42	0.872
Hunyuan3D (4 Views)	0.136	0.579	0.810	0.966	16.83	0.877	0.108	21.46	0.873
ReLi3D (Ours) (2 Views)	0.104	0.654	0.888	0.986	19.74	0.913	0.089	25.01	0.896
ReLi3D (Ours) (4 Views)	0.094	0.718	0.906	0.989	20.84	0.919	0.082	25.33	0.900
ReLi3D (Ours) (8 Views)	**0.089**	0.745	**0.914**	**0.991**	21.21	**0.921**	**0.080**	25.50	0.901
ReLi3D (Ours) (16 Views)	**0.089**	**0.749**	**0.914**	0.990	**21.29**	**0.921**	**0.080**	**25.58**	**0.902**

Table 3: **Real-world Evaluation on Stanford ORB.** Quantitative evaluation on Stanford ORB dataset showing 3D reconstruction, image quality, and basecolor material prediction performance. Our method outperforms baselines across all metrics and improves with more input views.

5.5 Limitations

Although rare, failure cases occur where the decomposition fails to disentangle lighting and materials, resulting in baked-in lighting affecting the material maps. This seems to occur when environment lighting is not in domain for the RENI++ prior, most notably when multiple very strong light sources are present.

The largest remaining weakness is the relatively limited resolution of the triplane, limiting texture and geometry resolution in practice, also visible in reconstruction examples against Hunyuan3D. While we do not claim to have the best geometry prediction, as other methods spend more time with high-quality diffusion processes, we are confident that our illumination disentanglement structure is a contribution that, with sufficient resources, could help larger methods.

6 Conclusion

We have enhanced the fundamental challenge of illumination disentanglement in feed-forward 3D reconstruction, enabling the first method to jointly predict spatially-varying PBR materials and coherent HDR environments from sparse image inputs. Through our novel two-path architecture and differentiable Monte Carlo training, we demonstrate that proper material-lighting separation is achievable at interactive speeds, delivering production-quality relightable assets in under one second.

This development in illumination disentanglement opens exciting avenues for future research and applications. The ability to rapidly generate physically accurate 3D assets from casual captures could transform content creation workflows, enabling real-time asset digitization. More broadly, our disentanglement framework could extend beyond reconstruction to enable in-the-wild material understanding; imagine training on objects captured under varying real-world illumination to learn material priors that generalize across lighting conditions.

We release all code, pretrained weights, and dataset generation scripts to accelerate adoption and enable the community to build upon this foundation for the next generation of 3D-aware vision systems.

References

Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17–24, 2013.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12684–12694, 2021.

Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022.

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024.

Brent Burley and Walt Disney Animation Studios. Physically-based shading at disney. In *Acm Siggraph*, volume 2012, pp. 1–7. vol. 2012, 2012.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia-xl: High-quality 3d pbr asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*, 2024.

Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4456–4465, 2023.

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pp. 628–644. Springer, 2016.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

Jan-Niklas Dihlmann, Arjun Majumdar, Andreas Engelhardt, Raphael Braun, and Hendrik P.A. Lensch. Subsurface scattering for gaussian splatting, 2024.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Jonathan Dupuy and Anis Benyoub. Sampling Visible GGX Normals with Spherical Caps. *Computer Graphics Forum*, 2023.

Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Ricardo Martin Brualla, Jonathan T Barron, Hendrik Lensch, et al. Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19636–19646, 2024.

James AD Gardner, Bernhard Egger, and William AP Smith. Reni++ a rotation-equivariant, scale-invariant, natural illumination prior. *arXiv preprint arXiv:2311.09361*, 2023.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

Poly Haven. Poly Haven • Poly Haven — polyhaven.com. `https://polyhaven.com/`, 2024. [Accessed 22-08-2024].

Eric Heitz. Sampling the ggx distribution of visible normals. *Journal of Computer Graphics Techniques (JCGT)*, 2018.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. 2025.

IHDRI. HDRI Skies - Download your favorite HDRI Sky for Free! https://www.ihdri.com/, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018a.

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 72–87, 2018b.

Zhengqin Li, Dilin Wang, Ka Chen, Zhaoyang Lv, Thu Nguyen-Phuoc, Milim Lee, Jia-Bin Huang, Lei Xiao, Cheng Zhang, Yufeng Zhu, et al. Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. *arXiv preprint arXiv:2504.20026*, 2025.

Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21644–21653, 2024.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7708–7717, 2019.

Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Uncommon objects in 3d. In *arXiv*, 2024.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *CVPR*, 2024.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20133–20143, October 2023.

Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Aditya Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. *arXiv preprint arXiv:2303.13009*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10901–10911, 2021.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Trans. Graph.*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20875–20886, 2023.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024a.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3D object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024b.

Shimon Vainer, Mark Boss, Mathias Parger, Konstantin Kutsy, Dante De Nigris, Ciara Rowles, Nicolas Perony, and Simon Donné. Collaborative control for geometry-conditioned PBR image generation. In *ECCV*, 2024.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *ECCV*, 2024.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–67, 2018.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2024a.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Lior Yariv, Omri Puny, Oran Gafni, and Yaron Lipman. Mosaic-sdf for 3d generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4630–4639, 2024.

Cheng Zhang, Zhao Dong, Michael Doggett, and Shuang Zhao. Antithetic sampling for monte carlo differentiable rendering. *ACM Trans. Graph.*, 2021a.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021b.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Appendix

This appendix provides technical details and additional experimental validation for our multi-view illumination disentanglement approach. We organize the material as follows: Appendix A presents extended experimental results including detailed PBR comparisons, real-world and synthetic reconstruction examples, and an ablation study that validate our architectural choices. Appendix B offers implementation specifics including loss formulations for mixed-domain training, the progressive training protocol that bridges volumetric and mesh-based rendering. Finally, Appendix C details our curated training data composition, covering both synthetic dataset construction with full PBR supervision and the extensive preprocessing pipeline required to integrate challenging real-world UCO3D captures for robust domain generalization.

A Further Experiments

This section extends our experimental validation of our multi-view illumination disentanglement approach, including detailed visual analysis of PBR decomposition quality, comprehensive reconstruction comparisons across synthetic and real-world datasets, and critical ablation studies that validate our architectural design choices.

A.1 Comparison

Figure 5 demonstrates the superior quality of our spatially varying material predictions compared to existing methods. Unlike previous approaches that predict global material properties or fail to achieve proper material-lighting separation, our method produces detailed albedo, roughness, and metallic maps that exhibit realistic spatial variation. Particularly noteworthy is our method's ability to handle mixed-material objects.

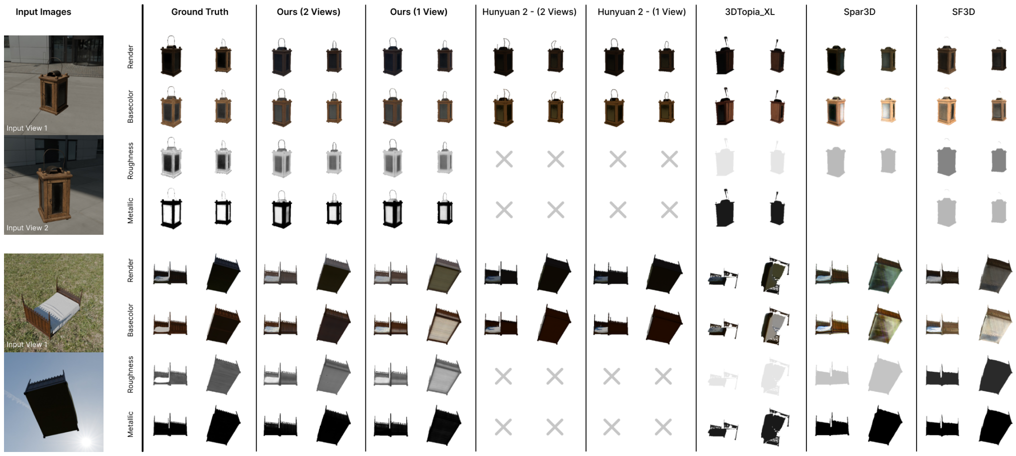

Figure 5: **PBR Decomposition Results.** Our method is capable of producing highly detailed textures and geometries even from a single view. It is also the only method capable of reproducing accurate spatially varying PBR parameters, which are essential for relighting.

