UNCERTAINTY- AND PARALLAX-AWARE GAUSSIAN SPLATTING FOR INDOOR OBJECT-CENTRIC REFLECTIVE SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Specular reflections remain a major challenge for novel view synthesis, requiring both reliable localization of reflective regions and accurate modeling of their view-dependent appearance. However, reflective regions are difficult to localize consistently across views, while existing reflection models often assume distant environment lighting. This assumption breaks down indoors, where finite-distance scene geometry induces strong reflection parallax, leading to baked-in appearance and view-inconsistent artifacts. To address these challenges, we leverage learned photometric uncertainty to derive metallic supervision through Uncertainty-Guided Metallic Supervision (UGMS), enabling stable diffuse-specular decomposition, and incorporate Parallax-Corrected Cubemaps (PCC) to account for finite-distance environment lighting. We further introduce IndoorRef12, a synthetic benchmark featuring near-field reflection parallax and cluttered backgrounds. It comprises 12 diverse indoor environments with varying scene layouts, reflective-object configurations, and background complexity. Extensive experiments demonstrate substantial improvements in reflection reconstruction, particularly in challenging indoor scenes with strong near-field parallax. We will release our code and IndoorRef12 dataset.

1 INTRODUCTION

Specular reflections on highly metallic objects remain a major bottleneck for novel view synthesis (NVS) in indoor scenes. While outdoor reflections often stem from distant illumination, the reflections on indoor objects are strongly influenced by near-field geometry, including nearby walls, ceilings, and furniture. This induces complex reflection parallax, causing the reflected content to shift with viewpoint and producing sharp view-dependent appearance changes that are difficult to model consistently. Recent reflection-aware methods adopt a physically motivated diffuse–specular decomposition and optimize material and lighting parameters via inverse rendering (Yao et al., 2024; Zhang et al., 2025b; Tong et al., 2025; Tang et al., 2025; Zhou et al., 2025; Zhang et al., 2025a; Yang et al., 2024; Gu et al., 2025; Xu et al., 2023). However, they rely on assumptions that are fragile in indoor scenes, leading to limitations. First, inaccurate identification of reflective objects leads to unstable diffuse–specular decomposition. Under such ambiguity, optimization tends to suppress specular reflections by reducing metallicity or specular strength to minimize photometric error. This causes view-dependent reflection effects to be absorbed into the diffuse appearance, resulting in *baked-in artifacts* that eliminate the dynamic nature of reflection (Fig. 1: Left). These failures are exacerbated in indoor scenes with mixed illumination. Mixed illumination from nearby sources and inter-reflections amplifies view-dependent errors, further obscuring reflective object boundaries Wu et al. (2025). Second, most methods assume an infinite distance environment map, where lighting depends only on direction. While this is reasonable outdoors, it breaks down indoors because dominant reflectors such as walls and furniture lie at finite distances. As a result, direction only lighting fails to capture spatially varying near-field reflection shifts, leading to flattened highlights and misaligned reflection anchoring across views (Fig. 1: Right). These observations highlight the need for stable reflective object localization and finite distance illumination modeling in indoor reflective NVS.

Motivated by these observations, we propose a reflection-aware GS framework for object-centric reflective scenes in indoor environments that stabilizes reflective-object localization and models finite-distance illumination. To stabilize localization, we introduce Uncertainty-Guided Metallic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

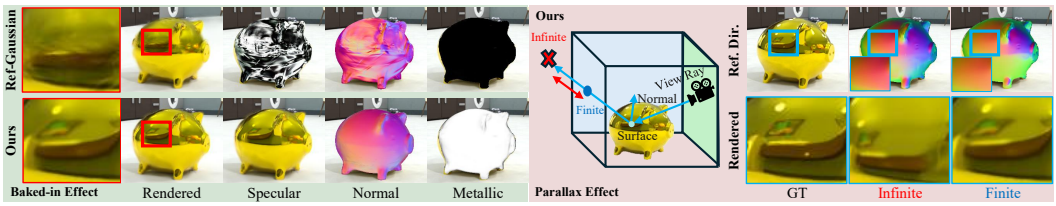

Figure 1: **Indoor reflection failure cases and our solution. Left:** Inaccurate reflective object localization causes specular collapse and baked-in appearance, which in turn corrupts the recovered normals. Our uncertainty-guided metallic supervision (UGMS) stabilizes diffuse–specular decomposition. **Right:** The infinite environment map assumption leads to reflection misalignment. Our parallax-corrected cubemaps model near-field lighting and improve reflection anchoring.

Supervision (UGMS). UGMS builds on our observation that reflective surfaces tend to exhibit elevated learned photometric uncertainty under the evaluated capture conditions. By converting uncertainty signals into metallic supervision, we prevent the optimizer from collapsing into diffuse-only solutions. This encourages more stable diffuse–specular decomposition and reduces the absorption of reflections into the diffuse component. To model finite distance indoor illumination, we further adopt parallax-corrected cubemaps (PCC) Sébastien & Zanuttini (2012) anchored by proxy geometry. By correcting the environment sampling direction based on spatial ray intersections, PCC enables parallax-aware lighting that reproduces near field reflection shifts across viewpoints. Finally, we present IndoorRef12, a new challenging benchmark for indoor reflective scenes specifically curated to feature intense near-field parallax and cluttered environments.

Our main contributions are summarized as follows.

- We explicitly address two fundamental challenges that are specific to highly specular indoor surfaces: unstable reflective object localization under multi-view ambiguity, and near-field reflection parallax under the infinite environment map assumption.
- We introduce two complementary modules: Uncertainty-Guided Metallic Supervision (UGMS) to stabilize diffuse–specular decomposition, and parallax-corrected cubemaps (PCC) to model finite distance indoor illumination in reflective scenes.
- We build IndoorRef12, a challenging indoor reflective benchmark with cluttered backgrounds and near field parallax, and show improved reflection fidelity and view consistency through extensive evaluations on both our dataset and GlossyReal Liu et al. (2023).

2 Related Work

Novel view synthesis. NVS renders unseen viewpoints from posed multi-view images. NeRF-based approaches represent scenes as implicit radiance fields optimized via volumetric rendering Mildenhall et al. (2020); Barron et al. (2022); Müller et al. (2022), while explicit point-based methods such as 3DGS enable competitive quality with real-time rasterization Kerbl et al. (2023). Recent extensions have improved robustness and rendering fidelity by addressing aliasing artifacts Huang et al. (2024); Ye et al. (2024). Yet specular reflections on metallic surfaces remain challenging in indoor scenes.

Reflection-aware rendering in 3DGS. Recent Gaussian-based methods extend 3DGS towards PBR-based formulations to estimate material and lighting properties. These include diffuse-specular decomposition under environment lighting Jiang et al. (2023); Ye et al. (2024); Yao et al. (2024); Tong et al. (2025); Tang et al. (2025); Zhou et al. (2025); Zhang et al. (2025a); Gao et al. (2024); Fan et al. (2025); Gu et al. (2025); Kouros et al. (2026); Xie et al. (2025) or learned specular features Zhang et al. (2025b); Fan et al. (2026); Shi et al. (2026). However, these methods are prone to baked-in artifacts under the mixed illumination of cluttered indoor scenes, and most query illumination by direction only, leaving finite-distance reflectors such as walls and furniture unmodeled. Closer to our localization problem, TransparentGS Huang et al. (2025) relies on text prompts (via GroundingDINO Liu et al. (2024)), which presume a known object category, and MaterialRefGS Zhang et al. (2025a) infers reflective regions from multi-view photometric variance via depth-based warping, which becomes circularly unreliable in highly specular regions where depth itself is ambiguous.

3 DATASET CONSTRUCTION: INDOORREF12

All data are synthetically generated using Blender 4.5.3 LTS Blender Online Community (2024), comprising twelve closed, cube-shaped indoor scenes that represent common residential spaces such as living rooms, kitchens, and laundry rooms. Unlike existing reflective benchmarks that capture isolated objects against a distant or empty background Verbin et al. (2022); Liu et al. (2023), our dataset places each reflective target inside a full room, where finite-distance walls and furniture induce noticeable parallax in specular reflections. Each scene contains highly reflective metallic objects such as gold, silver, and copper, and two scenes feature multiple reflective objects to evaluate multi-object generalization. For each scene, we render 200 viewpoints along a dome-like trajectory at 1024×1024 resolution, generating spatially aligned RGB images, world-space normal maps, and binary masks of reflective objects. While only RGB images are used for training, the aligned normals and masks enable rigorous quantitative evaluation of geometry and segmentation accuracy. Further details on scene composition, asset complexity, and rendering settings are provided in Appendix B.

4 METHOD

4.1 PRELIMINARIES

3DGS represents a scene as Gaussians $\{G_i\}_{i=1}^{K}$, parameterized by position μ_i, opacity α_i, and anisotropic scaling. For better surface alignment and stable normal estimation, we adopt 2DGS Huang et al. (2024), which flattens each Gaussian into a disk-like primitive by collapsing one scaling axis. Each is associated with physically based rendering (PBR) material attributes, including a surface normal $\mathbf{n}_i \in \mathbb{R}^3$, albedo $\mathbf{a}_i \in \mathbb{R}^3$, metallic $m_i \in [0, 1]$, and roughness $\rho_i \in [0, 1]$. Diffuse component is modeled with SH-based view independent appearance, and $\mathbf{a}_i$ modulates the specular term.

Differentiable splatting for map rendering. Given a target attribute ψ_i, we render its screen-space map via front-to-back alpha compositing.

$$\hat{\Psi} = \sum_{i=1}^{N} T_i \, w_i \, \psi_i, \quad T_i = \prod_{j=1}^{i-1} (1 - w_j), \quad w_i = \alpha_i \, \pi_i, \tag{1}$$

where T_i is the accumulated transmittance before the i-th element. N is the number of contributing Gaussian primitives for that pixel. w_i is its alpha-weight and π_i denotes the screen-space footprint of the i-th primitive. Given $\psi_i \in \{\mathbf{n}_i, \mathbf{a}_i, m_i, \rho_i\}$, we predict $\hat{\Psi}$, normal/albedo/metallic/roughness maps, respectively.

Deferred PBR shading (simplified Disney BRDF). We employ a deferred shading pipeline Ye et al. (2024) with a simplified Disney-style microfacet BRDF Burley & Studios (2012). Following prior works Yao et al. (2024); Zhang et al. (2025b;a), we model the diffuse appearance using SH and compute the specular term via environment lighting. For a viewing direction ω_o and surface normal $\mathbf{n}$, the outgoing color can be defined as

$$\hat{\mathbf{C}}(\omega_o) = (1 - m)\mathbf{c}_d + \mathbf{L}_s(\omega_o; \mathbf{a}, m, \rho, \mathbf{n}), \tag{2}$$

where $\mathbf{c}_d$ denotes the diffuse component represented with SH and $\mathbf{L}_s$ is the specular term computed by integrating incident illumination $\mathbf{L}_i(\omega_i)$ over the upper hemisphere.

$$\begin{aligned} \mathbf{L}_s(\omega_o) &= \int_{\Omega} \mathbf{L}_i(\omega_i) \; f_s(\omega_i, \omega_o) \, (\omega_i \cdot \mathbf{n}) \, d\omega_i, \\ f_s &= \frac{D \, F \, G}{4(\omega_i \cdot \mathbf{n})(\omega_o \cdot \mathbf{n})}, \end{aligned} \tag{3}$$

where D, F, and G are the microfacet distribution function, the Fresnel reflection coefficient, and the geometric attenuation or shadowing factor, respectively.

Directly calculating equation 3 is computationally expensive, so we adopt the split-sum approximation widely used in inverse rendering with environment lighting Yao et al. (2024).

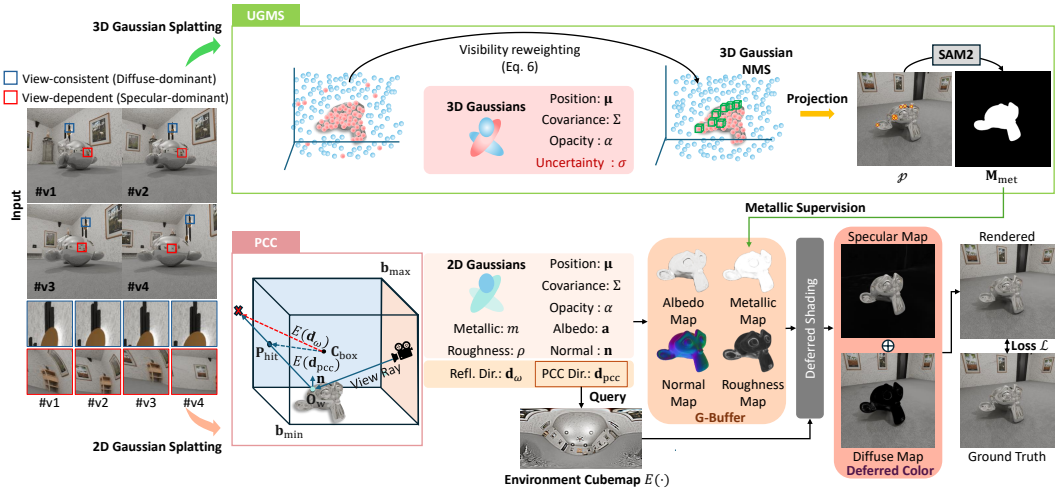

Figure 2: **Overview of our reflection-aware Gaussian splatting.** We enable physically grounded diffuse–specular decomposition and capture near-field parallax in indoor scenes. **UGMS** learns per-Gaussian uncertainty σ_i to localize reflective regions and suppress baked-in artifacts. **PCC** refines the environment query direction $\mathbf{d}_{\text{pcc}}$ via ray–proxy intersection to enable spatially correct reflection sampling under finite-distance illumination.

4.2 Uncertainty-Guided Metallic Supervision (UGMS)

Stable diffuse–specular decomposition requires accurate identification of reflective regions during training. As shown on the left of Fig. 2, however, in indoor scenes, the apparent color on reflective surfaces varies sharply with viewpoint, resulting in high photometric variance across views. To mitigate this issue, we leverage photometric uncertainty as a supervisory cue to generate pseudo metallic supervision. To capture persistent view-dependent photometric errors, we associate each Gaussian i with a learnable scalar uncertainty parameter $\sigma_i \geq 0$, shared across RGB channels. During rendering, these per-Gaussian uncertainties are accumulated along each ray to produce a per-pixel uncertainty map $U(\mathbf{p})$.

$$U(\mathbf{p}) = \sum_{i \in \mathcal{N}(\mathbf{p})} T_i w_i \sigma_i, \tag{4}$$

where $\mathcal{N}(\mathbf{p})$ is the set of Gaussians contributing to pixel $\mathbf{p}$.

