Geo-ID: Test-Time Geometric Consensus for Cross-View Consistent Intrinsics

Alara Dirik[1] and Stefanos Zafeiriou[1]

Imperial College London, UK
{a.dirik22,s.zafeiriou}@imperial.ac.uk
Project page: https://alaradirik.github.io/geoid/

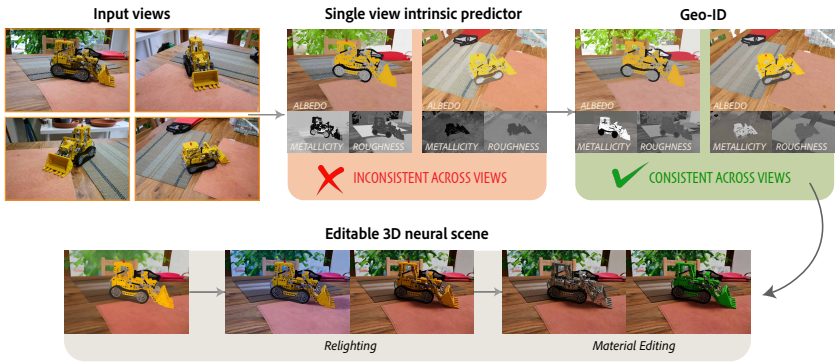

Fig. 1: Existing single-view intrinsic predictors estimate PBR parameters independently per view, leading to cross-view inconsistencies. Geo-ID produces consistent decompositions by coupling per-view predictions through geometric correspondences at test time, enabling coherent material editing and relighting in downstream neural scene representations.

Abstract. Intrinsic image decomposition aims to estimate physically based rendering (PBR) parameters such as albedo, roughness, and metallicity from images. While recent methods achieve strong single-view predictions, applying them independently to multiple views of the same scene often yields inconsistent estimates, limiting their use in downstream applications such as editable neural scenes and 3D reconstruction. Video-based models can improve cross-frame consistency but require dense, ordered sequences and substantial compute, limiting their applicability to sparse, unordered image collections. We propose Geo-ID, a novel test-time framework that repurposes pretrained single-view intrinsic predictors to produce cross-view consistent decompositions by coupling independent per-view predictions through sparse geometric correspondences that form uncertainty-aware consensus targets. Geo-ID is model-agnostic, requires no retraining or inverse rendering, and applies directly to off-the-shelf intrinsic predictors. Experiments on synthetic benchmarks and real-world scenes demonstrate substantial improvements in cross-view intrinsic consistency as the number of views increases, while maintaining comparable single-view decomposition performance. We further show that the resulting consistent intrinsics enable coherent appearance editing and relighting in downstream neural scene representations.

1 Introduction

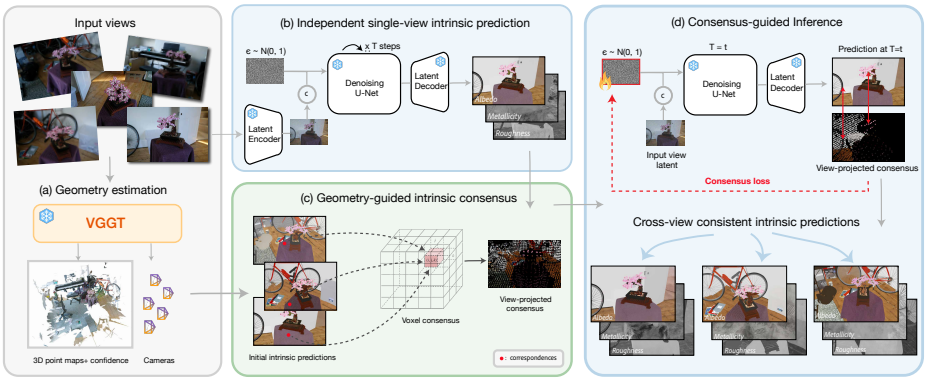

Fig. 2: Overview of Geo-ID. The pipeline consists of three phases. (1) Geometry-guided correspondence estimation: (a) We use a pretrained geometry transformer (VGGT) to predict camera parameters and dense 3D point maps with confidence from the input views. (2) Voxel-based intrinsic consensus: (b) A pretrained single-view diffusion model first produces independent intrinsic predictions for each view. (c) We voxelise high-confidence 3D points and aggregate the corresponding intrinsic values into a robust voxel-level consensus, which we reproject into each image. (3) Consensus-guided diffusion: (d) We run a second diffusion pass per view and inject the view-projected consensus as sparse guidance at selected denoising steps, producing cross-view consistent intrinsic predictions.

Intrinsic image decomposition is a long-standing computer vision problem that aims to recover illumination-invariant surface properties such as albedo, roughness, and metallicity from images. Recent diffusion-based and vision transformer models [5, 8, 12, 21, 24, 38] achieve high-quality predictions from a single view and generalize well across diverse scenes. However, when applied independently to multiple views of the same scene, these models often produce highly inconsistent intrinsic estimates. This inconsistency arises from a fundamental ambiguity: intrinsic decomposition from a single image is severely ill-posed, as geometry, illumination, and material appearance are entangled. Without explicit geometric context, multiple decompositions may be equally plausible, causing the same surface to be assigned different albedo or roughness values depending on viewpoint. Such inconsistencies are a major obstacle to building editable neural scene representations from sparse image collections, where coherent material predictions across views are essential.

Prior work has addressed this issue through two main directions. Video-based intrinsic decomposition models exploit temporal coherence to regularize predictions across frames [28, 32], while multi-view methods either employ architectures trained jointly on multi-view data [10] or rely on inverse rendering pipelines that optimize geometry and materials per scene [23, 29, 40]. While effec-

tive for dense captures or controlled settings, these approaches require specialized training data, heavy per-scene optimization, or accurate and complete geometry, assumptions that rarely hold for sparse, unordered image collections commonly encountered in practice.

In this work, we take a different perspective. Rather than designing a new intrinsic predictor or performing per-scene optimization, we introduce Geo-ID, an inference-time framework that repurposes pretrained single-view intrinsic models for cross-view consistent decomposition. Our key insight is that sparse geometric correspondences provide sufficient structure to resolve cross-view ambiguities, even when geometry is noisy or incomplete. By coupling independent per-view predictions through geometry-induced consensus constraints, we enforce agreement on shared surface properties without requiring dense geometry, retraining, or joint multi-image optimization, while preserving the original generative prior of the base model.