Our method's generalization capabilities are extensively validated across diverse synthetic and real-world scenarios. Figure 6 showcases reconstruction quality on synthetic objects. Figure 7 provides validation on real-world captures, where imperfect masks, camera estimation errors, and challenging lighting conditions test the robustness of our approach.

Real-world Material Prediction We demonstrate real-world performance on challenging UCO3D captures with motion blur and cluttered backgrounds (Figure 8). These examples show the benefit of our multi-view setting (e.g., recovering the front of objects given additional views) and improved material prediction as lighting aligns with ground truth. Our method successfully

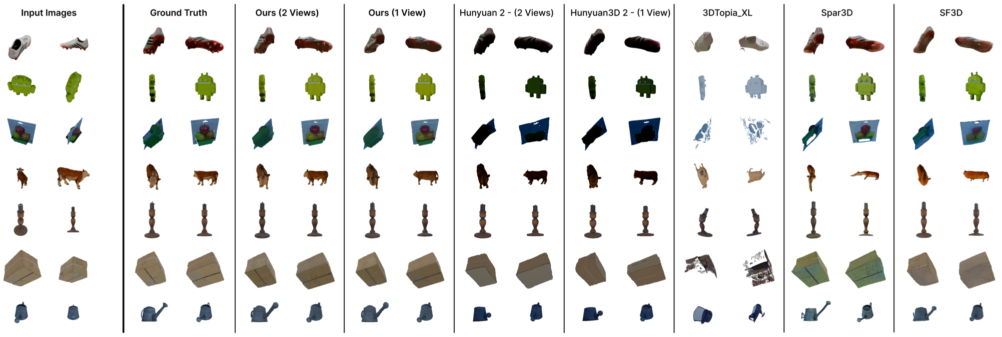

Figure 6: **Reconstruction Results (Synthetic).** Our method performs well across synthetic data and shows accurate reconstructions from a single view. Other methods show collaps with bend or flat predictions.

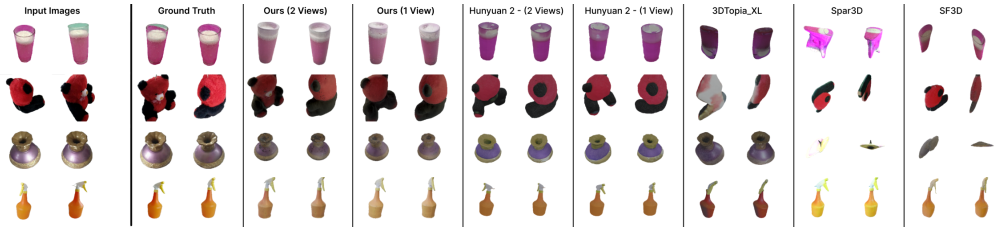

Figure 7: **Reconstruction Results (Real World).** Our method produces accurate reconstructions for real-world data, although challenging. Incorporating multiple views improves the performance further by clearing up uncertainties in unseen areas.

separates metallic and non-metallic materials even in challenging real-world settings with strong reflections and blur.

Complex Multi-material Objects We evaluate on complex, multi-material objects from the Blender Shiny dataset (Figure 9), demonstrating that our spatially varying PBR prediction generalizes to complex geometries and real materials. The figure shows predicted basecolor, roughness, metallic, and normal maps, along with relit renderings in novel environments, confirming robust material decomposition across diverse object types.

Illumination Disentanglement Quality Figure 10 provides detailed qualitative comparison of illumination prediction results between DiffusionLight, SPAR3D, and our method (ReLi3D). While DiffusionLight hallucinates completely different environments (e.g., predicting indoor scenes for outdoor inputs), and SPAR3D fails to recover meaningful illumination, ReLi3D accurately mimics the ground truth shape and color of the environment maps. This demonstrates the effectiveness of our dedicated illumination branch and multi-view reasoning for robust environment estimation.

Quantitative Evaluation of Illumination Disentanglement Table 4 provides quantitative results comparing ReLi3D, SPAR3D, and a DiffusionLight Phongthawee et al. (2023) baseline on the Polyhaven+HDRI dataset. ReLi3D achieves comparable relighting PSNR to DiffusionLight (20.88 vs 20.93 dB) while being significantly faster (0.34s vs 21.46s) and supporting multi-view input. SPAR3D achieves similar speed but significantly lower quality (17.10 dB PSNR), confirming the importance of our dedicated illumination branch architecture.