The per-pixel uncertainty $U(\mathbf{p})$ is a differentiable aggregation of per-Gaussian uncertainty parameters σ_i. Therefore, we optimize them using a NLL loss.

$$\mathcal{L}_{\text{nll}}(\mathbf{p}) = \frac{1}{2}\left(\frac{\|\hat{\mathbf{C}}(\mathbf{p}) - \mathbf{C}(\mathbf{p})\|_2^2}{U(\mathbf{p})^2 + \epsilon} + \log(U(\mathbf{p})^2 + \epsilon)\right). \tag{5}$$

Minimizing Eq. (5) encourages higher uncertainty in regions with persistent photometric errors (i.e., specular reflections), making it an effective proxy for identifying metallic surfaces. We analyze why the learned uncertainty concentrates on metallic surfaces, with supporting statistics, in App. C.1. However, raw uncertainty estimates can be noisy, as spurious high uncertainty may arise from transient artifacts that appear in only a few views. To suppress such unreliable signals, we introduce a visibility-aware filtering strategy at the Gaussian level. For this purpose, we compute a visibility score v_i for each Gaussian i by counting the number of views where it is rendered, and define a filtered importance score s_i as follows.

$$s_i = \sigma_i \cdot \frac{v_i}{\max_k v_k}. \tag{6}$$

Based on the filtered scores s_i, we select the top-K candidates and apply greedy 3D non-maximum suppression (NMS) Hosang et al. (2017) with a fixed minimum separation to obtain a compact and spatially diverse set of 3D candidates $\mathcal{P}$ corresponding to reliable reflective regions.

For each training view, we project $\mathcal{P}$ onto the image plane to obtain 2D point prompts, which are provided to SAM2 Ravi et al. (2024). SAM2 produces a dense metallic guidance map $\mathbf{M}_{\text{met}} \in [0, 1]$ (see Fig. 3). We use this as a pseudo supervision for metallic prediction.

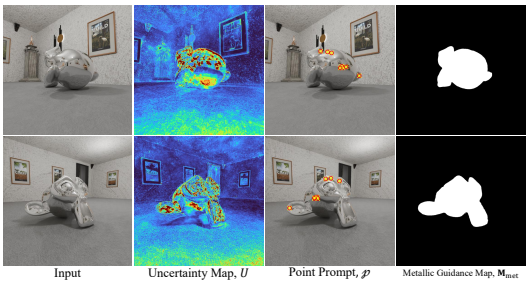

Figure 3: **UGMS examples.** Uncertainty-guided prompts produce metallic guidance maps.

Using equation 1, our model predicts a metallic map $\hat{\mathbf{M}}_{\text{met}} \in [0, 1]$. Since the pseudo metallic guidance $\mathbf{M}_{\text{met}}$ is derived from sparse point prompts, it may be incomplete. Therefore, we down-weight the supervision outside the mask and impose only weak background regularization.

$$\Lambda = \mathbf{M}_{\text{met}}(\mathbf{p}) + \lambda_{\text{bg}}(1 - \mathbf{M}_{\text{met}}(\mathbf{p})), \tag{7}$$

where $\lambda_{\text{bg}} = 0.05$ to control the strength of background.

Finally, the loss for metallic guidance map prediction $\mathcal{L}_{\text{met}}$ can be defined as

$$\mathcal{L}_{\text{met}} = \left\| \Lambda \odot \left(\hat{\mathbf{M}}_{\text{met}}(\mathbf{p}) - \mathbf{M}_{\text{met}}(\mathbf{p}) \right) \right\|_1. \tag{8}$$

4.3 Indoor Illumination Modeling with Parallax-Corrected Cubemaps (PCC)

To model indoor illumination more faithfully, we replace the standard infinite distance environment map with parallax-corrected cubemaps (PCC). Unlike conventional environment mapping that depends only on the reflection direction, PCC accounts for finite-distance structures by using a proxy box to obtain a parallax-aware sampling direction (Fig. 2).

A key challenge in indoor captures is that the world coordinate frame reconstructed by COLMAP Schonberger & Frahm (2016) is arbitrary. As a result, a proxy box defined in the world frame can be poorly aligned with the dominant room layout, leading to unstable ray–proxy intersections.

To obtain a robust proxy box, we estimate a rotation matrix $\mathbf{R}_{\text{box}}$ via PCA on the reconstructed point cloud Gottschalk et al. (1996) (Appendix A), aligning the box axes with the dominant scene directions. We set the box center as the point-cloud centroid $\mathbf{C}_{\text{box}} = \frac{1}{N}\sum_i \mathbf{x}_i$ and compute local coordinates $\mathbf{x}_{\ell,i} = \mathbf{R}_{\text{box}}^{\top}(\mathbf{x}_i - \mathbf{C}_{\text{box}})$. The axis-aligned bounds are then given by $\mathbf{b}_{\min} = \min_i \mathbf{x}_{\ell,i}$ and $\mathbf{b}_{\max} = \max_i \mathbf{x}_{\ell,i}$, defining a geometry-aligned proxy box that encloses the scene.

Given a surface point with ray origin $\mathbf{O}_{\text{w}}$ and reflected direction $\mathbf{d}_\omega$, PCC finds the exit point $\mathbf{P}_{\text{hit}}$ where the ray first intersects the proxy box boundary along the forward direction. Since the proxy box is constructed to enclose the cameras and scene geometry in our indoor captures, the ray typically originates inside the box, i.e., $\mathbf{O}_{\text{w}} + t\,\mathbf{d}_\omega, t \geq 0$.

To simplify ray–box intersection Kay & Kajiya (1986), we transform the ray into the box-aligned local frame.

$$\mathbf{O}_\ell = \mathbf{R}_{\text{box}}^{\top}(\mathbf{O}_{\text{w}} - \mathbf{C}_{\text{box}}), \qquad \mathbf{d}_\ell = \mathbf{R}_{\text{box}}^{\top}\mathbf{d}_\omega. \tag{9}$$

In this frame, the box is axis-aligned with bounds $[\mathbf{b}_{\min}, \mathbf{b}_{\max}]$. For each axis $k \in \{x, y, z\}$, the ray intersects the two boundary planes at t_1^k and t_2^k.

$$t_1^k = \frac{\mathbf{b}_{\min}^k - \mathbf{O}_\ell^k}{\mathbf{d}_\ell^k + \epsilon}, \qquad t_2^k = \frac{\mathbf{b}_{\max}^k - \mathbf{O}_\ell^k}{\mathbf{d}_\ell^k + \epsilon}. \tag{10}$$

Because the ray starts inside the box, t_1^k or t_2^k lies behind the origin and the other lies in front. We take the forward exit time along axis k as

$$t_{\text{far}}^k = \max(t_1^k, t_2^k). \tag{11}$$

The ray exits the box at the closest of these axial exit planes, i.e., the face reached first along the forward ray direction.

$$t_{\text{hit}} = \min_k t_{\text{far}}^k, \qquad \mathbf{P}_{\text{hit}} = \mathbf{O}_{\text{w}} + t_{\text{hit}}\,\mathbf{d}_\omega, \tag{12}$$

Table 1: Per-scene image quality comparison on synthesized test views on IndoorRef12 dataset.

Method	*Spartan*			*Rabbit*			*Chick*			*Vase*			*Beethoven*			*Suzanne & Utah teapot*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS-DR	27.27	0.948	0.337	30.73	0.935	0.309	30.66	0.963	0.112	31.54	0.966	0.103	25.64	0.901	0.142	28.31	0.939	0.171
Ref-GS	29.58	0.931	0.351	29.67	0.931	0.305	30.77	0.967	0.110	31.60	0.964	0.105	22.41	0.869	0.177	33.55	0.974	0.061
Ref-Gaussian	28.95	0.947	0.339	29.87	0.921	0.323	28.92	0.957	0.138	18.51	0.899	0.221	18.84	0.852	0.243	24.12	0.903	0.219
MaterialRefGS	29.75	0.948	0.331	31.79	0.946	0.318	29.25	0.960	0.128	30.63	0.966	0.112	26.35	0.908	0.145	33.58	0.972	0.076
MGSR	33.84	0.963	**0.293**	28.55	0.934	0.306	31.53	0.970	0.094	31.55	0.964	0.101	25.93	0.900	0.146	34.30	0.973	0.059
Ref-DGS	34.54	0.950	0.300	30.37	0.934	0.308	32.05	0.971	0.103	31.27	0.964	0.109	26.01	0.899	0.139	35.15	0.941	0.094
SpecGloss-GS	31.61	0.930	0.385	31.48	0.947	0.301	30.73	0.966	0.120	31.52	0.966	0.107	26.69	0.898	0.138	34.23	0.975	0.062
Ours	**34.79**	**0.965**	0.298	**32.53**	**0.950**	**0.294**	**34.45**	**0.978**	**0.089**	**34.86**	**0.976**	**0.096**	**27.11**	**0.910**	**0.134**	**35.36**	**0.977**	**0.058**

Method	*Apple*			*Skull*			*Pumpkin*			*Pig*			*Knight*			*Rabbit & Chick*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS-DR	33.97	0.967	0.067	34.39	0.955	0.265	30.03	0.952	0.079	32.90	0.967	0.219	32.98	0.948	0.321	29.85	0.943	0.378
Ref-GS	34.14	0.971	0.058	35.20	0.961	**0.235**	30.22	0.954	0.080	32.74	0.970	0.209	33.19	0.952	**0.307**	34.53	0.960	0.358
Ref-Gaussian	25.37	0.925	0.165	20.46	0.859	0.392	23.11	0.908	0.166	25.92	0.953	0.247	22.95	0.895	0.406	23.46	0.901	0.452
MaterialRefGS	32.15	0.964	0.067	34.10	0.956	0.293	29.32	0.952	0.091	33.91	0.972	0.214	31.07	0.936	0.333	33.35	0.957	0.339
MGSR	33.55	0.968	0.067	34.01	0.962	0.255	29.97	0.952	0.080	31.34	0.970	0.210	32.44	0.948	0.316	33.07	0.957	0.303
Ref-DGS	34.64	0.960	0.058	35.32	0.958	0.252	30.91	0.955	0.079	34.66	0.968	0.207	33.47	0.954	0.311	34.24	0.964	0.307
SpecGloss-GS	33.12	0.966	0.076	32.71	0.957	0.274	29.87	0.952	0.087	33.51	0.971	0.214	30.18	0.933	0.345	32.78	0.954	0.336
Ours	**34.85**	**0.973**	**0.056**	**35.83**	**0.969**	0.250	**31.02**	**0.957**	**0.072**	**35.31**	**0.976**	**0.205**	**34.01**	**0.956**	0.310	**34.65**	**0.966**	**0.251**

Table 2: Per-scene image quality comparison on synthesized test views on GlossyReal dataset.

Method	*Bear*			*Bunny*			*Coral*			*Maneki*			*Vase*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS-DR	19.19	0.710	0.283	24.54	0.856	0.211	19.02	0.741	0.275	21.85	0.832	0.238	21.68	0.790	0.269
Ref-GS	19.03	0.729	0.274	25.72	0.880	0.195	18.70	0.760	0.249	24.14	0.860	0.193	24.01	0.831	0.188
Ref-Gaussian	20.32	0.742	0.258	24.90	0.860	0.209	20.30	0.761	0.246	24.04	0.849	0.227	23.23	0.829	0.231
MaterialRefGS	19.78	0.731	0.268	25.13	0.866	0.198	19.92	0.768	0.252	23.71	0.852	0.215	23.45	0.815	0.221
MGSR	19.93	0.722	0.275	24.26	0.858	0.215	19.61	0.759	0.265	20.48	0.801	0.250	22.65	0.792	0.240
Ref-DGS	20.13	0.750	0.249	25.83	0.885	0.165	20.13	0.780	0.224	24.59	0.867	0.195	24.55	0.839	0.190
SpecGloss-GS	19.97	0.754	0.249	24.79	0.856	0.215	20.06	0.778	0.243	23.45	0.840	0.236	23.02	0.806	0.227
Ours	**21.04**	**0.779**	**0.247**	**25.98**	**0.891**	**0.162**	**20.83**	**0.803**	**0.219**	**24.84**	**0.879**	**0.190**	**24.80**	**0.848**	**0.186**

where $\mathbf{P}_{\text{hit}}$ denotes the 3D point where the reflected ray first intersects the proxy box boundary, and serves as the reference point for parallax-aware environment sampling.

Then PCC query direction $\mathbf{d}_{\text{pcc}}$ can be defined as

$$\mathbf{d}_{\text{pcc}} = \frac{\mathbf{P}_{\text{hit}} - \mathbf{C}_{\text{box}}}{\|\mathbf{P}_{\text{hit}} - \mathbf{C}_{\text{box}}\|_2 + \epsilon}. \tag{13}$$

We query the environment cubemap $E(\cdot)$ using $\mathbf{d}_{\text{pcc}}$ instead of $\mathbf{d}_{\omega}$. By linking the sampling direction to the ray–proxy intersection point $\mathbf{P}_{\text{hit}}$, PCC captures near-field reflection parallax and improves reflection anchoring under finite-distance indoor illumination (see Fig. 7, bottom).

Finally, we optimize our model for indoor scene using a photometric reconstruction loss and metallic supervision. The total loss $\mathcal{L}$ can be defined as $\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{met}}\mathcal{L}_{\text{met}}$.

With equation 2 and equation 3, $\mathcal{L}_{\text{photo}}$ can be defined as

$$\begin{aligned}\mathcal{L}_{\text{photo}} = \|\hat{\mathbf{C}}(\mathbf{p}) - \mathbf{C}(\mathbf{p})\|_1 \\ + \lambda_{\text{ssim}}\Big(1 - \text{SSIM}\big(\hat{\mathbf{C}}(\mathbf{p}), \mathbf{C}(\mathbf{p})\big)\Big),\end{aligned} \tag{14}$$

where $\lambda_{\text{ssim}} = 0.2$.

5 Results and Evaluation

Datasets. We evaluate our method on both synthetic and real-world benchmarks. For synthetic data, we use our IndoorRef12 with 12 highly reflective metallic objects in diverse indoor room configurations. For real-world data, we use GlossyReal Liu et al. (2023), containing five reflective object scenes with 100–130 multi-view images at 1024×768 resolution.

Metrics. For rendering quality, we report the standard PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018) metrics. For geometric fidelity, particularly on reflective surfaces, we measure normal prediction accuracy using Mean Angular Error (MAE) and normal LPIPS.

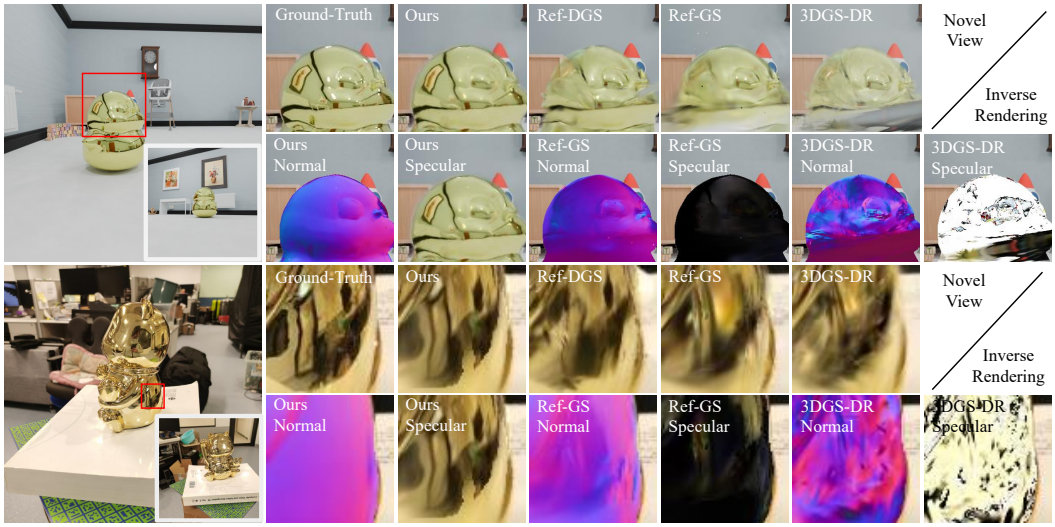

Figure 4: Qualitative comparison of novel-view synthesis and inverse rendering results.