Geo-ID operates entirely at test time and is applicable to existing diffusion-based intrinsic predictors. We demonstrate the approach using RGB↔X [38] and Marigold IID Appearance [21], and evaluate on synthetic benchmarks with ground-truth PBR parameters as well as real-world multi-view scenes. Our results show that Geo-ID substantially improves cross-view intrinsic consistency as the number of views increases, while maintaining comparable single-view decomposition quality. In summary, our contributions are:

- Geo-ID, an inference-time framework that produces cross-view consistent intrinsic decompositions from sparse, unordered image collections by coupling frozen single-view diffusion models through geometric correspondences, without retraining or per-scene optimization.
- A geometry-driven consensus formulation that aggregates per-view intrinsic predictions through sparse 3D correspondences using robust, confidence-weighted statistics.
- A consensus-guided diffusion strategy that injects sparse cross-view constraints into the denoising process while preserving the generative prior of the base model.

2 Related Work

2.1 Intrinsic Image Decomposition

Intrinsic image decomposition seeks to separate an image into geometry-, reflectance-, and illumination-related components, and more recently into physically motivated PBR parameters such as albedo, roughness, and metallicity. Early approaches relied on Retinex-style assumptions or optimization over hand-crafted priors [2, 14, 25]. Subsequent learning-based methods introduced convolutional and transformer architectures trained on synthetic or weakly supervised datasets [4, 11, 26, 30].

More recently, large-scale diffusion models and vision transformers have achieved strong single-view intrinsic predictions. RGB↔X [38], Marigold [20, 21], and

PRISM [8] repurpose pretrained generative priors to infer PBR parameters without task-specific retraining. Other approaches explore factorized representations to disentangle diffuse and specular components [5], or incorporate multi-modal reasoning to refine estimates [9]. While these methods produce high-quality per-view predictions, they operate independently on each image and remain sensitive to view-dependent illumination cues, often yielding individually plausible yet mutually inconsistent material estimates across views of the same surface.

2.2 Multi-View and Video Intrinsic Decomposition

Several lines of work aim to produce consistent intrinsic estimates across multiple views. Inverse rendering pipelines jointly optimize geometry and materials through neural radiance fields or Gaussian splatting under photometric constraints [29,37,40], but require dense captures and costly per-scene optimization. Intrinsic Image Fusion [23] aggregates stochastic single-view predictions into a consistent 3D PBR representation via iterative optimization and path-traced rendering, similarly relying on accurate geometry and dense inputs.

Feed-forward multi-view networks take a different approach. IDT [10] trains a transformer on multi-view inputs to jointly predict reflectance and shading, Event-ID [6] targets object-centric capture using event cameras, and IDArb [27] trains a diffusion model with cross-view attention on large-scale synthetic object data. While effective for isolated objects under controlled settings, these methods require multi-view training data and specialized architectures, limiting their applicability to scene-level, in-the-wild image collections.

Video-based methods such as Diffusion Renderer [28], Ouroboros [32] and UniRelight [16] exploit temporal coherence to regularize predictions across frames, achieving strong inter-frame consistency. However, they assume dense, ordered sequences and are not directly applicable to sparse, unordered multi-view inputs. In contrast, Geo-ID operates on unordered image collections and repurposes frozen single-view predictors without retraining.

2.3 Diffusion Guidance and Test-Time Optimization

Guidance mechanisms steer diffusion models at inference time without modifying their weights. Classifier guidance biases sampling using auxiliary objectives [7], ControlNet [39] introduces spatial conditioning during training, and blended diffusion techniques constrain subsets of the latent space to satisfy partial observations [1]. Test-time optimization has also been explored for inverse problems and image reconstruction [19,33]. Recent work incorporates geometric cues into generative pipelines through camera-aware conditioning or rendering-based supervision [13, 15, 31], but these approaches modify model architectures or require retraining. Geo-ID instead operates purely at inference time, injecting sparse geometric constraints into the denoising process while keeping all model parameters frozen.

3 Method

Our goal is to produce plausible and cross-view consistent intrinsic decompositions from a sparse, unordered set of images using a pretrained single-view diffusion model, without modifying its parameters. To this end, we introduce an inference-time framework that couples otherwise independent diffusion processes through geometry-guided consensus constraints derived from sparse multi-view correspondences. Our framework, which we denote as Geo-ID, consists of three stages: (i) geometry-guided correspondence estimation, (ii) voxel-based consensus initialization, and (iii) consensus-guided diffusion sampling. An overview of the pipeline is shown in Figure 2.

3.1 Problem Setup

Let $\{I_i\}_{i=1}^{V}$ denote V images of a static scene. Given a pretrained diffusion model defining a conditional prior $p_\theta(Y_i \mid I_i)$ over intrinsic maps Y_i (albedo, roughness, and metallicity) for each view independently, our objective is to estimate $\{\hat{Y}_i\}_{i=1}^{V}$ that remain faithful to the per-view prior while being mutually consistent at corresponding 3D scene locations. All consistency constraints are enforced purely at inference time.

Our goal can be viewed as approximating a joint posterior over intrinsic maps conditioned on all views, while retaining the per-view diffusion prior. Rather than learning a new multi-view model, we approximate this joint objective by enforcing agreement only at geometrically supported 3D correspondences. This yields a sparse, geometry-induced coupling between otherwise independent per-view predictions.

Base Intrinsic Predictors. Our framework is agnostic to the choice of diffusion-based intrinsic predictor and can be applied to any such model. We demonstrate it using two complementary pretrained predictors: (i) RGB↔X [38], which estimates each intrinsic modality via separate text-conditioned diffusion passes, and (ii) Marigold IID Appearance [21], which jointly predicts all intrinsic modalities in a single forward pass.

3.2 Geometry-Guided Correspondence Estimation

We obtain approximate multi-view correspondences and camera geometry using VGGT [34], a feed-forward geometry transformer. For each image I_i of an unordered set of images, VGGT predicts a world-frame point cloud $P_i \in \mathbb{R}^{H\times W\times 3}$, a depth map $D_i \in \mathbb{R}^{H\times W}$, camera extrinsics E_i, intrinsics K_i, and per-pixel confidence maps σ_i^P and σ_i^D. All 3D quantities are expressed in the coordinate frame of the first camera. We retain points with confidence $\sigma_i^P \geq \tau_c$ (default $\tau_c = 0.35$), and record the originating view i and pixel location for each retained point. When more images are available than the specified number of views, we randomly subsample V images.