A.2 Ablation

Table 5 shows validation of our key architectural choices, with particular emphasis on the critical role of Monte Carlo rendering in achieving high-quality material-lighting disentanglement. The ablation reveals that removing the Monte Carlo renderer (- MC-Render) significantly degrades image

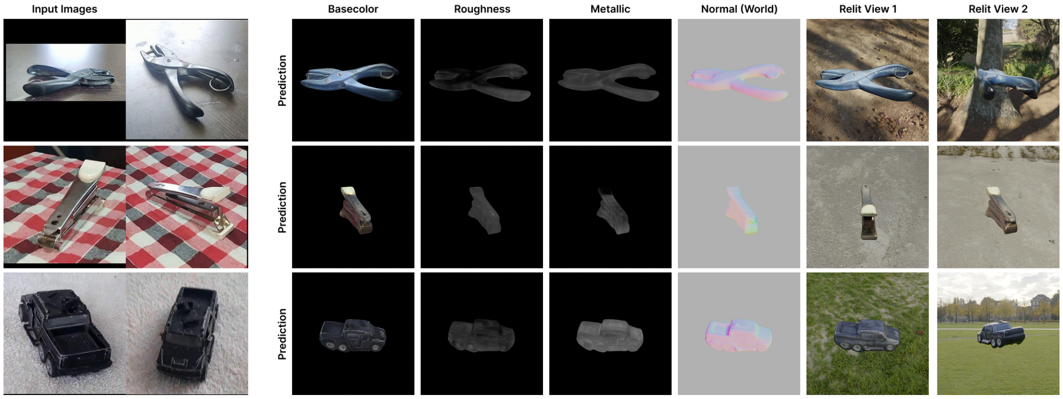

Figure 8: **Real-world material prediction.** Material maps (albedo, roughness, metallic, normal) for real-world objects from UCO3D dataset on very challenging settings, strong reflections and blur. Our method is still able to make a rough prediction and faithfully separates metallic and non-metallic materials.

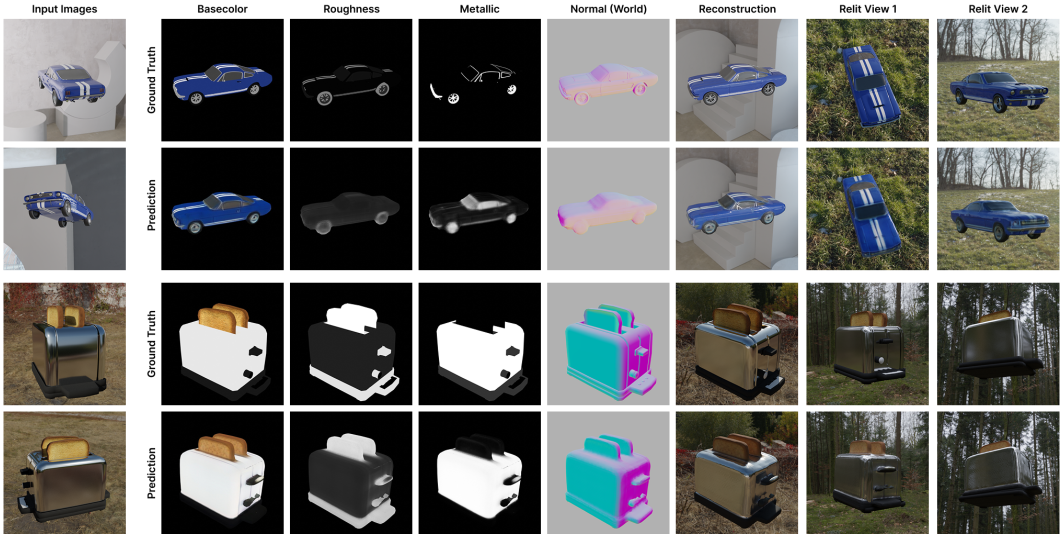

Figure 9: **Varying materials and complex objects.** Results on the Blender Shiny dataset showing spatially varying PBR material prediction on complex multi-material objects. The figure shows predicted basecolor, roughness, metallic, and normal maps, along with relit renderings in novel environments.

reconstruction quality (19.92 → 17.54 dB PSNR). This finding underscores a crucial insight: the Monte Carlo renderer with Multiple Importance Sampling is not merely an optimization detail but a fundamental component that enables proper physical disentanglement.