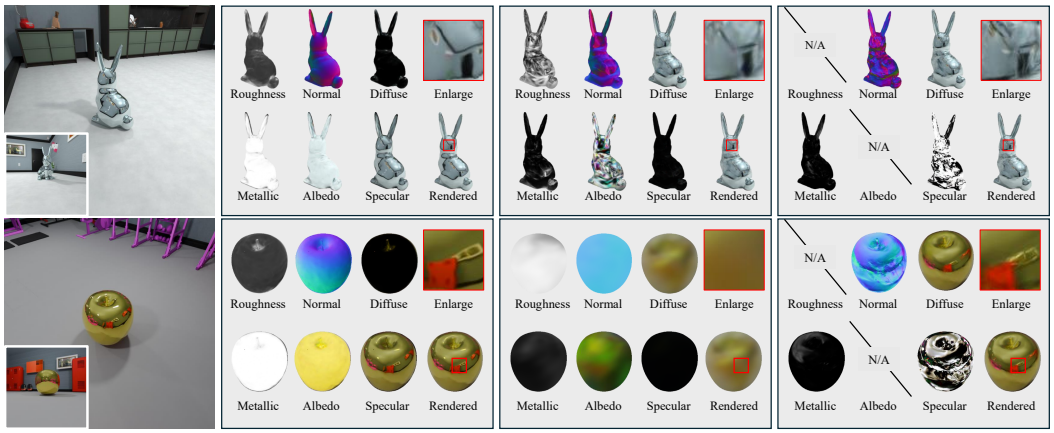

Figure 5: Qualitative comparison of illumination decomposition results on IndoorRef12.

Table 3: Normal reconstruction quality comparisons on IndoorRef12 dataset.

	Ref-Gaussian	Ref-GS	SpecGloss-GS	Ref-DGS	Ours
MAE° ↓	40.98	33.90	23.51	<u>19.25</u>	**10.92**
LPIPS ↓	0.599	0.481	0.356	<u>0.304</u>	**0.221**

Table 4: Ablation studies on the components of our model.

Components		IndoorRef12			GlossyReal		
UGMS	PCC	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
		24.29	0.911	0.264	22.56	0.808	0.234
✓		33.12	0.949	0.189	23.16	0.829	0.210
✓	✓	**33.72**	**0.963**	**0.176**	**23.50**	**0.840**	**0.201**

Baselines. We compare against reflection-aware gaussian splatting baselines such as 3DGS-DR Ye et al. (2024), Ref-GS Zhang et al. (2025b), Ref-Gaussian Yao et al. (2024), MaterialRefGS Zhang et al. (2025a), MGSR Zhou et al. (2025), Ref-DGS Fan et al. (2026), and SpecGloss-GS Kouros et al. (2026). These methods represent strong baselines for both general-purpose NVS and reflection-aware reconstruction in challenging reflective scenes.

Comparisons on IndoorRef12 Dataset. On IndoorRef12 (Tab. 1), our method achieves the best average performance across all metrics, with the gap further amplified on reflective regions, confirming the effectiveness of stabilizing reflective-region localization and modeling near-field reflection parallax. The first row of Fig. 4 shows consistent improvements in rendering fidelity with reduced baked-in artifacts, sharper normals, and more accurate specular appearance.

Comparisons on GlossyReal Dataset. On the real-world GlossyReal benchmark Liu et al. (2023) (Tab. 2), our method again achieves the best average performance, generalizing beyond synthetic data

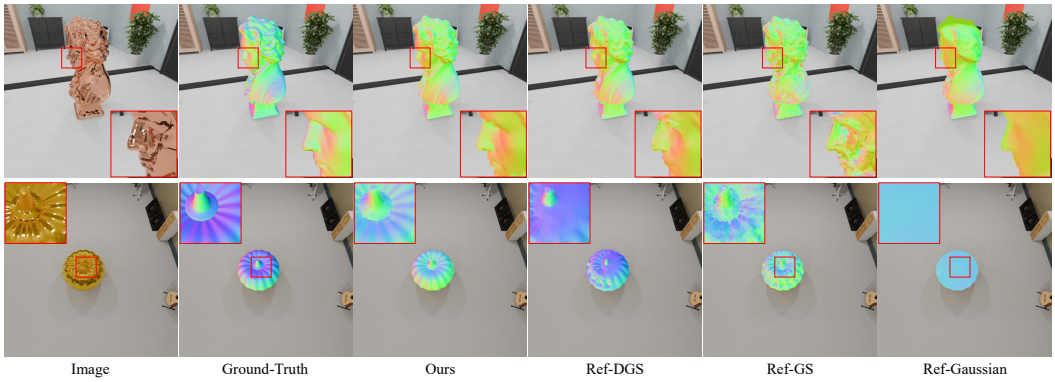

Figure 6: Qualitative comparison of normal produced by different methods.

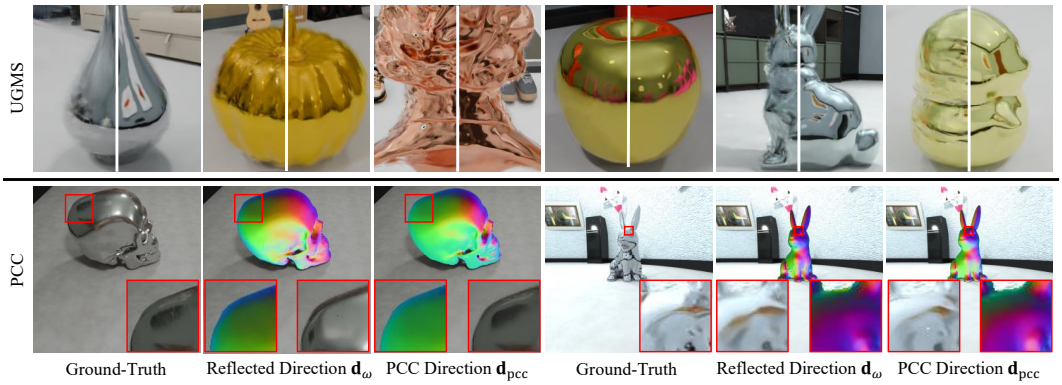

Figure 7: **Ablation of UGMS and PCC.** Top: without (left) and with (right) UGMS. Bottom: reflected direction versus PCC direction on the Skull scene and the dome-shaped Rabbit scene.

Table 5: Results on irregular layouts.

Layout	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Near-field	Ref-Gaussian	30.31	0.952	0.131
	MaterialRefGS	28.76	0.941	0.140
	Ref-DGS	33.11	0.962	0.115
	Ours	**33.91**	**0.966**	**0.109**
Dome	w/o PCC	34.15	0.959	0.141
	w/ PCC	**34.57**	**0.969**	**0.138**

Table 6: Mask quality and rendering. All rows except MaterialRefGS use our backbone, varying only the mask source.

Method	IoU ↑	PSNR ↑	SSIM ↑	LPIPS ↓
SAM2 (TopK)	0.039	24.38	0.913	0.263
Marigold	0.263	32.09	0.957	0.195
MaterialRefGS	0.040	31.27	0.953	0.204
Ours	**0.898**	**33.72**	**0.963**	**0.176**

to complex illumination and highly reflective materials. The second row of Fig. 4 confirms cleaner decomposition and better-anchored reflections across viewpoints.

Comparisons of Illumination Decomposition. We compare with inverse-rendering-based 3DGS baselines such as Ref-Gaussian Yao et al. (2024) and 3DGS-DR Ye et al. (2024). As shown in Fig. 5, our method produces substantially more reliable estimates of albedo, roughness, metallicity, and normals, avoiding common failure modes where specular reflections leak into the diffuse component.

Normal Estimation. We compare normal predictions on IndoorRef12 in Tab. 3, where our method achieves significantly better accuracy than reflection-aware baselines. Fig. 6 further shows closer alignment to the ground-truth normals, especially on reflective surfaces and fine details.

Ablation Study. As shown in Tab. 4, UGMS yields a large gain over the PBR-based 2DGS baseline, and PCC adds further improvement. Without UGMS, reflections become blurred due to baked-in specular appearance (Fig. 7, top). PCC improves reflection anchoring on both the standard Skull scene and the dome-shaped Rabbit scene, where the enclosure deviates from the cuboid proxy, indicating that PCC remains robust to proxy mismatch (Fig. 7, bottom, Tab. 5, App. I.2).

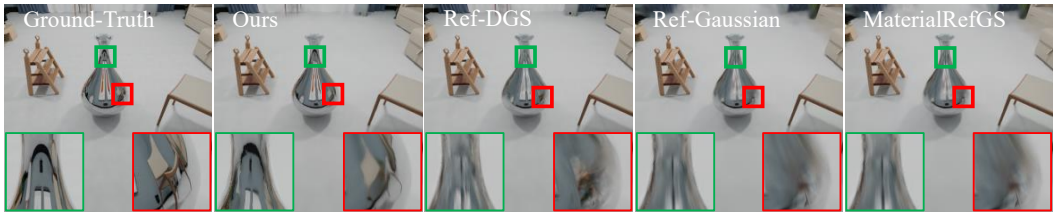

Figure 8: Near-field reflection comparison with adjacent furniture.

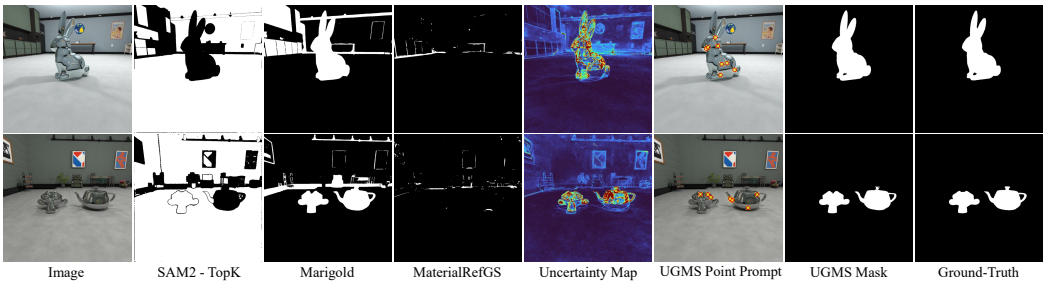

Figure 9: Qualitative comparison of metallic mask generation strategies.

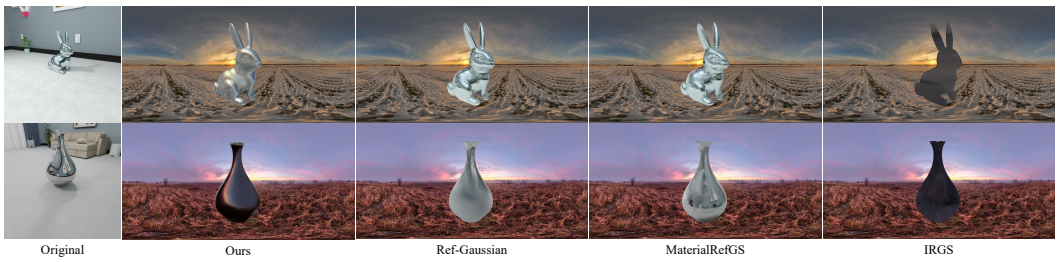

Figure 10: Relighting results under novel illumination.

Near-field reflection. With furniture placed directly in front of the metallic object (Fig. 8, Tab. 5, App. I.2), ray-tracing-based methods Yao et al. (2024); Zhang et al. (2025a) blur the reflected furniture (red boxes) and bake it into the diffuse term (green boxes), and Ref-DGS Fan et al. (2026) blurs it into an unstructured smear, losing detail. Ours recovers sharp, correctly placed reflections and achieves the highest PSNR (33.91 versus 33.11 dB for Ref-DGS), as PCC anchors near-field content by correcting the query direction without extra primitives or ray tracing.

Reflective Object Mask Quality. We compare UGMS with alternative mask sources in Tab. 6 and Fig. 9: SAM2-TopK, Marigold Ke et al. (2025) (material-based proxy), and MaterialRefGS Zhang et al. (2025a) (native architecture, included for reference). UGMS achieves both the highest mask IoU and the best downstream rendering quality, quantitatively and qualitatively. On partially metallic objects, where the guidance mask covers the whole object, UGMS still improves rendering by 7.7 to 9.2 dB over the unguided model and stays within 0.1 dB of a material-level oracle (App. C.2).

Relighting Comparison. In Fig. 10, Ref-Gaussian Yao et al. (2024) and MaterialRefGS Zhang et al. (2025a) remain unchanged in specular appearance under novel HDR illumination, as their specular components are baked into the diffuse term, and IRGS Gu et al. (2025) yields severely degraded renderings with missing highlights. In contrast, ours produces clear color shifts and reflection changes that respond to the illumination, evidencing successful decomposition.

6 Conclusion

We presented a reflection-aware GS framework for object-centric reflective scenes in indoor environments, where UGMS stabilizes diffuse–specular decomposition against baked-in artifacts and PCC models finite-distance illumination without ray tracing. Together with the IndoorRef12 benchmark, they yield consistent gains in reflection fidelity and geometry on both synthetic and real scenes.

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools to assist with translating and polishing the manuscript in English. We have not used generative AI tools to generate synthetic datasets, formulate or prove mathematical claims, implement methods, or propose the core research hypotheses, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. All numerical results were produced by our own experiments and verified by the authors, and all references were checked against their original publications. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF CVPR*, pp. 5470–5479, 2022.

Blender Online Community. Blender - a 3d modelling and rendering package, 2024. URL `http://www.blender.org`.

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

Jiahui Fan, Fujun Luan, Jian Yang, Milos Hasan, and Beibei Wang. Rng: Relightable neural gaussians. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26525–26534, 2025.

Ningjing Fan, Yiqun Wang, Dong-Ming Yan, and Peter Wonka. Ref-dgs: Reflective dual gaussian splatting. *arXiv preprint arXiv:2603.07664*, 2026.

Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pp. 73–89. Springer, 2024.

Stefan Gottschalk, Ming C Lin, and Dinesh Manocha. Obbtree: A hierarchical structure for rapid interference detection. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 171–180, 1996.

Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, and Li Zhang. Irgs: Inter-reflective gaussian splatting with 2d gaussian ray tracing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10943–10952, 2025.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4507–4515, 2017.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Letian Huang, Dongwei Ye, Jialin Dan, Chengzhi Tao, Huiwen Liu, Kun Zhou, Bo Ren, Yuanqi Li, Yanwen Guo, and Jie Guo. Transparentgs: Fast inverse rendering of transparent objects with gaussians. *ACM Transactions on Graphics (TOG)*, 44(4):1–17, 2025.

Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces, 2023.

Timothy L Kay and James T Kajiya. Ray tracing complex scenes. *ACM SIGGRAPH computer graphics*, 20(4):269–278, 1986.

Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Georgios Kouros, Minye Wu, and Tinne Tuytelaars. Spec-gloss surfels and normal-diffuse priors for relightable glossy objects. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4417–4426. IEEE, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *arXiv preprint arXiv:2305.17398*, 2023.

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. In *Proceedings of ECCV*, pp. 405–421, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):1–15, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Lagarde Sébastien and Antoine Zanuttini. Local image-based lighting with parallax-corrected cubemaps. In *ACM SIGGRAPH 2012 Talks*, pp. 1–1. 2012.

Ji Shi, Xianghua Ying, Bowei Xing, Ruohao Guo, and Wenzhen Yue. Rt-splatting: Joint reflection-transmission modeling with gaussian splatting. *arXiv preprint arXiv:2605.18263*, 2026.

Jiajun Tang, Fan Fei, Zhihao Li, Xiao Tang, Shiyong Liu, Youyu Chen, Binxiao Huang, Zhenyu Chen, Xiaofei Wu, and Boxin Shi. Spectre-gs: Modeling highly specular surfaces with reflected nearby objects by tracing rays in 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16133–16142, 2025.

Jinguang Tong, Xuesong Li, Fahira Afzal Maken, Sundaram Muthu, Lars Petersson, Chuong Nguyen, and Hongdong Li. Gs-2dgs: Geometrically supervised 2dgs for reflective object reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21547–21557, 2025.

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF CVPR*, pp. 5481–5490. IEEE, 2022.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Sean Wu, Shamik Basu, Tim Broedermann, Luc Van Gool, and Christos Sakaridis. Pbr-nerf: Inverse rendering with physics-based neural fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10974–10984, 2025.

Tao Xie, Xi Chen, Zhen Xu, Yiman Xie, Yudong Jin, Yujun Shen, Sida Peng, Hujun Bao, and Xiaowei Zhou. Envgs: Modeling view-dependent appearance with environment gaussian. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5742–5751, 2025.

Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. Relightable and animatable neural avatar from sparse-view video. *arXiv preprint arXiv:2308.07903*, 2023.

Ziyi Yang, Xinyu Gao, Yang-Tian Sun, Yihua Huang, Xiaoyang Lyu, Wen Zhou, Shaohui Jiao, Xiaojuan Qi, and Xiaogang Jin. Spec-gaussian: Anisotropic view-dependent appearance for 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:61192–61216, 2024.

Yuxuan Yao, Zixuan Zeng, Chun Gu, Xiatian Zhu, and Li Zhang. Reflective gaussian splatting. *arXiv preprint arXiv:2412.19282*, 2024.

Keyang Ye, Qiming Hou, and Kun Zhou. 3d gaussian splatting with deferred reflection. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–10, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Wenyuan Zhang, Jimin Tang, Weiqi Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. Materialrefgs: Reflective gaussian splatting with multi-view consistent material inference. *arXiv preprint arXiv:2510.11387*, 2025a.

Youjia Zhang, Anpei Chen, Yumin Wan, Zikai Song, Junqing Yu, Yawei Luo, and Wei Yang. Ref-gs: Directional factorization for 2d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26483–26492, 2025b.

Qingyuan Zhou, Yuehu Gong, Weidong Yang, Jiaze Li, Yeqi Luo, Baixin Xu, Shuhao Li, Ben Fei, and Ying He. Mgsr: 2d/3d mutual-boosted gaussian splatting for high-fidelity surface reconstruction under various light conditions. *arXiv preprint arXiv:2503.05182*, 2025.

A PCA-BASED OBB INITIALIZATION

This section describes how we initialize the proxy oriented bounding box (OBB) used in parallax-corrected cubemaps (PCC). Given a reconstructed point cloud $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ in the COLMAP world coordinate system, we estimate the box rotation $\mathbf{R}_{\text{box}}$ via PCA and derive axis-aligned bounds $[\mathbf{b}_{\min}, \mathbf{b}_{\max}]$ in the PCA-aligned local frame.

A.1 PCA ROTATION MATRIX

We first compute the point-cloud centroid (box center)

$$\mathbf{C}_{\text{box}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \tag{S1}$$

and center the points as $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{C}_{\text{box}}$. We then form the covariance matrix

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^{\top} \in \mathbb{R}^{3\times 3}. \tag{S2}$$

Let $(\lambda_k, \mathbf{e}_k)$ denote the eigenpairs of Σ with $\lambda_1 \geq \lambda_2 \geq \lambda_3$. We construct the PCA rotation as

$$\mathbf{R}_{\text{box}} = [\mathbf{e}_1\ \mathbf{e}_2\ \text{sign}(\det[\mathbf{e}_1\ \mathbf{e}_2\ \mathbf{e}_3]) \cdot \mathbf{e}_3] \in SO(3), \tag{S3}$$

where the sign correction on $\mathbf{e}_3$ ensures $\det(\mathbf{R}_{\text{box}}) = +1$, yielding a proper right-handed rotation matrix.

A.2 BOUNDS IN THE PCA-ALIGNED FRAME

Using $\mathbf{R}_{\text{box}}$, we transform points into the box-aligned local coordinates:

$$\mathbf{x}_{\ell,i} = \mathbf{R}_{\text{box}}^{\top}(\mathbf{x}_i - \mathbf{C}_{\text{box}}) \in \mathbb{R}^3. \tag{S4}$$

We then compute axis-aligned bounds (AABB) in this local frame:

$$\mathbf{b}_{\min} = \min_i\ \mathbf{x}_{\ell,i}, \qquad \mathbf{b}_{\max} = \max_i\ \mathbf{x}_{\ell,i}, \tag{S5}$$

where $\min\,/\,\max$ are taken element-wise over (x, y, z). This defines an OBB in the world frame with center $\mathbf{C}_{\text{box}}$ and orientation $\mathbf{R}_{\text{box}}$.

A.3 RAY TRANSFORM FOR PCC

For a world-space ray origin $\mathbf{O}_{\text{w}}$ and reflected direction $\mathbf{d}_\omega$, we transform the ray into the local box coordinates as

$$\mathbf{O}_\ell = \mathbf{R}_{\text{box}}^{\top}(\mathbf{O}_{\text{w}} - \mathbf{C}_{\text{box}}), \qquad \mathbf{d}_\ell = \mathbf{R}_{\text{box}}^{\top}\mathbf{d}_\omega. \tag{S6}$$

This reduces ray Oriented Bounding Box (OBB) intersection to a standard ray Axis-Aligned Bounding Box (AABB) intersection with bounds $[\mathbf{b}_{\min}, \mathbf{b}_{\max}]$ in the local frame.

B INDOORREF12 DATASET: ADDITIONAL DETAILS

This section provides additional details on the construction of the IndoorRef12 dataset, including scene constraints, viewpoint sampling, rendering outputs, and dataset statistics. IndoorRef12 will be released publicly with the paper.

B.1 SCENE DESIGN AND CONSTRAINTS

All scenes are synthetically generated in Blender (v4.5.3 LTS) as closed, cube-shaped indoor environments representing common residential spaces (e.g., living rooms, kitchens, laundry rooms). The dataset contains twelve distinct room layouts, comprising ten single-object scenes (D1–D10) and two multi-object scenes (D11–D12) for evaluating multi-object generalization.

Table S1: IndoorRef12 scene distribution plan used for controlled diversity across rooms. D11–D12 contain multiple reflective objects to evaluate generalization to multi-object scenes.

ID	Scene concept	Reflective object	Metal	BG detail	#Lights	Surface
D1	Kitchen	*Rabbit*	Chrome	Low	4	Simple
D2	Baby room	*Chick*	Gold	High	8	Simple
D3	Living room	*Vase*	Silver	Low	8	Simple
D4	Laundry room	*Pig*	Gold	High	4	Simple
D5	Practice room	*Pumpkin*	Gold	High	8	Complex
D6	Study	*Skull*	Silver	Low	4	Complex
D7	Porch	*Beethoven*	Copper	Low	8	Complex
D8	Bathroom	*Spartan*	Silver	Low	8	Complex
D9	Workshop	*Knight*	Copper	High	4	Complex
D10	Home gym	*Apple*	Gold	High	4	Simple
D11	Workshop	*Suzanne & Utah teapot*	Silver	High	4	Simple
D12	Practice room	*Rabbit & Chick*	Silver	High	8	Simple

Table S2: **Placement of reflective targets in IndoorRef12.** r_{wall} and r_{furn} denote the normalized distance from the target to the nearest wall and to the nearest large furniture, respectively. For multi-object scenes, distances are measured from the listed object. Normalized by the half-width of the room along the corresponding axis.

Scene	Target	r_{wall}	r_{furn}
D1 Kitchen	Rabbit	0.31	0.27
D2 Baby room	Chick	0.48	0.46
D3 Living room	Vase	0.35	0.29
D4 Laundry room	Pig	0.52	0.31
D5 Practice room	Pumpkin	0.55	0.44
D6 Study	Skull	0.45	0.63
D7 Porch	Beethoven	0.53	0.37
D8 Bathroom	Spartan	0.48	0.22
D9 Workshop	Knight	0.40	0.31
D10 Home gym	Apple	0.54	0.39
D11 Workshop	Suzanne	0.42	0.10
D12 Practice room	Rabbit	0.53	0.40
Mean		0.46	0.35
Scenes with furniture closer than wall		11 / 12	

Each single-object scene includes one reflective target object as the primary source of specular reflections, while multi-object scenes contain two reflective objects placed in close proximity to induce mutual reflection. Reflective targets are placed away from the room center and near surrounding structures, so that walls and large furniture induce near-field reflection parallax under indoor lighting. Tab. S2 reports the per-scene placement. The normalized distance to the nearest wall ranges from 0.31 to 0.55 (mean 0.46), and the normalized distance to the nearest large furniture ranges from 0.10 to 0.63 (mean 0.35). In 11 of the 12 scenes, large furniture is closer to the target than the nearest wall, making furniture the dominant near-field reflector. All other assets (furniture, wall decorations, props) are constrained to be non-reflective and non-transparent to avoid secondary reflection paths.

B.2 Material Rules

Reflective targets are assigned highly reflective metallic materials (e.g., gold, silver, copper) to induce view-dependent appearance. To isolate reflection behavior, we exclude transparent objects and other reflective/secondary specular surfaces (e.g., mirrors or glass) in the rest of the scene.

Floor as a non-reflective control surface. The floor is designed to be nearly diffuse to prevent it from becoming an additional reflector. We allow realistic texture, but suppress strong specular response and avoid high-frequency patterns (e.g., visible grout/tile lines).

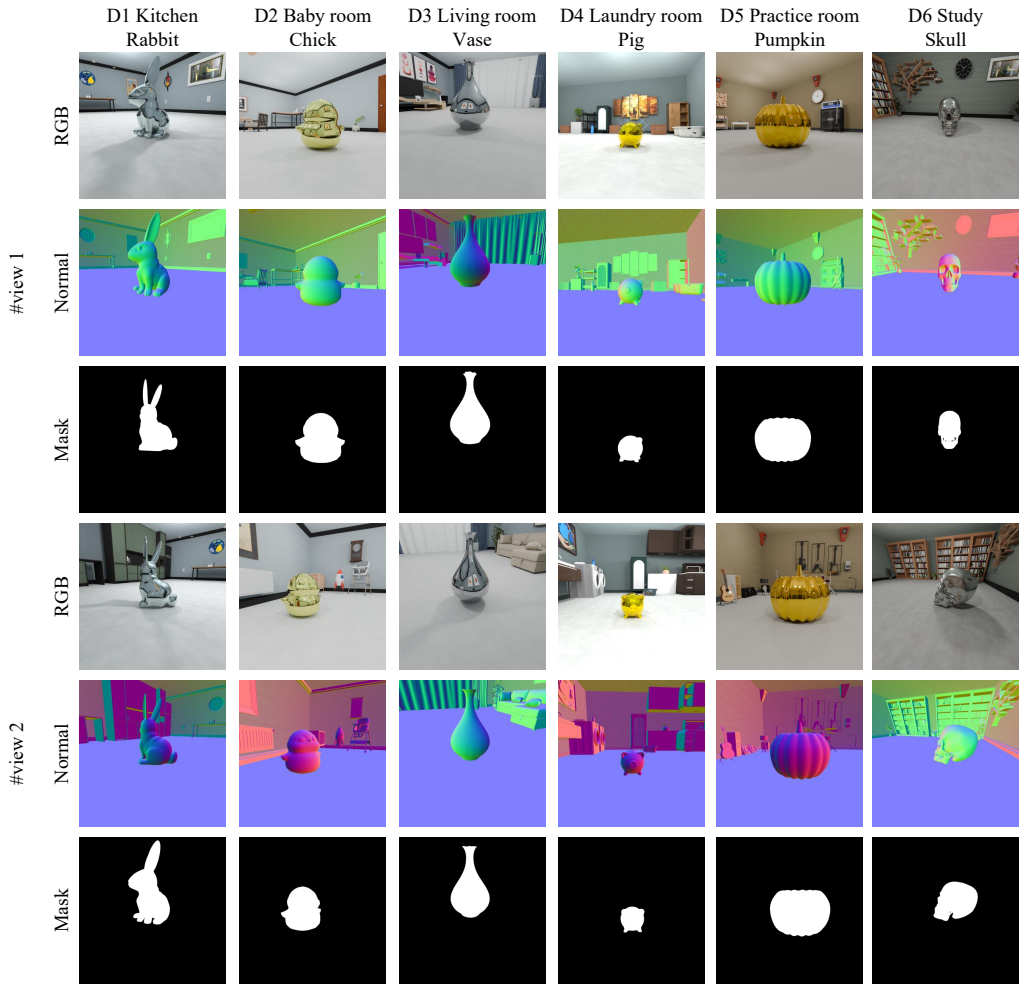

Figure S1: IndoorRef12 dataset examples (Rooms D1–D6). For each room, we show two viewpoints with the corresponding RGB image, world-space normal map, and object mask.

B.3 Lighting Setup

Scenes use multiple small point lights (bulb-like point sources). Lighting is treated as a controlled variable because it directly affects specular appearance. We keep light layout and overall intensity as consistent as possible across scenes, allowing only minor concept-driven adjustments. We use 4–8 lights per scene (some scenes with 4 and others with 8 for diversity), and fix a single color temperature per scene (warm 3000–4000K or neutral 4500–5500K).

B.4 Viewpoint Sampling

For each room, we render 200 views along a predefined circular (dome-like) trajectory centered on the reflective object(s). Camera intrinsics are fixed across all views; only camera pose varies along the path. No per-view optimization or parameter tuning is performed.