3.3 Voxel-Based Consensus Initialisation

After establishing geometric correspondences, we aggregate per-view intrinsic predictions at corresponding 3D locations into robust consensus estimates. Rather

than relying on VGGT's 2D tracking head, we operate directly on the predicted world-frame point clouds, which avoids error accumulation from track propagation, instability arising from visibility scores associated with tracklets, and naturally supports unordered inputs.

We pool high-confidence 3D points from all views into a single point cloud and partition it into axis-aligned voxels of side length δ. When unspecified, we set $\delta = \alpha \cdot \tilde{d}$, where $\tilde{d}$ denotes the median nearest-neighbour distance in the point cloud and $\alpha = 2.5$. We discard voxels containing points from fewer than $n_{\min}$ distinct views (default $n_{\min} = 2$). We then run the base diffusion model independently on each image I_i to obtain initial intrinsic predictions $\hat{Y}_i^{(0)}$. For each voxel v, we collect intrinsic values by sampling $\hat{Y}_i^{(0)}$ at the pixel locations corresponding to points that fall inside the voxel.

Let $\mathcal{V}(v)$ denote the set of observations contributing to voxel v, and let u_v^j be the image-space location of observation j in its originating view. From the initial per-view predictions $\hat{Y}_i^{(0)}$, we compute a robust consensus estimate and dispersion measure for each voxel:

$$\begin{aligned} s_v &= \text{weighted-median}_{j \in \mathcal{V}(v)} \hat{Y}_j^{(0)}(u_v^j), \\ \hat{\sigma}_v &= 1.4826 \cdot \text{median}_{j \in \mathcal{V}(v)} \big| \hat{Y}_j^{(0)}(u_v^j) - s_v \big|. \end{aligned} \tag{1}$$

where the weighted median uses observation weights $w_v^j = \sigma_i^P(u_v^j) \log\big(1 + |\mathcal{V}(v)|\big)$ combining the per-point geometric confidence $\sigma_i^P(u_v^j)$ with a term that grows with the number of distinct views $|\mathcal{V}(v)|$ contributing to the voxel. We use the weighted median to ensure robustness to misaligned correspondences or stochastic prediction outliers. The dispersion estimate $\hat{\sigma}_v$ provides a robust measure of cross-view disagreement. First, observations whose deviation from the consensus exceeds a multiple of $\hat{\sigma}_v$ are treated as outliers and excluded from guidance. Second, the inverse variance $\hat{\sigma}_v^{-1}$ modulates the strength of the guidance signal, assigning lower weight to voxels with high intrinsic ambiguity.

To construct per-view guidance targets, we project voxel centers into each camera using the predicted intrinsics and extrinsics, and verify visibility using the predicted depth maps. A voxel is considered visible in view i if its projected pixel location lies within the image bounds and its projected depth z_v^i agrees with the predicted depth D_i within a relative tolerance $\epsilon = 0.05$:

$$m_v^i = 1\left[\text{in-bounds}(u_v^i) \ \wedge \ \frac{|z_v^i - D_i(u_v^i)|}{D_i(u_v^i)} < \epsilon \right]. \tag{2}$$

This yields, for each view i, a sparse set of visible consensus targets $\{(u_v^i, s_v, w_v^i)\}$, with weight w_v^i combining the voxel view count and inverse dispersion $\hat{\sigma}_v^{-1}$.

3.4 Inference-Time Consensus Guidance

We obtain the final intrinsic predictions by re-running diffusion sampling for each view while softly enforcing agreement with the consensus targets derived in Section 3.3. Let $x_{t,i}$ denote the diffusion latent corresponding to view i at denoising timestep t. At selected timesteps $t \in \mathcal{S}$, we estimate the corresponding

clean intrinsic prediction $\hat{Y}_{t,i}$ from the current latent using the Tweedie formula. Specifically, given the model's predicted noise (or v-prediction), we recover an estimate of the denoised sample, decode it through the VAE, and evaluate a consistency loss over visible consensus targets:

$$\mathcal{L}_{t,i} = \sum_{v:\, m_v^i = 1} w_v^i \, \rho\big(\hat{Y}_{t,i}(u_v^i) - s_v\big), \tag{3}$$

where s_v denotes the voxel consensus value, u_v^i its projected pixel location in view i, w_v^i the corresponding confidence weight, and $\rho(\cdot)$ is a Huber loss with threshold $\delta_H = 1.0$. We incorporate this loss by applying a single gradient-based update to the latent,

$$x_{t,i} \leftarrow x_{t,i} - \eta_t \, \nabla_{x_{t,i}} \mathcal{L}_{t,i}, \tag{4}$$

followed by the standard denoising step of the base diffusion scheduler. Only the latent variables are updated; all model parameters remain frozen.

Early diffusion steps primarily determine global structure and coarse material layout, while later steps refine high-frequency detail. Applying guidance during early steps can therefore override structural decisions made by the pretrained prior. We apply guidance only during the last 80% of denoising steps (i.e., at lower noise levels) and only at voxel-supported pixel coordinates. This allows the generative model to establish coarse structure and material layout in the initial high-noise steps before we enforce cross-view agreement. Although the constraints are shared across views, sampling remains independent per view, avoiding joint multi-image optimization.

3.5 Implementation Details

We conduct all experiments on a single NVIDIA A40 GPU. For RGB↔X, we use 50 DDIM sampling steps; for Marigold IID Appearance, which is LCM-distilled, we use 10 steps. In both cases, consensus guidance is applied only during the last 80% of the denoising steps. Visibility is verified using predicted depth with tolerance $\epsilon = 0.05$. All intrinsic predictions and losses are computed in linear RGB space. All reported results use a single fixed configuration ($\alpha = 2.5$, $\tau_c = 0.35$, $n_{\min} = 2$, $\epsilon = 0.05$, $\delta_H = 1.0$) across every scene and dataset, with no per-scene tuning. We reserve 20% of voxels as a held-out set for consistency evaluation.

4 Experiments

We evaluate Geo-ID along three axes: (i) cross-view intrinsic consistency, (ii) per-view decomposition quality, and (iii) downstream applicability to editable 3D scene representations. We first describe the datasets, baselines, and evaluation metrics used in our study. We then quantify improvements in cross-view agreement across varying numbers of input views, while verifying that per-view decomposition accuracy is preserved. Finally, we demonstrate the practical impact of consistent intrinsics through relightable neural scene reconstruction and conduct ablations to analyze the contribution of key design choices.