Training Stage Contributions Our progressive training pipeline transitions from volumetric rendering through spherical Gaussian approximation stages (128 → 256 → 512 Gaussians) to full Monte Carlo integration, as detailed in Appendix B.2. Table 6 reports the share of the total improvement (Phase 1 → Full MC) contributed by each intermediate stage. The Gaussian stages with larger batch sizes explain 70–80% of the 3D coverage gains (CD/FS), confirming they mainly stabilize geometry before expensive rendering. The Monte Carlo stage accounts for the majority of remaining improvements in material disentanglement (basecolor, roughness, metallic). The 512-Gaussian stage provides the sweet spot for geometry+runtime, while the final MC finetuning sharpens material maps and relighting without regressing 3D accuracy.

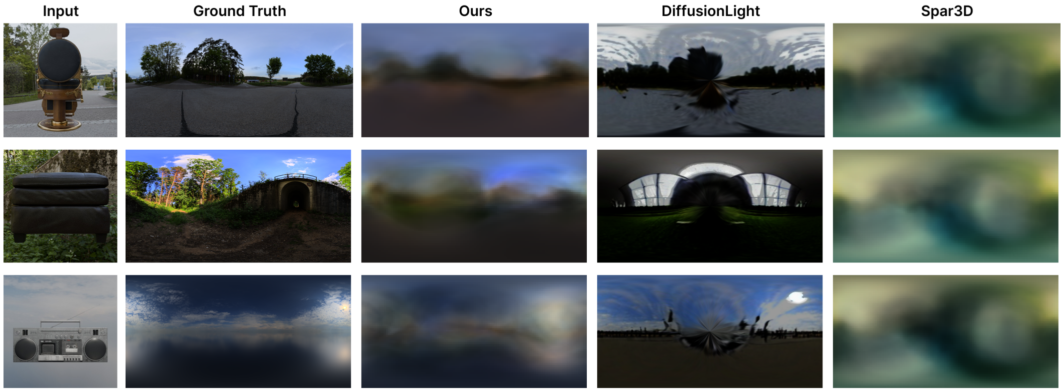

Figure 10: **Illumination Comparison.** Comparison of illumination prediction results between DiffusionLight, SPAR3D, and our method (ReLi3D). Predicted environmens vary vastly while ours mimics the ground truth shape and color, DiffusionLight hallucinates a completely different environment, SPAR3D fails.

Time (s)↓			PSNR↑		
Diffusion Light	ReLi3D	SPAR3D	Diffusion Light	ReLi3D	SPAR3D
21.46	0.34	0.33	20.93	20.88	17.10

Table 4: **Quantitative evaluation of illumination disentanglement.** Comparison of environment map prediction and relighting quality on Polyhaven+HDRI dataset.

	3D				Image
Method	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑
ReLi3D	**0.110**	**0.676**	**0.870**	**0.975**	**19.92**
- MC-Render	0.114	0.668	0.865	0.971	17.54

Table 5: **Ablation study.** Impact of removing the differentiable Monte-Carlo renderer (- MC-Render).

B Implementation Details

This section provides comprehensive implementation details for our multi-view illumination disentanglement architecture, including loss formulations, training protocols, architectural design choices.

B.1 Loss Functions

Our training objective combines physically-based image reconstruction with material and illumination supervision, designed to handle mixed-domain datasets with varying levels of ground truth availability.

Image Reconstruction Loss The primary training signal compares rendered reconstructions against ground truth images not used as input:

$$\mathcal{L}_{\text{img}} = 10.0\mathcal{L}_{\text{MSE,im}} + 2.0\mathcal{L}_{\text{LPIPS,im}} \tag{13}$$

This combination ensures both pixel-level accuracy and perceptual quality.