B.5 Rendered Modalities and File Formats

At each viewpoint, we export three spatially aligned modalities via Blender Render Layers and compositing: (i) RGB images, (ii) normal maps, and (iii) binary masks for the reflective object(s). All images are rendered at 1024×1024 resolution. RGB images are stored as 16-bit PNGs in sRGB color space. Normal maps are stored as full-precision floating-point EXR files in non-color space

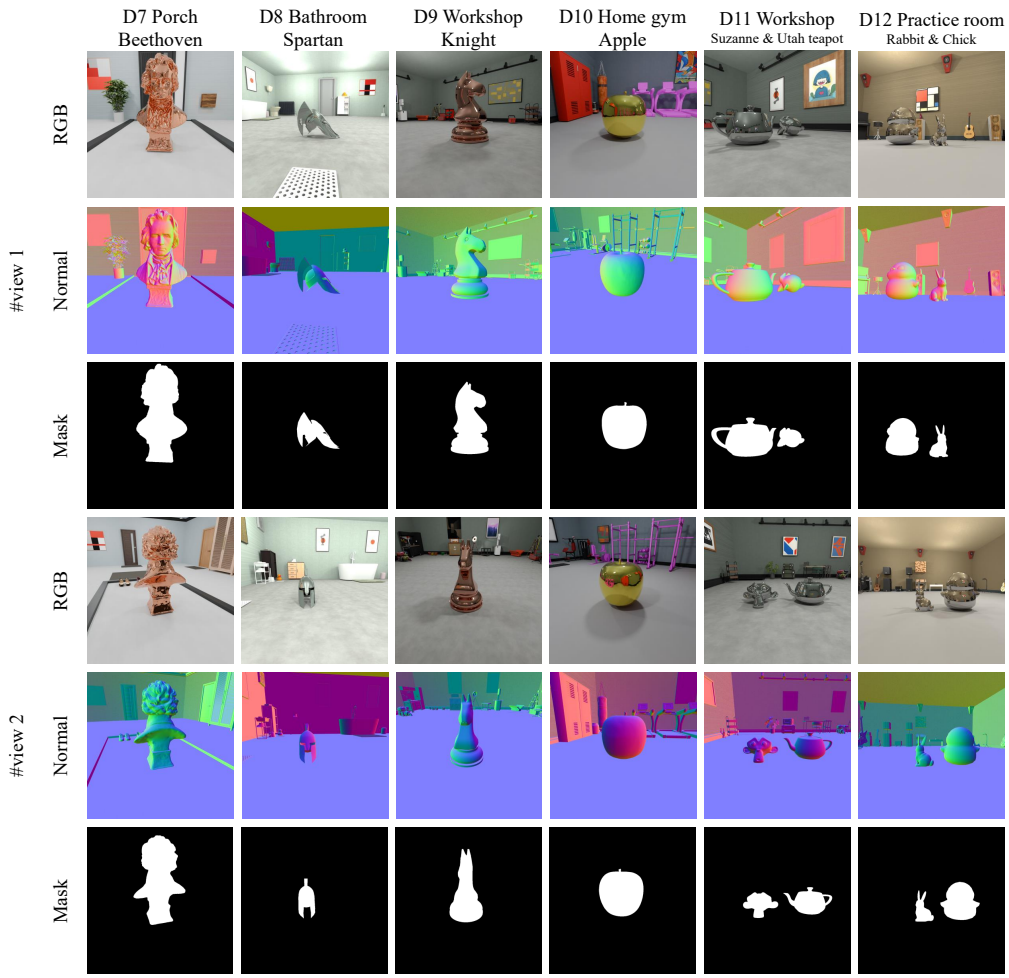

Figure S2: IndoorRef12 dataset examples (Rooms D7–D12). For each room, we show two viewpoints with the corresponding RGB image, world-space normal map, and object mask.

and represent world-space surface normals from Blender's default normal pass. Masks are stored as 16-bit grayscale PNGs in non-color space. To ensure consistent color management across modalities, the "Save as Render" option is disabled.

B.6 Dataset Summary

The final dataset consists of **12 rooms** $\times$ **200 viewpoints** $\times$ **3 modalities, resulting in 7,200 images total**. For each camera pose, RGB, normal, and mask are strictly aligned and share the same resolution and coordinate correspondence.

B.7 Scene Distribution (Construction Plan)

To diversify geometry and appearance while keeping controlled factors, we distribute scenes across: (i) background detail complexity (High/Low), (ii) number of lights (4/8), (iii) reflective object surface complexity (Simple/Complex), **(iv) single- vs. multi-object configuration**. Table S1 summarizes the construction plan.

Train/Test Split. For each scene in IndoorRef12, we hold out every 8th view as the test set (25 test views) and use the remaining 175 views for training, following the standard 3DGS evaluation protocol Kerbl et al. (2023).

B.8 Scope and Limitations

The dataset intentionally excludes transparent materials and real-world sensor noise. All scenes are indoor-only with closed room layouts, designed to isolate reflection behavior. Robustness to non-cuboid layouts is evaluated separately in App. I.2, and extensions to transparent materials or hybrid synthetic–real captures remain future work.

B.9 Asset and Annotation Notes

We use assets with redistribution-friendly licenses (e.g., CC0) when possible. We also avoid object choices that tend to produce unstable instance masks (e.g., objects with detachable bases, multi-part assemblies that are hard to separate, or extremely fine fur/feather-like details).

C Analysis of UGMS

This section analyzes why the uncertainty-guided metallic supervision (UGMS) localizes reflective objects and examines its behavior beyond uniformly metallic targets.

C.1 Learned Uncertainty

Table S3: **Statistics of the learned uncertainty σ on IndoorRef12.** Top: fraction of pixels exceeding each threshold on the metallic target and on the diffuse background, aggregated over all 12 scenes. Bottom: location of the uncertainty-guided prompts after visibility weighting, top-K selection, and 3D NMS.

Region	$P(\sigma > 0.1)$	$P(\sigma > 0.5)$
Metallic target	22.9%	5.7%
Diffuse background	4.8%	0.46%
Ratio (target / background)	4.8×	12.5×

Prompt statistic	Value
Top-8 prompts on the metallic target (average per scene)	8 / 8
Scenes with all 8 prompts on the target	12 / 12
Prompts falling outside the target	0 / 96

UGMS relies on the learned uncertainty σ to locate reflective objects. Since high uncertainty can in principle also arise from geometry errors, occlusions, or poorly observed regions, we analyze why σ nonetheless concentrates on metallic surfaces from two perspectives and verify it statistically.

Optimization perspective. Minimizing the negative log-likelihood in equation 5 of the main paper with respect to σ drives its optimum toward the magnitude of the photometric residual in each region. Residuals caused by geometry errors or occlusions are transient. They are confined to a subset of views and diminish as the geometry converges. Residuals caused by specular reflection cannot be explained by any view-independent representation and therefore persist across views even after convergence. The learned σ thus selectively captures this persistent component.

Physical perspective. The persistent component is strongest on metallic surfaces. Cross-view color variation is governed by the Fresnel reflectance of the material, which at normal incidence is about 0.04 for common dielectrics and 0.5 to 1.0 for metals. Metals also exhibit little diffuse reflection, so the view-invariant part of their appearance is weak while the view-dependent specular part dominates. This is consistent with the material swap experiment in Sec. C.3, where replacing a metallic target with a smooth dielectric lowers its uncertainty to the background level.

Statistical verification. Tab. S3 confirms this separation on IndoorRef12. Strong uncertainty ($\sigma > 0.5$) occurs 12.5× more often on the metallic target than on the diffuse background, and the gap is 4.8× even at the lower threshold. The prompt selection further suppresses the remaining

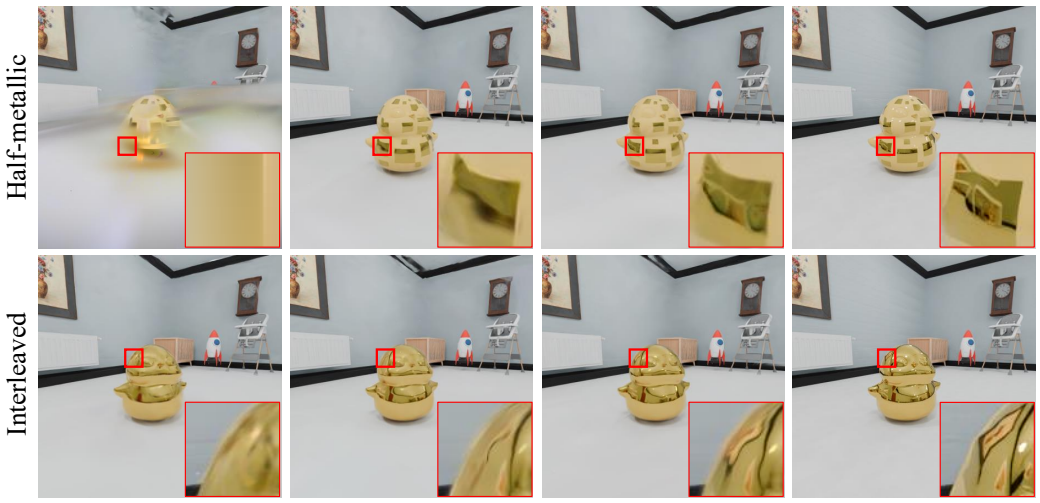

Figure S3: **Partially metallic objects.** Top: half-metallic object. Bottom: interleaved metallic patches. For each case, we show the ground truth, and novel-view renderings without guidance, with the oracle material-level mask, and with our object-level guidance. Although the guidance mask covers the whole object in both cases, our rendering avoids the baked-in artifacts of the unguided model and closely matches the oracle.

Table S4: **Partially metallic objects.** *Half-metallic* splits a single object vertically into a metallic half and a matte dielectric half. *Interleaved* embeds scattered metallic patches in a matte dielectric object. *Oracle* is trained with the ground-truth material-level metallic mask. Mask IoU is measured against the metallic region.

Guidance	*Half-metallic*		*Interleaved*	
	Mask IoU ↑	PSNR ↑	Mask IoU ↑	PSNR ↑
None (w/o UGMS)	–	28.92	–	28.01
Ours (object-level)	0.45	36.63	0.23	37.20
Oracle (material-level)	1.00	36.73	1.00	36.61

high-uncertainty regions in the background. Visibility weighting, top-K selection, and 3D NMS retain only strong and spatially consistent candidates, so isolated artifacts are rarely selected over persistently reflective regions. As a result, all 96 selected prompts across the 12 scenes fall on the metallic target.

Limitation and future direction. A single scalar σ does not separate aleatoric uncertainty, which reflects persistent appearance variation under sufficient observations, from epistemic uncertainty caused by insufficient observations. We currently rely on visibility weighting to suppress the latter. Modeling the two explicitly would distinguish reflective regions from poorly observed regions in a more principled way, and we leave this to future work.

C.2 Partially Metallic Objects

UGMS produces a metallic guidance mask at the object level, because SAM2 segments objects rather than materials. To examine how this granularity affects objects that are only partially metallic, we construct two controlled scenes (Tab. S4, Fig. S3).

Half-metallic object. A single object is vertically split into a metallic half and a matte dielectric half, with all other properties unchanged. All eight uncertainty-guided prompts land on the metallic half, consistent with the material selectivity of σ (Sec. C.3). SAM2, however, expands the mask to the whole object, giving an IoU of 0.45 against the metallic half. Despite this object-level guidance,

Table S5: **Material selectivity of the learned uncertainty.** We take three IndoorRef12 scenes and change only the target material from metallic to a smooth dielectric (metallic $= 0$, roughness $= 0.05$), keeping geometry, illumination, and camera trajectory identical. σ denotes the learned uncertainty. The dielectric target falls to the background level (1.22% in the same scenes), and the uncertainty-guided prompts no longer land on it.

Scene	$P(\sigma > 0.5)$, metallic	$P(\sigma > 0.5)$, dielectric	Ratio
Kitchen (Rabbit)	4.35%	0.73%	6.0×
Baby room (Chick)	9.19%	2.24%	4.1×
Living room (Vase)	6.79%	0.86%	7.9×
All three scenes	6.35%	1.03%	6.2×
Top-8 prompts on the target	24 / 24	2 / 24	

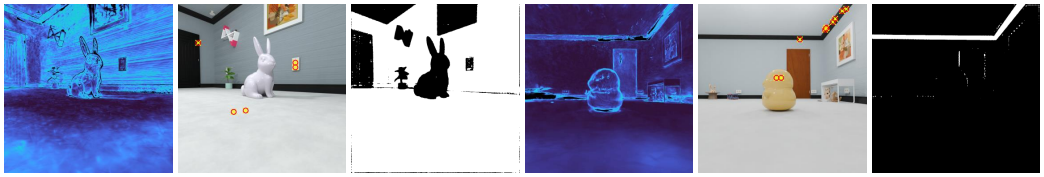

Figure S4: **UGMS on dielectric targets.** The Rabbit (left) and Chick (right) scenes of Tab. S5 with the target material swapped to a smooth dielectric. For each scene, we show the learned uncertainty map, the selected point prompts, and the resulting metallic guidance mask. The uncertainty no longer concentrates on the target, most prompts fall on background structures, and the guidance mask no longer corresponds to the target object.

novel-view synthesis improves by 7.7 dB over the model without guidance (28.92 to 36.63 dB) and stays within 0.1 dB of the oracle.

Interleaved patches. To test a harder case, we embed spatially scattered metallic patches in an otherwise matte object. The object-level mask covers the whole object and reaches an IoU of only 0.23 against the patches, yet it improves novel-view synthesis by 9.2 dB over the model without guidance (28.01 to 37.20 dB) and even slightly exceeds the oracle trained with the patch-level mask (36.61 dB). This does not mean that the object-level mask assigns materials more accurately, but it indicates that object-level guidance is easier to optimize for three reasons. First, the patch boundaries are finer than the spatial extent of individual Gaussians, so a Gaussian that straddles a boundary receives conflicting metallic and non-metallic supervision under the patch-level mask, whereas the object-level mask supervises every Gaussian on the object consistently. Second, thin patches projected into each 2D view are affected by occlusion, resolution, and slight misalignment, so the patch-level mask may provide supervision that varies across views, while the object-level mask remains coherent across all views. Third, the guidance is soft and jointly optimized with the photometric loss, so an object-level mask marks where reflections may occur and leaves the per-Gaussian material to be resolved by the multi-view observations, whereas a patch-level mask imposes a rigid spatial prior that the optimization must follow even where it conflicts with the observed appearance.

Summary. Object-level guidance is sufficient to prevent baked-in reflections and to preserve both novel-view rendering quality on partially metallic objects. It does not recover material boundaries within an object, which limits per-material editing, and we regard fine-grained material segmentation as a direction for future work.

C.3 Material Selectivity

UGMS does not classify materials. It identifies regions whose appearance varies persistently across views, and this signal is governed by the Fresnel reflectance of the surface (Sec. 4). To verify that the signal is selective rather than merely correlated with foreground objects, we swap the target material

from metallic to a smooth dielectric in three scenes while holding everything else fixed (Tab. S5). The strong-uncertainty rate on the target drops from 6.35% to 1.03%, statistically indistinguishable from the diffuse background of the same scenes, and the prompt selection follows accordingly (24/24 versus 2/24 prompts on the target). As visualized in Fig. S4, the uncertainty on the dielectric target is no stronger than on the surrounding room, and the prompts no longer concentrate on it. This confirms that the metallic guidance is driven by material, not by object saliency. The same property defines the scope of our method. Weakly specular materials such as polished plastic or marble produce photometric inconsistency too mild to activate UGMS, and the optimization may absorb their subtle specular effects into the diffuse term.