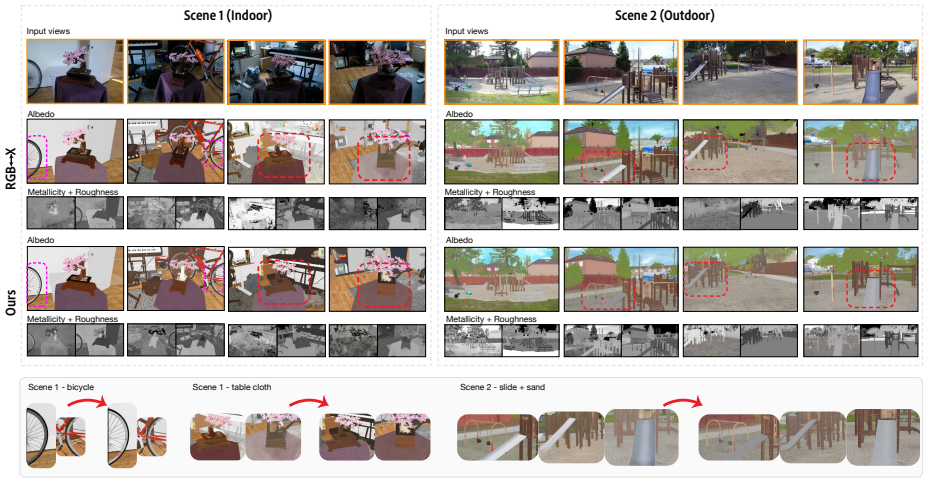

Fig. 3: Qualitative comparison. For two scenes (one indoor MipNeRF-360, one outdoor Tanks & Temples), we show input views alongside albedo, metallicity, and roughness predictions from the base model (RGB↔X) applied independently per view and with our Geo-ID guidance. Without guidance, the base model produces plausible but inconsistent decompositions: note the color drift across views on the same surfaces (highlighted regions). Bottom row: zoomed-in crops of corresponding surface regions across two views, showing how Geo-ID reduces cross-view disagreement while preserving fine detail and decomposition quality.

4.1 Experimental Setup

Datasets. We evaluate decomposition quality on the synthetic InteriorVerse [41] and HyperSim [30]. InteriorVerse provides ground-truth albedo, metallicity and roughness maps, whereas HyperSim only provides albedo maps. We use the official test split of HyperSim and follow the same train/test split as RGB↔X for InteriorVerse. For cross-view consistency on real-world data, we use scenes from MipNeRF-360 [3] and Tanks&Temples [22]. From MipNeRF-360, we evaluate seven scenes (four indoor: *bonsai*, *counter*, *kitchen*, *room*; three outdoor: *bicycle*, *garden*, *stump*), each containing approximately 150 images. From Tanks&Temples [22], we use the intermediate and advanced scene splits, which provide unordered multi-view captures without ground-truth intrinsics. For each scene and view count $V \in \{4, 8, 16, 32\}$, we sample $k = 5$ random view subsets and report averages.

Baselines. We compare against the unguided single-view predictions of our two base models, RGB↔X [38] and Marigold IID Appearance [21], as well as Colorful Diffusion (CID) [5], a SOTA single-view model for albedo and irradiance estimation. We additionally compare against the intrinsic decomposition module of the video-based Diffusion Renderer [28], and evaluate it both in our target unordered, sparse view setting and in its original temporally ordered setting, which serves as our video-based performance upper-bound. All baselines are evaluated using the official model checkpoints and implementations.

Fig. 4: Qualitative comparison with single-view, multi-view, and video intrinsic decomposition baselines on the MipNeRF-360 Garden scene. Each row shows the albedo prediction for a different input view. Single-view methods (RGB↔X, IDArb) produce detailed but cross-view inconsistent estimates, while video-based approaches (Diffusion Renderer) generates multi-view consistent results, albeit fails to remove lighting effects. Geo-ID (rightmost column, 16-view setting) maintains the detail of single-view predictions while achieving cross-view consistency.

Metrics. For decomposition quality, we report PSNR, SSIM, and LPIPS for albedo, and RMSE for roughness and metallicity, computed against ground truth on InteriorVerse and HyperSim test sets. For cross-view consistency, we use geometry-aligned correspondences from the held-out 20% of voxels not used during guidance. For each held-out voxel v observed in views $\mathcal{V}(v)$, we compute the median absolute deviation (MAD) of predicted intrinsic values:

$$\mathcal{C}_v = \mathrm{MAD}_{i \in \mathcal{V}(v)}\big(\hat{Y}_i(u_v^i)\big). \tag{5}$$

We report the average $\mathcal{C}_v$ across held-out voxels, separately for each modality. Lower values indicate better cross-view agreement. For Diffusion Renderer and Colorful Diffuse (CID), which do not use our pipeline, we run VGGT on the same 32-view subsets and evaluate consistency on the same held-out correspondences.

4.2 Cross-View Consistency

We first evaluate cross-view intrinsic consistency on real-world multi-view datasets and report results on MipNeRF-360 and Tanks&Temples in Table 1. When applied independently to each view, both base models exhibit substantial disagreement across views, particularly for roughness and metallicity, which are more sensitive to view-dependent ambiguities. Geo-ID consistently reduces cross-view disagreement across all modalities and view counts. On MipNeRF-360, consistency improves monotonically with the number of views for both base models. For Marigold Appearance, indoor albedo MAD decreases from 0.091 (unguided) to 0.076 at 32 views, while outdoor metallicity drops from 0.070 to 0.044.

Table 1: Cross-view intrinsic consistency (mean per-correspondence MAD, ↓). MipNeRF-360 Indoor: *bonsai, counter, kitchen, room*; Outdoor: *bicycle, garden, stump*. Tanks & Temples and MipNeRF-360 use VGGT correspondences; InteriorVerse uses ground truth. Diffusion Renderer is shown *ordered* (video upper bound) and *unordered* (matched to Geo-ID's subsets). Geo-ID improves consistency across all view counts without retraining.