Geometry and Mask Supervision During volumetric training stages, we employ mask binary cross-entropy loss $10.0\mathcal{L}_{\text{mask}}$ for foreground segmentation. Geometry losses $\mathcal{L}_{\text{geom}}$ follow the Flexicubes implementation and weighting scheme for robust mesh extraction.

Method	3D Coverage Share (%)	Image Quality Share (%)	Basecolor Share (%)	Roughness Share (%)	Metallic Share (%)
Phase 2 (256) vs Phase 1 (128)	20.2	6.8	22.6	6.3	7.7
Phase 3 (512) vs Phase 2 (256)	**70.3**	**50.0**	23.9	31.3	41.0
MC (Full) vs Phase 3 (512)	9.5	43.2	**53.5**	**62.4**	**51.3**

Table 6: **Training stage contribution analysis.** Average share of the total improvement from Phase 1 (128 Gaussians) to the full Monte Carlo stage that is attributable to each intermediate stage. Columns aggregate the metrics shown in Table 1: (1) 3D coverage (CD and FS@0.05–0.5), (2) image quality (PSNR, SSIM, LPIPS), (3) basecolor (PSNR, SSIM, LSSIMSE), (4) roughness (PSNR, SSIM, RMSE), and (5) metallic (PSNR, SSIM, RMSE). Early Gaussian stages mainly expand 3D coverage, while the Monte Carlo refinement sharpens PBR material disentanglement.

Material Property Supervision Given the mixed nature of our training data, material supervision adapts to ground truth availability:

$$\mathcal{L}_{\text{mat}} = 10.0\mathcal{L}_{\text{MSE,PBR}} + 4.0\mathcal{L}_{\text{cos,nrm}} + 0.05\mathcal{L}_{\text{flat}} \tag{14}$$

where basecolor, roughness, and metallic parameters use MSE supervision when available, surface normals employ cosine similarity loss, and bump maps are regularized toward flatness using local normal direction $\mathbf{n}_{\text{up}} = (0, 0, 1)^T$.

Environment Supervision Direct RENI++ latent supervision provides illumination guidance:

$$\mathcal{L}_{\text{env}} = 0.1\mathcal{L}_{\text{MSE,RENI}} + 0.02\mathcal{L}_{\text{demod}} \tag{15}$$

When RENI++ ground truth is unavailable, demodulation regularization biases the environment toward neutral white lighting.

B.2 Training Protocol

Our multi-stage training protocol progressively transitions from volumetric to mesh-based rendering, culminating in full Monte Carlo integration.

Multi-stage Rendering Pipeline We execute three distinct training phases:

1. **Volumetric rendering** of the implicit field using NeRFAcc for initial shape learning
2. **Mesh rendering with spherical Gaussian approximation**, progressively increasing image resolution (128 → 256 → 512) for efficient lighting approximation
3. **Full Monte Carlo integration** with VNDF sampling, spherical caps, and antithetic sampling for physically accurate shading

Each stage spans 60,000 training steps. This progressive approach ensures stable convergence while gradually increasing rendering fidelity.

Stage-specific Losses and Training All stages employ the same loss formulation combining image reconstruction ($\mathcal{L}_{\text{img}}$), material supervision ($\mathcal{L}_{\text{mat}}$), geometry regularization ($\mathcal{L}_{\text{geom}}$), and environment supervision ($\mathcal{L}_{\text{env}}$) as detailed in Appendix B.1. Stages 1-3 use spherical Gaussian approximation for lighting, while stage 4 employs full Monte Carlo integration. All network components remain trainable throughout all stages no modules are frozen. The background module is excluded from weight loading when transitioning between stages to allow adaptation to new rendering configurations.

Training Configuration We utilize 512×512 input resolution and randomly sample 1–4 conditioning views per training iteration. The entire pipeline trains end-to-end with a learning rate of 5×10^{-5}. Batch sizes adapt to computational demands: 64 during volumetric rendering, 192 during spherical Gaussian stages, and 32 during Monte Carlo integration.

B.3 Architectural Design Choices

Hero View Selection and Sensitivity The hero view serves as the query stream for the cross-conditioning transformer, providing a stable reference for geometry and appearance alignment. In our reported metrics (Tables 1 and 2), the hero view is selected uniformly at random, ensuring our results reflect robust performance independent of viewpoint choice, unlike methods relying on canonical frontal views. To test sensitivity, we compared random selection against fixed frontal-view selection (Table 7). Results show only marginal differences, with slight perceptual gains for random views likely due to parallax information from side views.