D Implementation Details

Our training procedure consists of two stages, executed sequentially on a single NVIDIA RTX A6000 GPU. We provide full timing and hyperparameter details below to facilitate reproduction.

Stage 1: Uncertainty Learning on 3DGS. We first train a standard 3DGS reconstruction model augmented with a learnable per-Gaussian uncertainty scalar $\sigma_i \geq 0$ for $30{,}000$ iterations.

- Uncertainty learning is activated at iteration $1{,}000$ (after initial geometry stabilizes). Before this iteration, only the standard photometric loss is applied.
- Each σ_i is parameterized through a softplus activation to ensure non-negativity and is optimized jointly with the standard Gaussian parameters via the NLL loss in Eq. (5) of the main paper.
- After Stage 1 training completes, we compute a visibility score v_i for each surviving Gaussian by performing a single forward pass over all training views and counting the number of views in which Gaussian i is rendered (i.e., its projected radius is positive). This post-hoc measurement ensures that all Gaussians, regardless of when they were spawned during training, are evaluated under identical conditions on the same view set.

Stage 1→2 Transition: Pseudo Mask Generation. After Stage 1 completes, SAM2 is invoked *once* to generate the metallic guidance map $\mathbf{M}_{\text{met}}$. This is not an iterative process during training. The role of Stage 1 is solely to produce $\mathbf{M}_{\text{met}}$; the learned 3D Gaussian parameters and per-Gaussian uncertainty σ_i are not carried over to Stage 2.

- We rank all Gaussians by the visibility-reweighted score $s_i = \sigma_i \cdot v_i / \max_k v_k$ (Eq. 6 in the main paper).
- Greedy 3D non-maximum suppression (NMS) with a fixed minimum separation of 0.1 in scene units is applied to obtain the top $K = 8$ spatially diverse candidates $\mathcal{P}$.
- For each training view, $\mathcal{P}$ is projected to 2D after visibility filtering and used as positive point prompts for SAM2.1 (Hiera-L backbone). Visibility filtering combines two checks: (i) a multi-view aggregate filter that retains only Gaussians rasterized with positive radius in at least 30% of training views ($v_i / N_{\text{view}} \geq 0.3$), removing spurious candidates that survive in too few views to be reliable; and (ii) a *per-view depth-consistency check* that compares each candidate's camera-space depth z_i against the rendered surface depth $D(u, v)$ at its projected pixel, rejecting the candidate for that view if $z_i > D(u, v) + \tau$. This prevents occluded Gaussians from being projected as foreground prompts that would mislead SAM2.
- The resulting $\mathbf{M}_{\text{met}}$ is generated once and held *static* throughout Stage 2.

Stage 2: PBR Optimization with Metallic Supervision and PCC. We then train the reflection-aware PBR 2DGS model for an additional $30{,}000$ iterations.

- Metallic supervision $\mathcal{L}_{\text{met}}$ (Eq. 8 in the main paper) activates at iteration $3{,}000$ with weight $\lambda_{\text{met}} = 0.2$, applied as an L1 loss on the rendered reflectivity map.
- Background regularization weight $\lambda_{\text{bg}} = 0.05$ (Eq. 7) is held fixed across all scenes.
- PCC sampling (Sec. 4.3 of the main paper) is enabled from iteration $3,000$ of Stage 2.

- PCC proxy box parameters: a uniform scale factor of 1.1 is applied to the PCA-aligned bounds, and quantile thresholds $[0.01, 0.99]$ are used to remove outliers when estimating the proxy from the SfM point cloud. These parameters are *fixed across all scenes*; no per-scene tuning is performed.
- The environment illumination $E(\cdot)$ is represented as a cubemap with 6 faces of 128×128 resolution per face (RGB, 32-bit float, linear HDR space). It is initialized as a uniform mid-gray (0.5 across all pixels) and optimized jointly with the 2DGS parameters. Specular shading queries the cubemap via the split-sum approximation Yao et al. (2024).
- SSIM weight $\lambda_{\text{ssim}} = 0.2$ in $\mathcal{L}_{\text{photo}}$ (Eq. 14).

Inference. The per-Gaussian uncertainty parameters σ_i and SAM2 are used *only during training*. At test time, only the PBR 2DGS representation with PCC environment sampling is required, yielding real-time rendering.

E Efficiency: Detailed Runtime Analysis

We compare end-to-end training time and inference speed against representative reflection-aware 3DGS baselines on IndoorRef12. All measurements are conducted on a single NVIDIA RTX A6000 GPU. For our method, training time corresponds to Stage 1 + Stage 2 of the full pipeline.

Table S6: Training time and inference speed on IndoorRef12. Training time is reported in hours. For our method, training time includes Stage 1 (uncertainty learning on 3DGS, $20{,}000$ iterations) and Stage 2 (PBR optimization on 2DGS, $30{,}000$ iterations). The PCC overhead at inference is approximately 1.15 ms per frame ($\sim$6% of total).

Method	Training (h) ↓	Inference (FPS) ↑
Ref-Gaussian Yao et al. (2024)	0.73	76
Ref-GS Zhang et al. (2025b)	1.17	30
MGSR Zhou et al. (2025)	1.30	82
IRGS Gu et al. (2025)	2.78	0.05
SpecGloss-GS Kouros et al. (2026)	0.63	39
Ref-DGS Fan et al. (2026)	0.43	46
Ours	1.19	51

Discussion. Table S6 reports end-to-end training time and inference speed on IndoorRef12. Our end-to-end pipeline (Stage 1 and Stage 2) requires 1.19 hours of training, comparable to Ref-GS Zhang et al. (2025b) (1.17 h) and faster than MGSR Zhou et al. (2025) (1.30 h) and IRGS Gu et al. (2025) (2.78 h). Ref-DGS Fan et al. (2026) (0.43 h), Ref-Gaussian Yao et al. (2024) (0.73 h), and SpecGloss-GS Kouros et al. (2026) (0.63 h) train faster, as they use a single optimization stage without explicit reflective-region supervision, which however leads to less stable diffuse–specular decomposition on highly specular indoor surfaces. At test time, our method runs in real time at 51 FPS, faster than the recent reflection-aware methods Ref-DGS (46 FPS) and SpecGloss-GS (39 FPS), and orders of magnitude faster than ray-tracing-based inverse rendering such as IRGS (0.05 FPS). The deferred PBR shading with PCC sampling introduces a moderate overhead compared to lighter pipelines such as Ref-Gaussian (76 FPS) and MGSR (82 FPS), of which PCC itself accounts for only about 1.15 ms per frame. The uncertainty learning, NLL optimization, and SAM2 pass impose no inference-time cost, as they are used only during training.

E.1 Runtime Breakdown

Tab. S7 breaks down the cost of the components added by our method. None of them is a bottleneck. The uncertainty σ is a single scalar per Gaussian that is optimized jointly within the standard Stage 1 training, so it adds no separate stage. SAM2 is invoked only once, after Stage 1, to produce the metallic guidance mask, which then remains fixed throughout Stage 2. It therefore adds no per-iteration cost and no inference cost, and its one-time inference takes about 80 seconds, less than 2% of the end-to-end training time. PCC adds 1.15 ms per frame at inference for the ray–box

Table S7: **Component-wise runtime of our method on IndoorRef12.** Training time is measured end-to-end per scene, and inference time per frame. All measurements are on a single NVIDIA A6000 GPU.

Component	Runtime	Fraction
Training		
Stage 1 (Gaussian optimization with per-Gaussian σ)	34 min	$\sim$48%
SAM2 guidance mask (one-time, 0.46 s/view $\times$ 175 views)	80 s	$\sim$2%
Stage 2 (PBR optimization with PCC)	$\sim$36 min	$\sim$50%
Total	1.19 h	100%
Inference		
PCC direction correction	1.15 ms/frame	$\sim$6%
Full rendering	51 FPS	–

intersection and direction correction, about 6% of the frame time. The dominant cost lies in the two-stage Gaussian optimization itself, which is shared with existing methods. As a result, our total training time is comparable to Ref-GS (1.17 h), and the full model runs in real time at 51 FPS, whereas ray-tracing-based inverse rendering such as IRGS runs at 0.05 FPS (Tab. S6).

F Additional Qualitative Results on IndoorRef12 and GlossyReal

We provide additional qualitative results for each scene in the IndoorRef12 dataset. For every scene, we render novel views under the same evaluation protocol as in the main paper and include side-by-side comparisons with representative baselines. These results highlight the effect of PCC on near-field reflection parallax and the improved reflection anchoring in cluttered indoor environments. Figure S5 presents additional per-scene qualitative comparisons. Figure S6 provides additional novel-view synthesis and inverse rendering results on both IndoorRef12 and GlossyReal Liu et al. (2023), including normal and specular maps that further illustrate the geometric and view-dependent fidelity of our method. Figure S7 compares illumination decomposition against MaterialRefGS Zhang et al. (2025a) and IRGS Gu et al. (2025), showing more reliable estimates of roughness, normal, metallic, albedo, and environment maps with our method.

G Applications

G.1 Material Editing

Our reflection-aware decomposition enables intuitive material editing by directly manipulating the estimated PBR attributes while keeping the scene geometry and illumination fixed. Importantly, our method supports *object-centric* edits even in cluttered indoor environments: although the reconstructed scene contains complex backgrounds, we can modify the target object's material without spilling the edits into the surrounding walls, floor, or furniture.

As shown in Fig. S8, editing the intrinsic albedo changes the perceived base color of the reflective object (e.g., copper to gold/red) while maintaining consistent reflection behavior under indoor lighting. Similarly, increasing the roughness produces a smooth transition from sharp mirror-like reflections to more blurred highlights, matching the expected microfacet behavior. These results demonstrate that our method yields disentangled, spatially localized material parameters that facilitate stable material editing in indoor scenes.

G.2 Object-Level Relighting

We further demonstrate an object-level relighting application enabled by our UGMS module and reflection-aware material decomposition. Given an indoor scene with a cluttered background, UGMS provides an object mask for the reflective target, allowing us to isolate the metallic object from the background. We then re-render only the segmented object under novel environment illumination (HDR maps) while keeping the scene geometry fixed. This design avoids the physically inconsistent

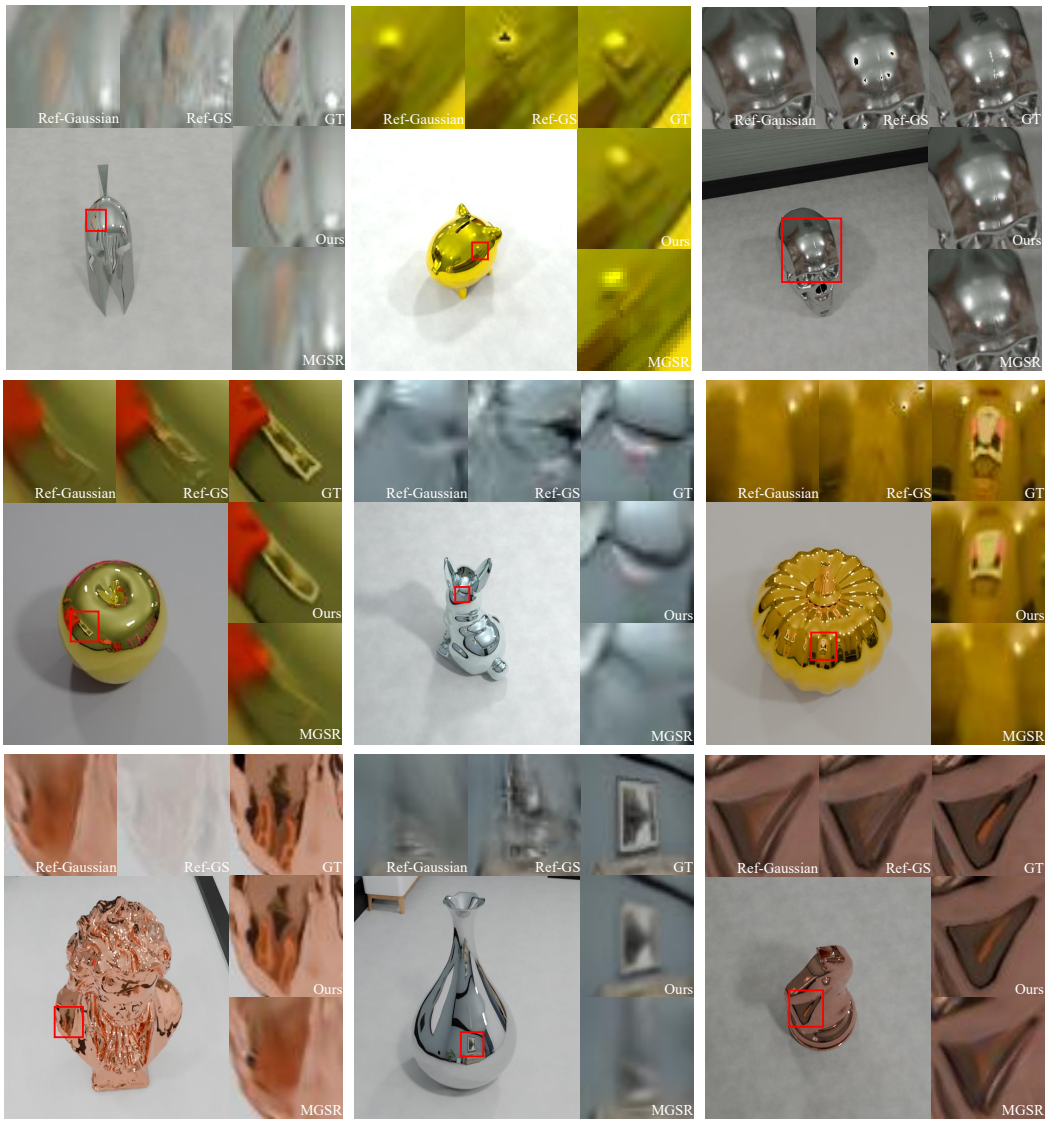

Figure S5: **Additional qualitative comparisons on IndoorRef12.** Columns follow the order in the header.

setting of changing the illumination of an object while keeping the original indoor background unchanged. As a result, our method supports clean relighting of metallic objects even in complex indoor scenes by localizing the edit to the target object.

Comparison with prior methods. We additionally compare relighting results with three recent reflection-aware baselines (Fig. 10). IRGS Gu et al. (2025) fails to produce coherent renderings under novel HDR illumination, exhibiting severely degraded appearance and missing specular response. Ref-Gaussian Yao et al. (2024) and MaterialRefGS Zhang et al. (2025a), while producing structurally complete renderings, show minimal change in specular appearance across different HDR environments. This indicates a baked-in artifact, in which the view-dependent reflection has been absorbed into the diffuse term, so changing the environment map has no visible effect on the final appearance. In contrast, our method produces clear color shifts and reflection changes that respond to the novel illumination, evidencing successful diffuse–specular decomposition.