Method	Views	Mip-NeRF Ind.			Mip-NeRF Out.			Tanks & Temples			InteriorVerse (GT)		
		Albedo	Rough.	Metal.	Albedo	Rough.	Metal.	Albedo	Rough.	Metal.	Albedo	Rough.	Metal.
IDArb [27]	16	0.258	0.316	0.289	0.224	0.310	0.280	0.256	0.309	0.299	0.241	0.298	0.275
CID [5]	32	0.101	–	–	0.070	–	–	0.103	–	–	0.095	–	–
Diff. Renderer [28] (ordered)	32	0.068	0.038	0.043	0.043	0.041	0.019	0.060	0.040	0.028	0.058	0.034	0.037
Diff. Renderer [28] (unordered)	32	0.104	0.075	0.087	0.089	0.100	0.102	0.110	0.100	0.129	0.108	0.086	0.120
RGB↔X [38]	32	0.114	0.100	0.206	0.070	0.144	0.192	0.109	0.105	0.219	0.107	0.096	0.198
Marigold Appr. [21]	32	0.091	0.096	0.111	0.073	0.080	0.070	0.098	0.100	0.118	0.085	0.089	0.103
RGB↔X + Geo-ID	4	0.110	0.074	0.138	0.067	0.121	0.115	0.106	0.096	0.195	0.103	0.072	0.132
	8	0.107	0.072	0.128	0.061	0.118	0.101	0.102	0.089	0.180	0.099	0.068	0.121
	16	0.103	0.071	0.115	0.058	0.104	0.097	0.100	0.082	0.174	0.095	0.065	0.109
	32	0.098	0.065	0.114	0.054	0.079	0.085	0.097	0.076	0.166	0.091	0.060	0.105
Marigold Appr. + Geo-ID	4	0.086	0.088	0.101	0.067	0.072	0.062	0.089	0.090	0.104	0.080	0.082	0.094
	8	0.081	0.086	0.103	0.064	0.069	0.059	0.085	0.085	0.099	0.075	0.078	0.091
	16	0.078	0.079	0.099	0.063	0.064	0.049	0.083	0.082	0.097	0.072	0.073	0.087
	32	0.076	0.082	0.100	0.061	0.062	0.044	0.082	0.080	0.095	0.070	0.075	0.085

RGB↔X follows similar trends, with the largest relative gains on roughness and metallicity.

We evaluate Diffusion Renderer under two settings: *ordered*, where we input the images in their original temporal sequence and read as a video upper bound, and *unordered*, applied to the same randomized subsets as Geo-ID. As shown in Table 1, removing temporal order degrades the Diffusion Renderer's performance sharply: on MipNeRF-360 outdoor, albedo, roughness, and metallicity MAD rise by roughly 2×, 2.4×, and 5.4×. Being permutation-invariant by construction, our Geo-ID beats this matched unordered baseline on albedo across all real-world splits, on every outdoor modality, and on all three InteriorVerse modalities; on the more challenging Tanks&Temples scenes it leads on albedo and roughness, with metallicity the hardest modality. The ordered upper bound retains an edge only on indoor roughness and metallicity, where dense temporal supervision is strongest. Consistency improvements scale approximately linearly with view count across all modalities and datasets without saturating in the range we evaluate. We note that the gains hold on InteriorVerse with ground-truth correspondences (Table 1, rightmost columns), where for Marigold Appearance, albedo MAD drops from 0.085 to 0.070 at 32 views, and hence ruling out mere alignment with VGGT's geometric biases. See the supplementary for per-scene breakdowns and variance across subsets.

4.3 Decomposition Quality

We next evaluate whether enforcing cross-view consistency degrades per-view intrinsic accuracy. Table 2 shows that Geo-ID preserves the decomposition quality of both base models on InteriorVerse and HyperSim. For Marigold Appearance, PSNR, SSIM, and LPIPS remain within ±0.1 dB, ±0.01, and ±0.01 of the unguided baseline across all view counts; RGB↔X exhibits a similar pattern. These results indicate that enforcing cross-view agreement through sparse geometric

Table 2: Per-view intrinsic decomposition quality on InteriorVerse (left) and HyperSim (right). Geo-ID achieves cross-view consistency (Table 1) without degrading single-view accuracy.

Method	Num. Views	InteriorVerse					HyperSim		
		Albedo			Metallic	Roughness	Albedo		
		PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
IDArb [27]	1	9.7	0.66	0.61	0.61	0.43	9.4	0.62	0.63
CID [5]	1	17.7	0.79	0.27	–	–	17.1	0.78	0.21
Kocsis et al. [24]	1	17.4	0.80	0.22	0.21	0.26	12.1	0.72	0.41
Diffusion Renderer [28]	24	21.9	0.87	0.17	0.28	0.35	22.2	0.78	0.21
RGBX [38]	1	16.4	0.78	0.19	0.44	0.38	17.4	0.81	0.18
Marigold Appr. [21]	1	19.5	0.85	0.19	0.20	0.25	18.5	0.78	0.20
RGB↔X + Geo-ID	4	16.2	0.77	0.20	0.44	0.38	17.2	0.80	0.19
	8	16.3	0.78	0.19	0.42	0.36	17.3	0.81	0.18
	16	16.4	0.78	0.19	0.42	0.35	17.4	0.81	0.18
	32	16.4	0.78	0.19	0.42	0.35	17.3	0.81	0.18
Marigold Appr. + Geo-ID	4	19.3	0.84	0.20	0.21	0.26	18.3	0.77	0.21
	8	19.4	0.85	0.19	0.20	0.24	18.4	0.78	0.20
	16	19.5	0.85	0.19	0.20	0.24	18.5	0.78	0.20
	32	19.5	0.86	0.19	0.19	0.23	18.5	0.78	0.20

constraints does not degrade the intrinsic estimates produced by the underlying models.

In some cases (e.g., 16 views), guided predictions slightly improve per-view metrics, indicating that the consensus signal from multiple views can correct isolated single-view errors. Diffusion Renderer [28], a video-based intrinsic decomposition method, achieves strong per-view accuracy in our evaluation, reflecting the benefits of temporal supervision and joint reasoning across dense frame sequences. However, such approaches assume ordered video input and are trained specifically for multi-frame inference. In contrast, Geo-ID is designed as a lightweight inference-time wrapper for existing single-view intrinsic predictors. It introduces sparse geometry-aligned constraints during sampling, enabling cross-view consistency while preserving the original per-view generative prior of the base model.

4.4 Ablation Studies

We ablate key design choices on MipNeRF-360 indoor scenes at 16 views using RGB↔X as the base model and summarize our results in Table 3. We first study the effect of the consensus aggregation strategy. Replacing the weighted median with a weighted mean yields slightly better consistency but at the cost of decomposition quality, as the mean is more sensitive to outlier correspondences that pull predictions away from the base model's prior.