	Polyhaven Subset						
	3D				Image		
Method	CD↓	FS@0.1↑	FS@0.2↑	FS@0.5↑	PSNR↑	SSIM↑	LPIPS↓
ReLi3D (Random Hero)	0.102	0.697	0.883	0.982	20.25	0.919	0.083
ReLi3D (Frontal Hero)	0.123	0.641	0.840	0.965	19.06	0.909	0.095

Table 7: **Hero view selection sensitivity.** Comparison of metrics using random hero view selection versus always selecting the most frontal view on the Polyhaven dataset.

Illumination Prior and Alternative Representations Our framework is compatible with alternative lighting representations: we use spherical Gaussian approximations in the intermediate training stage (Appendix B.2) before switching to Monte Carlo rendering with RENI++ envmaps. In those stages, we train with a low frequency Gaussian representation and observe that it fails to capture sharp highlights and directional suns, leading to worse relighting metrics as shown in Table 3.

RENI++ provides a compact but high-frequency representation critical for photorealistic relighting and accurate material and lighting separation. While nothing in our architecture prevents using SH or Gaussians, we found RENI++ to be the best trade-off between expressiveness and efficiency. We choose this compact representation to fit our memory limitations; expanding into a more memory intensive representation (e.g., ENV Map HDR prediction) would not be possible with our constraints.

C Datasets

Our training leverages a carefully curated mix of synthetic and real-world data to achieve robust generalization while maintaining physical plausibility. This mixed-domain approach enables learning from both controlled synthetic environments with full material supervision and challenging real-world captures that provide crucial domain adaptation.

C.1 Synthetic Data Composition

Following established protocols while extending coverage, we combine multiple synthetic datasets to maximize training diversity. Our synthetic corpus extends the TripoSR dataset protocol with Amazon Berkeley Objects (ABO) Collins et al. (2022) and ARIA Pan et al. (2023), providing comprehensive material and geometric variation.

Rendering Protocol Each object is rendered under three distinct illumination environments, randomly rotated around the vertical axis to prevent lighting bias. Camera focal lengths are sampled from a scaled normal distribution between 22° and 37° to match real-world capture conditions. Objects are normalized to unit scale and centered, with cameras positioned to fill the frame with appropriate padding, followed by slight positional augmentation.

We render significantly more views for objects with PBR ground truth (100 images) compared to RGB-only objects (30 images), providing richer supervision where material information is available. This asymmetric sampling strategy maximizes learning efficiency while accommodating varying supervision levels.

Illumination Environments Our synthetic rendering employs 1000 HDRI environments sourced from iHDRI IHDRI (2024) and Polyhaven Haven (2024) datasets. These environments are preprocessed to extract RENI++ latent codes, enabling direct illumination supervision during training. This diverse illumination set ensures robust material-lighting disentanglement across varied lighting conditions.