As shown in Fig. 10, changing the environment illumination yields plausible variations in specular appearance and highlight placement on metallic surfaces. Across HDR maps, the object geometry

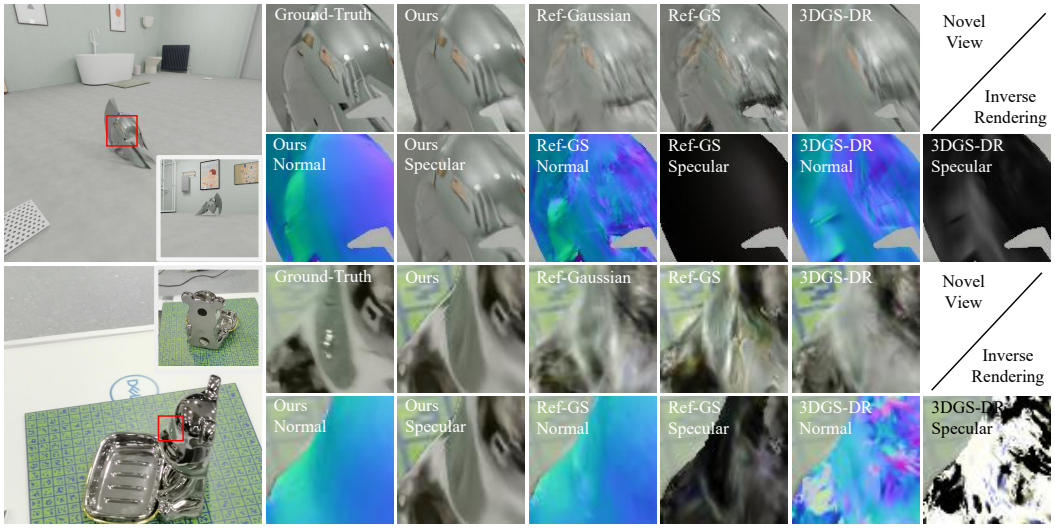

Figure S6: **Additional novel-view synthesis and inverse rendering comparisons on IndoorRef12 and GlossyReal Liu et al. (2023).** Each row shows novel-view rendering (top) and the corresponding normal/specular maps (bottom) across baselines.

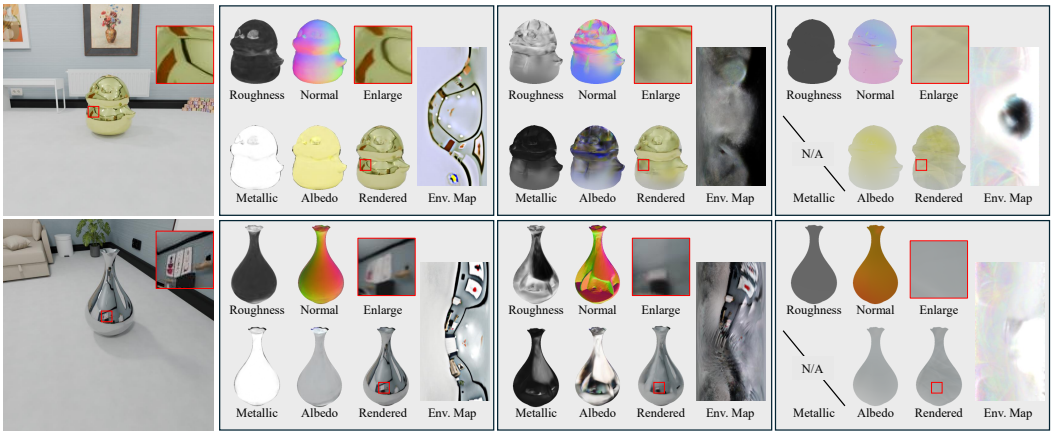

Figure S7: **Additional illumination decomposition comparisons against MaterialRefGS Zhang et al. (2025a) and IRGS Gu et al. (2025) on IndoorRef12.** For each method, we visualize roughness, normal, metallic, albedo, rendered image, and environment map. "N/A" indicates that the method does not predict the corresponding component.

is preserved and the reflection behavior remains coherent. Since only the object segmented by the UGMS guidance mask is re-rendered, the edit is confined to the target and leaves the surrounding scene unchanged. These results indicate that UGMS not only stabilizes material optimization but also provides practical object masks that facilitate downstream editing applications such as relighting.

H UGMS-Based Metallic Guidance Mask Generation and Comparisons

In this supplementary section, we describe how we obtain the pseudo metallic guidance map $\mathbf{M}_{\text{met}}$ from uncertainty-guided prompts (UGMS), and compare it with alternative mask acquisition strategies that do not use UGMS. All comparisons use the same SAM2.1 checkpoint/backbone (Hiera-L) Ravi et al. (2024) and identical preprocessing (RGB conversion); inference is run on GPU with mixed precision when available.

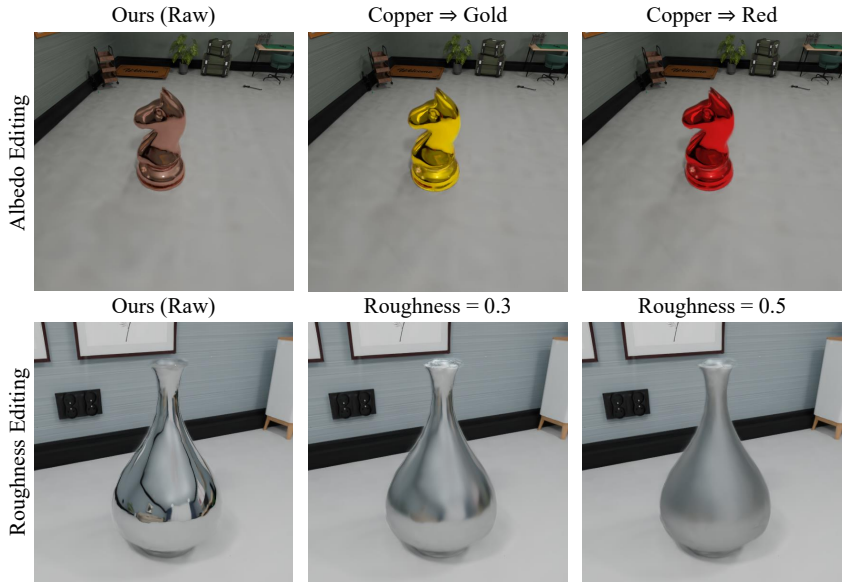

Figure S8: **Material editing results enabled by our reflection-aware 2DGS.** Top: **albedo editing** of a reflective chess piece (copper → gold/red) while preserving view-dependent reflections. Bottom: **roughness editing** of a metallic *vase* by increasing roughness ($\rho = 0.3, 0.5$), which progressively attenuates sharp specular highlights. Despite cluttered backgrounds, edits are localized to the target object and do not affect surrounding structures.

H.1 Ours: Uncertainty-Guided Metallic Guidance Mask via SAM2

Reflective regions tend to exhibit larger cross-view appearance variations in cluttered indoor scenes. UGMS leverages this cue to generate sparse prompts that guide SAM2 toward reflective/metallic regions. Specifically, we obtain a compact set of reliable prompts from the uncertainty signal, project them to each training view, and run SAM2 conditioned on these prompts to produce a dense pseudo metallic guidance map $\mathbf{M}_{\text{met}} \in [0, 1]$.

H.2 Alternative Mask Acquisition without UGMS

To evaluate the impact of UGMS-guided prompting, we compare against the following strategies:

(A) Prompt-free segmentation (AMG, Top-K union). For each RGB frame, we run SAM2.1's Automatic Mask Generator (AMG) without any external prompts. We aggregate the Top-K proposals ($K = 15$) into a single binary mask per frame by taking their pixel-wise union.

(B) Material-prior-guided prompting (Marigold). For each RGB frame, we obtain a material prior using a foundation model for material estimation (Marigold Ke et al. (2025)), which provides a coarse indication of regions likely to exhibit reflective/metallic behavior. We convert this prior into SAM-style mask prompts and run SAM2.1 conditioned on these prompts. Among the returned candidates, we select the highest-confidence prediction as the final binary mask.

(C) Photometric variance via depth-based warping (MaterialRefGS). MaterialRefGS Zhang et al. (2025a) infers reflective regions by computing multi-view photometric variance through depth-based warping, treating pixels with high cross-view appearance disagreement after warping as candidates for reflective regions. This approach, however, suffers from a circular dependency in highly specular regions, where depth estimation itself becomes unreliable due to view-dependent appearance, leading to inaccurate variance estimates. Furthermore, the supervision is derived from the model's own predictions. Even with an accurate object mask, the metallic values within the mask are regularized toward their average over the object, so once reflections are already baked into the diffuse term and the predicted metallic is low, this averaging preserves the low value instead of raising

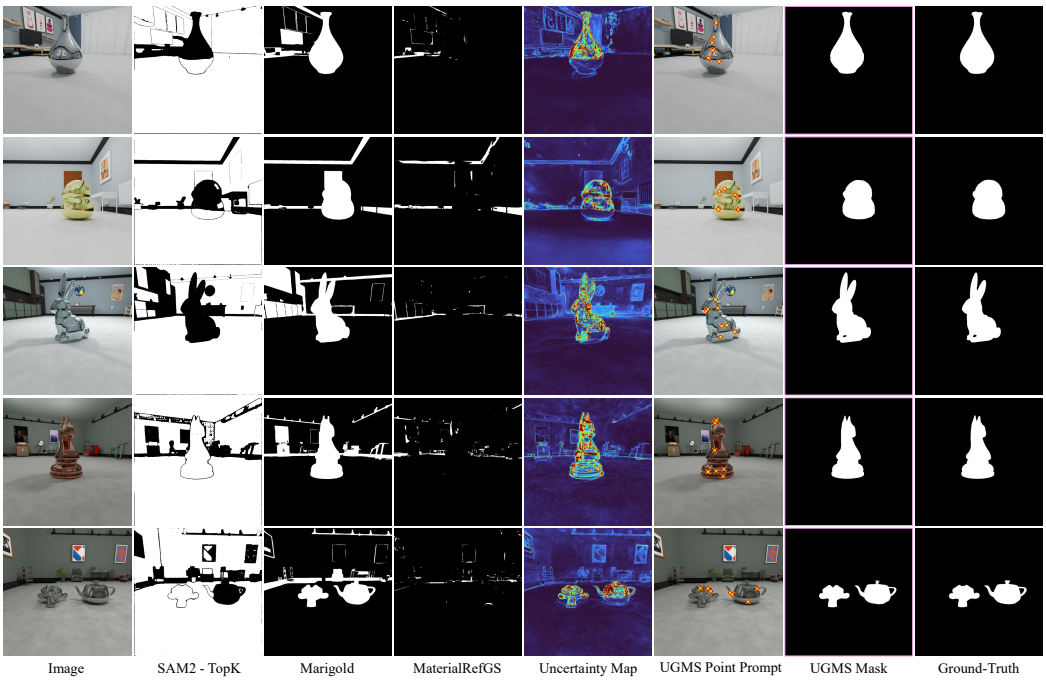

Figure S9: **Comparison of metallic/foreground mask acquisition strategies. SAM2 TopK/AMG (prompt-free)** aggregates the union of Top-K proposals ranked by predicted IoU, which often over-includes background structures in cluttered indoor scenes. **SAM2 with Marigold prior (material-guided prompting)** converts a material-estimation prior into a bounding box and sparse positive/negative point prompts; its output is sensitive to the reliability of the prior. **Ours (UGMS-guided prompting)** converts uncertainty-guided prompts into 2D point prompts and densifies them with SAM2, producing cleaner and more stable masks that better match the ground truth.

it, reinforcing the baked-in solution. In contrast, UGMS supervises the metallic prediction with a target that does not depend on the current material estimate.

Discussion. As shown in Fig. S9, prompt-free AMG often over-includes background structures in cluttered indoor scenes (and unioning Top-K proposals can further amplify this issue), while Marigold-guided prompting is sensitive to the reliability of the estimated material prior under complex illumination and strong reflections. MaterialRefGS, in turn, is limited by its reliance on warping-based variance, which becomes unstable on highly specular surfaces where depth itself is unreliable. In contrast, UGMS bypasses these dependencies by learning per-Gaussian uncertainty directly during optimization, independent of explicit depth correspondence or external material priors. This produces cleaner and more stable masks that better match the ground truth.

I Additional Ablation Studies

I.1 Sensitivity to λ_{MET}

We evaluate the sensitivity of our method to the metallic supervision weight λ_{met} by varying it over the range $[0.1, 0.3]$ on IndoorRef12. As shown in Tab. S8, performance remains stable across this range, with PSNR varying by only 0.05 dB. This indicates that our method is not sensitive to λ_{met}, and we use the default value of 0.2 in all main experiments. Other coefficients (λ_{bg}, λ_{ssim}) follow standard settings from prior work and are not tuned per scene.

Table S8: Sensitivity to the metallic supervision weight λ_{met} on IndoorRef12. Performance is stable across the tested range, with only 0.05 dB PSNR variation.

λ_{met}	PSNR ↑	SSIM ↑	LPIPS ↓
0.1	33.67	0.958	0.183
0.2	**33.72**	**0.962**	**0.176**
0.3	33.69	0.959	0.179

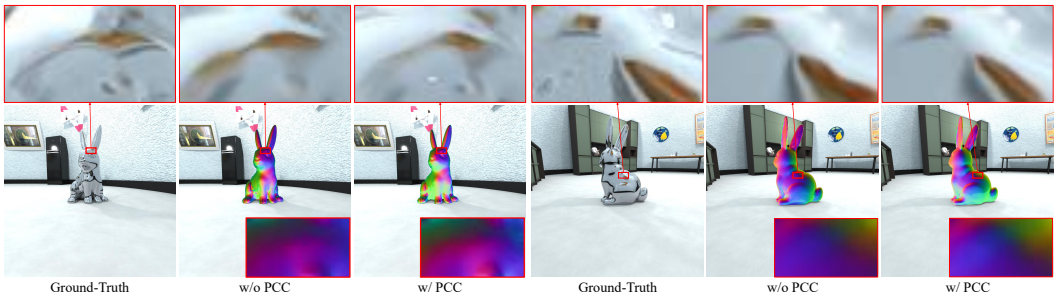

Figure S10: **Robustness of PCC in a dome-shaped room.** Two views of the dome-shaped scene, where the cuboidal proxy box does not match the actual ceiling geometry. For each view, we compare ground truth, rendering without PCC, and with PCC. Insets show zoomed-in reflections on the rabbit surface (top) and the corresponding normal maps (bottom). PCC produces more accurate reflection anchoring and sharper specular highlights, despite the proxy–geometry mismatch.

I.2 PCC Robustness under Irregular Layouts

The PCC module relies on a cuboidal proxy box, which is a coarse approximation of the surrounding geometry. To evaluate PCC beyond this assumption, we construct two stress-test scenes that are separate from the 12 IndoorRef12 scenes and report their quantitative results in Tab. 5 of the main paper.

Dome-shaped room. We construct a non-rectangular indoor structure with a domed ceiling. In this scene, the cuboidal proxy intentionally mismatches the dominant geometry, since a dome cannot be accurately enclosed by an axis-aligned box.