Next, we examine the impact of the guidance schedule. Applying guidance throughout the entire denoising process improves consistency but degrades per-view decomposition quality, suggesting that early-step constraints interfere with structural decisions made by the diffusion prior. Conversely, restricting guidance to only the final 20% of steps yields weaker consistency gains. Guiding the intermediate-to-late stages provides the best balance between consistency and fidelity.

Finally, we compare our voxel-based 3D correspondence formulation with VGGT's tracking head, which establishes 2D correspondences by propagating point tracks across views. In this ablation, we replace voxelized 3D aggregation

with track-based correspondence and enforce consistency directly at tracked pixel locations. This variant yields worse cross-view consistency, likely because 2D track propagation accumulates drift and relies on reliable visibility ordering, assumptions that are often violated in sparse, unordered multi-view settings. Additional ablation experiments on confidence threshold and voxel size choice, and qualitative examples are provided in the supplementary.

4.5 Runtime Analysis

We report per-view and total inference times in Table 4. The geometry extraction with VGGT is a one-time cost that amortises over views (3–6s depending on view count). The dominant cost is the guided diffusion pass, which adds ∼30% overhead per view compared to unguided inference due to the VAE decode and gradient computation at guided steps. For full PBR prediction, Diffusion Renderer requires ∼1.77 min per modality (∼5.3 min total at 16 views). Geo-ID with Marigold Appearance, which predicts all three modalities jointly, completes in ∼3 min (roughly 40% faster), while additionally supporting sparse, unordered inputs.

4.6 Qualitative Results

We present qualitative results on two representative scenes in Figure 3: an indoor scene from MipNeRF-360 and an outdoor scene from Tanks&Temples. We show albedo, roughness, and metallicity predictions from RGB↔X applied independently per view alongside predictions refined with Geo-ID. For both scenes, we run our method using 16 views. Without guidance, the base model produces plausible per-view decompositions but exhibits clear cross-view inconsistencies, such as color drift on the tablecloth and bicycle frame (Scene 1) and shifting sand and slide tones (Scene 2). We note that Geo-ID substantially reduces these discrepancies while preserving fine detail, as the zoomed-in crops in the bottom row confirm.

In Figure 4, we compare against single-view (RGB↔X, IDArb), multi-view (Colorful Diffuse), and video (Diffusion Renderer) intrinsic decomposition baselines on the MipNeRF-360 Garden scene. We observe that single-view methods

Table 3: Ablation study (16 views, RGB↔X). Consistency: mean MAD (↓) on held-out correspondences (MipNeRF-360 indoor). Quality: decomposition accuracy on the InteriorVerse test set. Aggressive guidance (no outlier rejection, 100% steps) improves consistency but degrades decomposition quality, while our full model balances both.

	Consistency (MAD ↓)			**Decomposition Quality (InteriorVerse)**				
				Albedo			Metal.	Rough.
Variant	Albedo	Rough.	Metal.	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓	RMSE ↓
Full model (Geo-ID)	**0.103**	**0.071**	**0.115**	**16.4**	**0.78**	**0.19**	**0.44**	**0.38**
Mean consensus (instead of median)	0.098	0.066	0.099	16.1	0.76	0.20	0.47	0.41
No outlier rejection	0.100	0.067	0.112	16.2	0.75	0.20	0.45	0.41
Guidance at 100% of steps	0.096	0.067	0.106	15.8	0.71	0.22	0.48	0.44
Guidance at last 20% of steps	0.111	0.093	0.188	16.4	0.78	0.19	0.44	0.38
2D tracking head (instead of 3D voxels)	0.109	0.074	0.135	16.2	0.75	0.21	0.47	0.42
Unguided baseline	0.114	0.100	0.206	16.4	0.78	0.19	0.44	0.38

Table 4: Runtime analysis on a single NVIDIA A40 GPU (16 views). Geo-ID adds moderate overhead to the base model, remains competitive with video-based Diffusion Renderer for full PBR prediction, and is orders of magnitude faster than optimisation-based multi-view intrinsic methods. †Requires known geometry and dense captures.

	Ours (16 views)			
Component	Marigold Appr.	RGB↔X	Diff. Renderer	Int. Img. Fusion†
Geometry estimation	6 s	6 s	–	known
Unguided init (per view)	0.9 s	3 s	–	–
Consensus computation	32 s	32 s	–	–
Guided pass (per view)	7.6 s	40 s	–	–
Inference (per modality)	–	–	1.77 min	–
Per-scene optimisation	–	–	–	>1 hr
Total	**∼3 min**	**∼22 min**	**∼5.3 min**	**>1 hr**

produce high-quality per-frame predictions but suffer from cross-view inconsistencies, while Diffusion Renderer generates consistent, albeit poorer performance with respect to albedo quality. Our method combines the strengths of both: it maintains the detail of single-view predictions while enforcing cross-view consistency through geometric correspondences. We provide additional examples and failure cases in the supplementary.

4.7 Applications

We demonstrate the utility of cross-view consistent intrinsics by enabling relightable and editable neural scene representations. To this end, we train a triangular mesh-based neural scene representation using MeshSplatting [17] on albedo maps predicted by Geo-ID for indoor scenes from Mip-NeRF 360. To isolate intrinsic appearance, we replace RGB inputs with predicted albedo maps and disable spherical-harmonic lighting effects by setting `sh_degree` = 0. We follow the original MeshSplatting training setup using monocular depth from Depth Anything V2 [35] and surface normals from StableNormals [36], without further modifications. While a direct quantitative evaluation of relighting quality would require ground-truth intrinsics and lighting that are unavailable for real-world scenes, we quantify the downstream benefit of consistency through albedo novel-view synthesis (Table 5) and provide qualitative demonstrations of relighting and material editing in Figure 5.