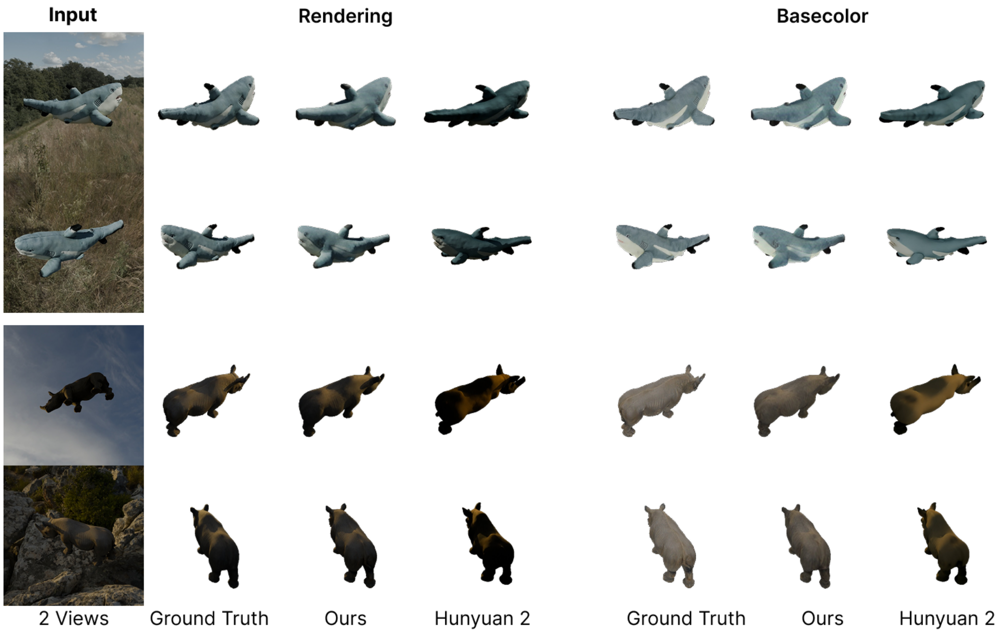

Figure 11: **Failure cases.** Failure cases showing challenges with baked in lighting for objects with strong self shadowing (fin of shark) and basecolor prediction difficulties in dark scenes (Rhino). Comparison includes results from Hunyuan3D and our method.

C.2 REAL-WORLD DATA PREPARATION

The unCommon Objects in 3D (UCO3D) Liu et al. (2024) dataset provides real-world training data, but requires extensive preprocessing to achieve training compatibility with our synthetic data pipeline.

Quality Filtering UCO3D contains numerous challenging samples including motion blur, inaccurate masks, and poor camera estimates. We apply strict quality filtering based on reconstruction and camera estimation scores provided by the dataset's Gaussian Splatting optimization, retaining only objects with scores ≥ 1.0. This filtering dramatically reduces the dataset size but ensures training stability and prevents degraded supervision signals.

Data Preprocessing Pipeline Our preprocessing pipeline, illustrated in, applies several critical transformations:

1. **Square cropping and centering**: Objects are consistently cropped to square aspect ratios and centered within frames
2. **Intrinsic calibration**: Camera intrinsics are carefully adjusted to account for cropping transformations
3. **Valid region tracking**: Due to square cropping, we maintain masks for valid view regions and foreground objects
4. **Surface normal estimation**: Monocular normal estimation provides additional geometric supervision
5. **Scale normalization**: Scene boundaries are rescaled to align with synthetic example scales

This comprehensive preprocessing ensures seamless integration with synthetic training data while preserving the challenging real-world characteristics that drive domain generalization.

Training Integration The processed UCO3D data provides RGB-only supervision without material or illumination ground truth. Our mixed-domain training protocol accommodates this through image-space reconstruction losses while synthetic data provides direct material supervision. This combination enables robust real-world generalization while maintaining physical material properties learned from synthetic supervision.

D LIMITATIONS AND FAILURE CASES

Although rare, failure cases occur where the decomposition fails to disentangle lighting and materials, resulting in baked-in lighting affecting the material maps (Figure 11). This primarily occurs when (i) environment lighting falls outside the RENI++ prior distribution, especially with multiple extremely bright, localized light sources, or (ii) strong self-shadowing leads to baked-in lighting in material maps, or (iii) dark scenes make basecolor prediction challenging. However, even in these challenging cases, ReLi3D still outperforms strong baselines like Hunyuan3D Zhao et al. (2025).

The largest remaining weakness is the relatively limited resolution of the triplane ($3 \times 40 \times 384 \times 384$), limiting texture and geometry resolution in practice, also visible in reconstruction examples against Hunyuan3D. Current blur results primarily stem from this resolution constraint and the DINOv2 fine-tuning bottleneck, not the disentanglement framework itself.

Transparent objects present another limitation: while our density-based NeRF pre-training handles transparency, explicit mesh reconstruction of transparent surfaces remains an open research challenge outside our current scope.

ReLi3D assumes known camera poses and physically plausible materials, which are often violated by generated images from diffusion models. While single-image inputs generally work well when pose estimation is accurate (e.g., from DUST3R Wang et al. (2024a)), severely bad pose estimation leads to blur artifacts. Multi-view generations sometimes degrade performance due to pose and appearance inconsistencies, though pairing generated multi-view images with proxy 3D reconstructions could enable adaptation to this regime in future work.