As shown in Tab. 5 of the main paper and Fig. S10, PCC improves over the infinite-environment baseline in all three metrics even under this geometry mismatch. This indicates that PCC remains effective as long as coarse spatial information is preserved in the proxy, since occluded regions beyond the room boundary do not significantly contribute to first-order reflections on the target object.

Near-field furniture. In a living room, we place furniture directly next to a metallic vase, so that the furniture occupies a large portion of the reflection on the object surface and lies well inside the proxy box. Compared with the standard IndoorRef12 scenes, this configuration brings the reflector much closer to the target. All baselines are trained with the same training and test views as in the main experiments. As shown in Fig. 8 and Tab. 5 of the main paper, PCC reproduces the reflected furniture more faithfully than both ray-tracing-based methods and Ref-DGS, although the furniture is not represented explicitly by the proxy. This is because PCC corrects the environment query direction through the proxy intersection, and the learned cubemap stores the near-field content at the location that best explains the multi-view observations.

Why PCC remains effective under proxy mismatch. PCC links the environment query direction to a ray–proxy intersection point $\mathbf{P}_{\text{hit}}$, which serves as a spatial anchor for sampling. This anchoring is robust to local geometric inaccuracies of the proxy because (i) regions beyond the visible scene (e.g., outside a domed ceiling) do not contribute to reflections on the object surface, and (ii) the dominant near-field reflectors (walls, furniture, floor) are still approximately enclosed by the cuboidal proxy.

Table S9: **Masked-region evaluation on IndoorRef12.** We report full-image PSNR, masked-region PSNR, the PSNR drop $\Delta = \text{Full} - \text{Masked}$, and masked-region LPIPS. Although all methods degrade on reflective regions, our method achieves the best masked PSNR and LPIPS while showing a smaller drop than the strong recent baselines.

Method	Full PSNR ↑	Masked PSNR ↑	Δ drop	Masked LPIPS ↓
Ref-Gaussian Yao et al. (2024)	24.21	17.50	6.71	0.342
MaterialRefGS Zhang et al. (2025a)	31.27	20.54	10.73	0.222
Ref-GS Zhang et al. (2025b)	31.47	20.85	10.62	0.215
MGSR Zhou et al. (2025)	31.67	21.20	10.47	0.208
Ref-DGS Fan et al. (2026)	32.72	21.97	10.75	0.198
Ours	**33.72**	**23.82**	9.90	**0.192**

Standard full-image PSNR/SSIM/LPIPS metrics are dominated by background pixels, which are easier to reconstruct and can therefore underestimate the performance gap on reflective objects. To better isolate the behavior on glossy surfaces, we evaluate reconstruction quality *only within the reflective object regions*, using the ground-truth masks provided by IndoorRef12.

I.3 Comparison with MaterialRefGS: Metallic Guidance Mask IoU

We further compare metallic localization accuracy using IoU against the ground-truth reflective object masks on IndoorRef12. Tab. 6 shows that prompt-free SAM2 and strong material foundation priors (Marigold) achieve limited metallic IoU, and MaterialRefGS's depth-warping-based variance signal also fails in highly specular regions. UGMS-guided prompts substantially outperform all alternatives, confirming that the gain comes from the *learned uncertainty signal* rather than the segmentation model itself.

Implementation note on MaterialRefGS comparison. We use the official MaterialRefGS implementation Zhang et al. (2025a) for the photometric variance map (reflectance score map) computation, which is the core component of their method. We additionally implement the SAM2-based segmentation step following the paper's description, as this part was not included in the released code. All other components, including the depth-warping pipeline and variance score generation, follow the official implementation.

Why MaterialRefGS struggles in highly specular regions. As discussed in Sec. H.2, the depth-based warping of MaterialRefGS becomes unreliable where view-dependent appearance violates photoconsistency, and its supervision is derived from its own predictions. UGMS avoids both issues by learning per-Gaussian uncertainty directly during optimization, independent of explicit depth correspondences and of the current material estimate.

I.4 Masked-Region Evaluation

Standard full-image PSNR, SSIM, and LPIPS are dominated by background pixels, which are easier to reconstruct and can therefore underestimate the performance gap on reflective objects. To better isolate the behavior on glossy surfaces, we evaluate reconstruction quality *only within the reflective object regions*, using the ground-truth masks provided by IndoorRef12 (Tab. S9).

This evaluation reveals two findings. First, full-image metrics can obscure errors on reflective objects because large background regions dominate the aggregate score. Even when evaluation is restricted to the glossy object mask, only a subset of pixels corresponds to strongly specular highlights or high-frequency reflected content, so the region-level metric still provides a conservative estimate of reflection quality. Second, recent strong baselines in Tab. 1, including MaterialRefGS, Ref-GS, MGSR, and Ref-DGS, exhibit large PSNR drops on reflective regions despite competitive full-image performance. This indicates that their full-image scores are partly supported by accurate background reconstruction rather than faithful reflection modeling. Ref-Gaussian shows the smallest drop only because its full-image PSNR is already low.

Table S10: Quantitative results on multi-object scenes in IndoorRef12. Our method maintains strong performance and produces stable reflections across multiple reflective objects, indicating generalization beyond single-object settings.

Method	Multi-obj1 (Suzanne + Teapot)			Multi-obj2 (Rabbit + Chick)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MGSR	34.30	0.973	<u>0.059</u>	33.07	0.957	<u>0.303</u>
Ref-GS	33.55	<u>0.974</u>	0.061	<u>34.53</u>	0.960	0.358
Ref-Gaussian	24.12	0.903	0.219	23.46	0.901	0.452
MaterialRefGS	33.58	0.972	0.076	33.35	0.957	0.339
Ref-DGS	<u>35.15</u>	0.941	0.094	34.24	<u>0.964</u>	0.307
Ours	**35.36**	**0.977**	**0.058**	**34.65**	**0.966**	**0.251**

In contrast, our method achieves the highest masked PSNR and the lowest masked LPIPS. Compared with the strongest masked-region baseline, Ref-DGS, our method improves masked PSNR by 1.85 dB and reduces masked LPIPS from 0.198 to 0.192. Moreover, our PSNR drop is smaller than those of MaterialRefGS, Ref-GS, MGSR, and Ref-DGS, suggesting that the proposed uncertainty-guided metallic supervision and parallax-corrected reflection modeling better preserve reconstruction quality on reflective surfaces. This consistent pattern supports our central claim that stable diffuse–specular decomposition is a key differentiator for highly specular indoor novel-view synthesis.

J Multi-Object Scenes and Inter-Reflection

J.1 Quantitative Results on Multi-Object Scenes

To evaluate generalization beyond single-object settings, IndoorRef12 includes two multi-object scenes, each containing two reflective objects placed in close proximity to induce mutual reflection between them. Tab. S10 reports per-scene results against recent reflection-aware baselines.

J.2 Inter-Reflection Between Two Reflective Objects

Both UGMS and PCC operate at the per-Gaussian level and are not tied to a specific object. As a result, our method maintains strong performance and produces stable reflections across multiple objects.

While our method is based on a first-order reflection formulation and does not explicitly model multi-bounce inter-reflection, we observe that the quality of diffuse–specular decomposition strongly influences how inter-reflection regions are rendered. This is particularly evident when comparing against baselines that *do* attempt to capture inter-reflection. Ref-Gaussian Yao et al. (2024) and MaterialRefGS Zhang et al. (2025a) trace secondary rays against the Gaussian scene, and Ref-DGS Fan et al. (2026) assigns near-field reflections to a dedicated set of local reflection Gaussians. Despite these mechanisms, none of them outperforms our first-order model on the multi-object scenes (Tab. S10). The ray-tracing-based methods gain nothing over methods without ray tracing, as MaterialRefGS performs on par with Ref-GS and Ref-Gaussian falls more than 9 dB below it, and both exhibit noticeable artifacts in inter-reflection regions (Fig. S11, red boxes). We attribute this to their unstable diffuse–specular decomposition, in which specular components are baked into the diffuse term, so that ray tracing applied on top of this contaminated decomposition does not resolve and can even amplify the errors. Ref-DGS reaches a PSNR close to ours on Multi-obj1 but yields the lowest SSIM and the highest LPIPS among the compared methods, indicating structural artifacts that the pixel-wise error does not reflect. Since the reflected content of the neighboring object is stored as the appearance of separate primitives, it is not guaranteed to follow the viewpoint consistently.

In contrast, our method first stabilizes the diffuse–specular separation through UGMS, yielding a cleaner first-order reflection that is visually closer to the ground truth even at inter-reflection regions, despite using only first-order PCC sampling. This suggests that decomposition stability is a *prerequisite* for effective multi-bounce modeling, since explicit inter-reflection modeling on top of an unstable decomposition does not improve inter-reflection quality and can even degrade it. We view

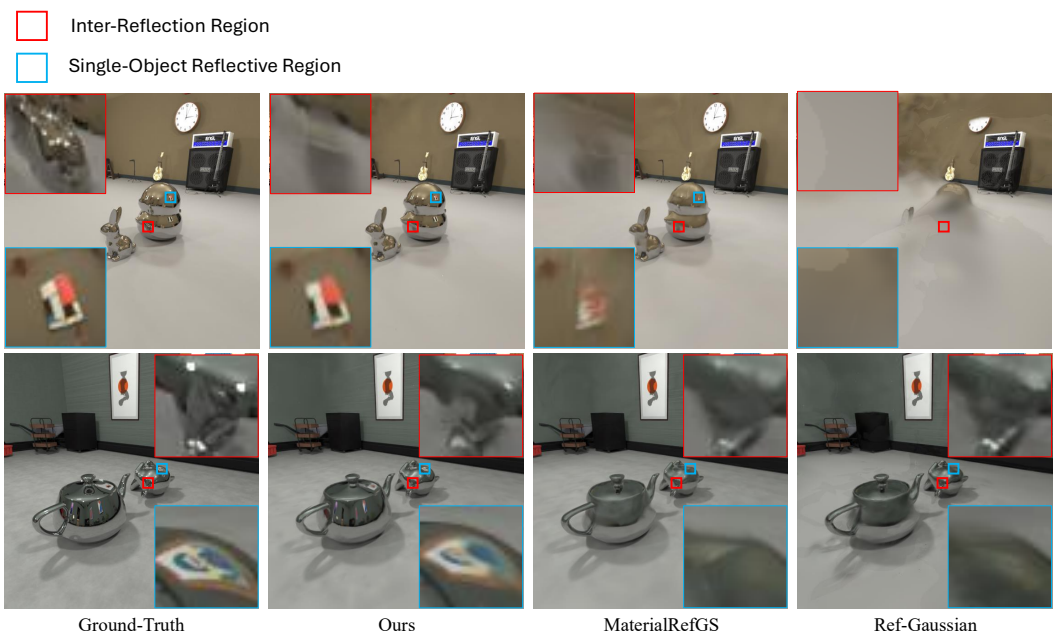

Figure S11: **Inter-reflection analysis on multi-object scenes.** Two scenes with closely placed reflective objects (Rabbit + Chick, Suzanne + Utah teapot). Red boxes mark inter-reflection regions where one object reflects another, and blue boxes mark single-object reflective regions for reference. Although our method is based on a first-order reflection formulation, the stable diffuse–specular decomposition produced by UGMS yields more coherent inter-reflection appearance than baselines that trace secondary rays or assign near-field content to additional reflection Gaussians.

stable first-order decomposition as a necessary foundation for future work on explicit multi-bounce modeling.

K Robustness and Limitations

K.1 Robustness of UGMS in Real Scenes

UGMS relies on persistent multi-view photometric inconsistency to identify reflective regions. We discuss how two sources of perturbation affect this signal.

Pose noise. Camera pose errors typically affect only a subset of views and are not view-consistent across the trajectory. Since UGMS aggregates uncertainty over many views via the NLL loss (Eq. 5), localized pose errors contribute weakly compared to systematic specular inconsistency.

Repetitive textures. Repetitive non-reflective textures are largely view-consistent under correct geometry, and therefore yield low photometric variance and low learned uncertainty. They do not strongly activate UGMS.

When UGMS may fail. Under strong perturbations (e.g., severe pose noise across the entire trajectory, or moving objects in the scene), uncertainty may also increase in non-reflective regions. Possible remedies include incorporating multi-view geometric consistency filtering or motion masking; we leave these extensions to future work.



Figure S12: **Photographer reflection handling on GlossyReal Liu et al. (2023).** Two scenes where the photographer is visible in the captured reflections (insets, ground truth). Since photographer reflections are inconsistent across views, the multi-view photometric objective naturally suppresses these transient signals during optimization, and our renderings produce the underlying object surface without the photographer's silhouette.

K.2 Limitations

We summarize the main limitations of our method, motivated by both the underlying assumptions of the formulation and our empirical observations. The material scope of UGMS, including dielectric targets and partially metallic objects, is analyzed separately in Sec. C.3 and Sec. C.2.

(1) First-order environment representation. PCC corrects where the environment is queried but still stores illumination as a single cubemap indexed by direction from the box center. Reflectors that are occluded when seen from the box center, or large mirror-like surfaces such as a metallic wall that themselves reflect the scene, cannot be represented faithfully by this formulation. This limitation is shared by environment-map-based methods, and replacing the single cubemap with a position-dependent representation is a natural extension.

(2) Multi-bounce inter-reflection. Strong multi-bounce inter-reflection between two highly reflective objects placed close together is not explicitly modeled. As shown in Fig. S11, our stable decomposition mitigates the resulting artifacts in inter-reflection regions, producing more coherent appearance than baselines whose specular components are baked into the diffuse term. Nevertheless, explicit multi-bounce modeling, for example via differentiable ray tracing on top of a stable decomposition, is a natural future direction.

(3) Dependence on external components and prompt selection. Our pipeline relies on COLMAP for camera poses and SAM2 for dense mask generation. Severe failures of these components, such as catastrophic SfM failure or degenerate SAM2 masks under extreme inputs, propagate to our method, although we did not observe such failures on IndoorRef12 or GlossyReal. In addition, prompts are selected by relative ranking of the uncertainty without an absolute threshold, so a scene without a strongly reflective object still yields prompts on the highest-uncertainty background regions (Fig. S4).

(4) Static-scene assumption and view-inconsistent reflections. The current method targets static indoor scenes. Dynamic objects, time-varying illumination, or significant scene motion violate the multi-view consistency assumption underlying UGMS and would require extensions such as motion-aware uncertainty modeling.

A related practical issue is the presence of view-inconsistent reflections in real captures, such as the photographer or capture rig appearing in the reflective surface. As shown in Fig. S12, since these transient reflections are not view-consistent, the multi-view photometric objective implicitly suppresses them during optimization, and our renderings recover the underlying object surface without the photographer's silhouette. While beneficial in this case, this behavior also indicates that genuinely view-dependent dynamic content cannot be preserved by the current formulation.

L BROADER IMPACTS

This work advances novel view synthesis for indoor reflective scenes, enabling applications in virtual reality, augmented reality, real-estate visualization, and digital twin reconstruction. The method requires multi-view captures of physical scenes and does not generate synthetic content from prompts, limiting its misuse potential. We do not foresee direct negative societal impacts beyond those already present in photorealistic 3D reconstruction (e.g., privacy considerations when capturing real environments without consent). Mitigations include following standard data collection ethics in capture-based applications.