Quantitative Downstream Evaluation. To quantify how cross-view consistency affects downstream reconstruction, we measure albedo novel-view synthesis. For each method, we train MeshSplatting on a subset of its predicted albedo maps and render at held-out poses, evaluating the rendered views against that method's own predictions at those poses. This deliberately isolates multi-view coherence from absolute correctness: it asks whether a method's per-view albedos form a coherent 3D signal representable by a single editable neural scene, which is the downstream use case our Geo-ID method targets, rather than how close they are to ground truth. As reported in Table 5, Geo-ID improves over its unguided bases by +1.8 dB (RGB↔X) and +1.9 dB (Marigold Appearance) in PSNR, with LPIPS reductions of 0.07 and 0.06 respectively. Without guidance, inconsistent per-view albedos provide contradictory supervision for the same 3D surface;

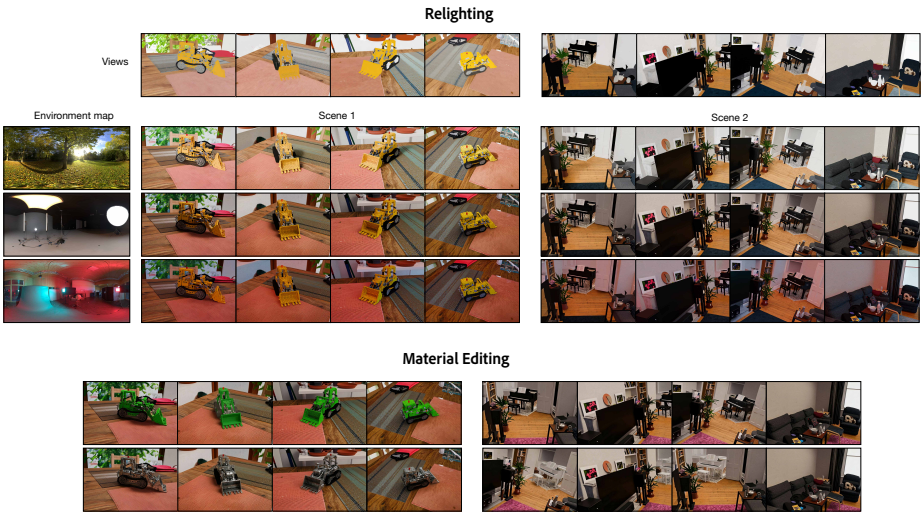

Fig. 5: Downstream applications. *Top:* We train MeshSplatting [17] on Geo-ID-predicted albedo maps and render the reconstructed surface meshes under novel HDR environment maps. Rows correspond to different lighting conditions; columns show rendered views. *Bottom:* We manually segment target regions on the extracted meshes and modify their albedo to demonstrate material editing.

Geo-ID supplies a more coherent multi-view signal. Both Geo-ID variants also outperform ordered Diffusion Renderer, whose residual baked-in lighting (Figure 4) introduces spurious view-dependence that our view-independent albedo training setup (spherical-harmonic lighting disabled, as described above) cannot absorb.

Relighting. We render the reconstructed surface meshes under novel HDR environment maps using physically based rendering in Blender [18]. As shown in Figure 5, our Geo-ID framework produces consistent albedo across views, the resulting meshes exhibit coherent appearance changes under varying illumination. Inconsistent per-view estimates would instead risk introducing visible seams and color discontinuities, as the mesh would bake conflicting color information into adjacent faces.

Material Editing. In addition to relighting, we show that cross-view consistent albedo also enables straightforward scene-level material editing. We manually segment target regions on the extracted meshes and modify their albedo values (e.g., recoloring an object or changing its surface finish). Since Geo-ID ensures that the underlying albedo is view-independent, these edits propagate coherently across all rendered viewpoints. Without consistent estimates, such edits would produce view-dependent color artifacts, as each face would encode a different estimate of the same surface.

Table 5: Downstream albedo novel-view synthesis. We train MeshSplatting [17] on each method's predicted albedo maps and render at held-out poses, comparing against that method's own held-out predictions. This isolates multi-view coherence from absolute correctness. Geo-ID yields the most coherent multi-view albedo signal, outperforming both its unguided bases and ordered Diffusion Renderer.

Training input	PSNR ↑	SSIM ↑	LPIPS ↓
Diffusion Renderer [28] (ordered)	20.4	0.79	0.26
Unguided RGB↔X [38]	19.4	0.71	0.28
RGB↔X + Geo-ID	21.2	0.78	0.21
Unguided Marigold Appr. [21]	20.1	0.74	0.25
Marigold Appr. + Geo-ID	**22.0**	**0.81**	**0.19**

5 Conclusion

We introduced Geo-ID, a training-free framework for enforcing cross-view consistency in intrinsic image decomposition via geometry-guided diffusion at inference time. By coupling independent single-view predictors through sparse geometric correspondences, our approach repurposes pretrained intrinsic models for multi-view settings without retraining or per-scene optimization. Experiments on synthetic benchmarks and real-world scenes show that Geo-ID substantially improves cross-view consistency while preserving per-view decomposition quality, and enables downstream applications such as coherent relighting and material editing in neural scene representations.

Limitations. Our approach inherits the failure modes of both its upstream components. First, when the geometry estimator (VGGT) produces severely incorrect correspondences (e.g., in textureless or repetitive regions), the consensus targets become unreliable and can degrade rather than improve predictions. Second, Geo-ID cannot correct systematic errors in the base intrinsic predictor — if the underlying model consistently misattributes lighting as albedo, geometric consensus will reinforce rather than resolve the error. Finally, while we demonstrate improvements on both indoor and outdoor scenes, current single-view intrinsic predictors remain biased toward indoor, Lambertian settings, which bounds the quality achievable by any test-time method including ours.

Acknowledgments. Stefanos Zafeiriou and Alara Dirik have been partially funded by Turing AI Fellowship (EP/Z534699/1).

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18187–18197 (2021), https://api.semanticscholar.org/CorpusID:244714366 4
2. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. ICCV pp. 334–341 (2012), https://api.semanticscholar.org/CorpusID:1342950 3

3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5460–5469 (2021), https://api.semanticscholar.org/CorpusID:244488448 8
4. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM TOG **33**(4), 159:1–12 (2014) 3
5. Careaga, C., Aksoy, Y.: Colorful diffuse intrinsic image decomposition in the wild. ArXiv **abs/2409.13690** (2024), https://api.semanticscholar.org/CorpusID:272770308 2, 4, 8, 10, 11
6. Chen, Z., Lu, Z., Ma, D., Tang, H., Jiang, X., Zheng, Q., Pan, G.: Event-id: Intrinsic decomposition using an event camera. Proceedings of the 32nd ACM International Conference on Multimedia (2024), https://api.semanticscholar.org/CorpusID:273642870 4
7. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. arXiv preprint arXiv:2105.05233 (2021) 4
8. Dirik, A., Wang, T., Ceylan, D., Zafeiriou, S., Frühstück, A.: PRISM: A unified framework for photorealistic reconstruction and intrinsic scene modeling. arXiv preprint arXiv:2504.14219 (2025) 2, 4
9. Dirik, A., Wang, T., Ceylan, D., Zafeiriou, S., Fruhstuck, A.: Reasonx: Mllm-guided intrinsic image decomposition. ArXiv **abs/2512.04222** (2025), https://api.semanticscholar.org/CorpusID:283557607 4
10. Du, K., Guan, Y., Wang, Z.: Idt: A physically grounded transformer for feed-forward multi-view intrinsic decomposition. ArXiv **abs/2512.23667** (2025), https://api.semanticscholar.org/CorpusID:284311259 2, 4
11. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: Revisiting deep intrinsic image decompositions. In: CVPR. pp. 8944–8952 (2018) 3
12. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. ArXiv **abs/2403.12013** (2024), https://api.semanticscholar.org/CorpusID:268532174 2
13. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. ArXiv **abs/2405.10314** (2024), https://api.semanticscholar.org/CorpusID:269791465 4
14. Gehler, P., Rother, C., Kiefel, M., Zhang, L., Scholkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. In: NeurIPS (2011), https://api.semanticscholar.org/CorpusID:6835647 3
15. Guo, H., Zhu, H., Peng, S., Lin, H., Yan, Y., Xie, T., Wang, W., Zhou, X., Bao, H.: Multi-view reconstruction via sfm-guided monocular depth estimation. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5272–5282 (2025), https://api.semanticscholar.org/CorpusID:277103982 4
16. He, K., Liang, R., Munkberg, J., Hasselgren, J., Vijaykumar, N., Keller, A., Fidler, S., Gilitschenski, I., Gojcic, Z., Wang, Z.: Unirelight: Learning joint decomposition and synthesis for video relighting. ArXiv **abs/2506.15673** (2025), https://api.semanticscholar.org/CorpusID:279447579 4
17. Held, J., Son, S., Vandeghen, R., Rebain, D., Gadelha, M., Zhou, Y., Cioppa, A., Lin, M.C., Droogenbroeck, M.V., Tagliasacchi, A.: Meshsplatting: Differentiable rendering with opaque meshes. ArXiv **abs/2512.06818** (2025), https://api.semanticscholar.org/CorpusID:283693921 13, 14, 15
18. Hess, R.: Blender Foundations: The Essential Guide to Learning Blender 2.6. Focal Press (2010) 14

19. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. ArXiv **abs/2201.11793** (2022), https://api.semanticscholar.org/CorpusID:246411364 4
20. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024) 3
21. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE transactions on pattern analysis and machine intelligence **PP** (2025), https://api.semanticscholar.org/CorpusID:278602942 2, 3, 5, 8, 10, 11, 15
22. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics **36**(4) (2017) 8
23. Kocsis, P., Höllein, L., Nießner, M.: Intrinsic image fusion for multi-view 3d material reconstruction. ArXiv **abs/2512.13157** (2025), https://api.semanticscholar.org/CorpusID:283896051 2, 4
24. Kocsis, P., Sitzmann, V., Nießner, M.: Intrinsic image diffusion for indoor single-view material estimation. CVPR pp. 5198–5208 (2024) 2, 11
25. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical Society of America **61 1**, 1–11 (1971), https://api.semanticscholar.org/CorpusID:14430259 3
26. Li, Z., Yu, T.W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.Y., Zhu, R., Gundavarapu, N., Shi, J., Bi, S., Yu, H.X., Xu, Z., Sunkavalli, K., Hašan, M., Ramamoorthi, R., Chandraker, M.: OpenRooms: An open framework for photorealistic indoor scene datasets. In: CVPR (2021) 3
27. Li, Z., Wu, T., Tan, J., Zhang, M., Wang, J., Lin, D.: Idarb: Intrinsic decomposition for arbitrary number of input views and illuminations. ArXiv **abs/2412.12083** (2024), https://api.semanticscholar.org/CorpusID:274789449 4, 10, 11
28. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, Z.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., Wang, Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 26069–26080 (2025), https://api.semanticscholar.org/CorpusID:275993796 2, 4, 8, 10, 11, 15
29. Liang, Z., Zhang, Q., Feng, Y., Shan, Y., Jia, K.: Gs-ir: 3d gaussian splatting for inverse rendering. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 21644–21653 (2023), https://api.semanticscholar.org/CorpusID:265466404 2, 4
30. 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: ICCV. pp. 10912–10922 (2021) 3, 8
31. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. ArXiv **abs/2308.16512** (2023), https://api.semanticscholar.org/CorpusID:261395233 4
32. Sun, S., Wang, Y., Zhang, H., Xiong, Y., Ren, Q., Fang, R., Xie, X., You, C.: Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse rendering. ArXiv **abs/2508.14461** (2025), https://api.semanticscholar.org/CorpusID:280692572 2, 4
33. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1921–1930 (2022), https://api.semanticscholar.org/CorpusID:253801961 4

34. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: Vggt: Visual geometry grounded transformer. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5294–5306 (2025), https://api.semanticscholar.org/CorpusID:277043968 5
35. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. ArXiv **abs/2406.09414** (2024), https://api.semanticscholar.org/CorpusID:270440448 13
36. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: StableNormal: Reducing diffusion variance for stable and sharp normal. ACM TOG **43**, 1 – 18 (2024), https://api.semanticscholar.org/CorpusID:270702725 13
37. Ye, K., Gao, C., Li, G., Chen, W., Chen, B.: Geosplatting: Towards geometry guided gaussian splatting for physically-based inverse rendering. ArXiv **abs/2410.24204** (2024), https://api.semanticscholar.org/CorpusID:273707771 4
38. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: RGBX: Image decomposition and synthesis using material- and lighting-aware diffusion models (2024) 2, 3, 5, 8, 10, 11, 15
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. ICCV pp. 3813–3824 (2023), https://api.semanticscholar.org/CorpusID:256827727 4
40. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P.E., Freeman, W.T., Barron, J.T.: Nerfactor. ACM Transactions on Graphics (TOG) **40**, 1 – 18 (2021), https://api.semanticscholar.org/CorpusID:235313848 2, 4
41. Zhu, J., Luan, F., Huo, Y., Lin, Z., Zhong, Z., Xi, D., Wang, R., Bao, H., Zheng, J., Tang, R.: Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. ACM (2022), https://doi.org/10.1145/3550469.3555407 8