GLINT: Modeling Scene-Scale Transparency via Gaussian Radiance Transport

Supplementary Material

This supplementary material provides additional details on (i) the proposed synthetic 3D-Front-T dataset (Sec. A), (ii) the baseline methods used in our experiments (Sec. B), and (iii) implementation details (Sec. C). In addition, we include qualitative results on ablation studies (Sec. D), comparative analysis with baselines (Sec. E), discussion on foundation models (Sec. F), detailed discussion on limitations and future work (Sec. G), and additional qualitative results on both real and synthetic datasets (Sec. H).

A. 3D-FRONT-T Dataset

To enable rigorous quantitative evaluation of scene-scale transparency reconstruction, we introduce 3D-FRONT-T, a new synthetic benchmark for scene-scale transparency reconstruction. We construct our dataset built upon the 3D-FRONT (3D Furnished rooms with layouts and semantics) dataset [5]. While existing real-world datasets contain transparent scenes [14], they lack ground-truth geometry annotations for quantitative evaluation on geometry reconstruction. Our 3D-FRONT-T addresses this by providing depth and normal ground truth alongside RGB renderings of transparent scenes.

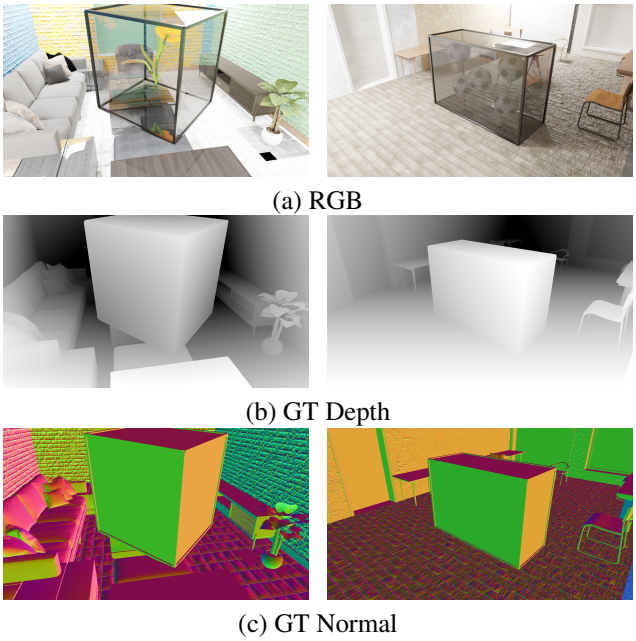
A.1. Dataset Construction

We collect 5 indoor scenes from the 3D-FRONT [5] datasets with diverse configurations. For each scene, we first identify the largest room by floor area and retain only the objects within that room, including walls, ceilings, doors, windows, and furniture placed on the floor. Then, we texture the floor, wall and ceiling in the scene with the diverse random materials following [15].

On top of this setup, we create and put a glass display container with a black metal frame, enclosing opaque objects. The container consists of six thin glass panels with a supporting frame structure. The container size is randomly determined to create different sizes of transparency regions.

To achieve realistic placement, we utilize Blender's physics-based simulation [7] to settle objects into physically plausible configurations. The glass container is initialized at a random position above the floor and released using rigid-body dynamics, allowing it to naturally come to rest on the floor or on top of existing furniture. This process ensures physically consistent placement while generating diverse occlusion patterns with the surrounding scene geometry.

All scenes are rendered using the Blender Cycles path tracer with 4096 samples per pixel and a maximum of 200 light bounces across all channels (diffuse, glossy, and transmission) to accurately capture complex light interactions through transparent surfaces. For our experiments, we render all images at a resolution of 960×540. For each scene, we generate camera trajectories which samples continuous viewpoints that ensure visibility of key objects including the transparent container. The example images of the dataset are shown in Fig. 1.



(a) RGB

(b) GT Depth

(c) GT Normal

Figure 1. **3D-FRONT-T dataset.** RGB, depth, and normal ground truth examples. As can be seen in the transparent surface (glass), the outgoing radiance for each pixel is entanglement of radiance from interface surface and the transmitted radiance.

A.2. Ground Truth Annotations

For each rendered viewpoint, we provide a complete set of physically accurate ground-truth annotations tailored for evaluating scene-scale transparency reconstruction:

RGB Images. High-fidelity renderings produced via multi-bounce path tracing in Blender Cycles. These images capture intricate light behavior including interreflections, refractions, and transmission through transparent materials, serving as a challenging benchmark for appearance modeling.

Depth Maps. Metric depth is extracted directly from Blender's rendering pipeline, ensuring physically consistent geometry even in regions partially occluded or viewed through transparent media. This enables rigorous evaluation of depth recovery in transparent scenes, a regime un-

derexplored in existing benchmarks.

Normal Maps. Per-pixel surface normals are rendered for all visible surfaces. These provide robust supervision signals for assessing fine-grained geometric accuracy, independent of texture or lighting cues.

Taken together, these annotations establish a comprehensive benchmark for transparent-scene reconstruction. To encourage reproducibility and extensibility, we will release our full dataset-generation pipeline, implemented on top of BlenderProc [4], enabling automatic synthesis of large-scale, diverse scenes with configurable and physically plausible object placement.

B. Overview of Baseline Methods

We compare GLINT with representative Gaussian-splatting–based approaches that cover planar-constrained geometry modeling, reflective radiance modeling, and transparent-surface reconstruction. Below, we briefly summarize each baseline to clarify their modeling assumptions and limitations in the context of scene-scale transparency.

2DGS [9] adopts 2D Gaussians to improve geometric accuracy over volumetric 3DGS. While effective for opaque surfaces, its monolithic α-compositing cannot separate interface geometry from secondary effects, causing transparent surfaces to be ignored or entangled into a single depth layer.

PGSR [3] extends planar-aligned Gaussian primitives with additional geometric constraints. Despite achieving sharper surface reconstruction, it shares the same opacity-based rendering pipeline and therefore struggles with multi-depth radiance, often collapsing glass regions into the background.

Ref-GS [21] models view-dependent appearance through directional factorization, enabling more expressive specular materials. However, its formulation implicitly assumes that all non-Lambertian behavior arises from surface reflection, without accounting for light transmission. As a result, transparent materials whose appearance is governed by both interface reflectance and background transmission are incorrectly treated as purely reflective surfaces, producing biased material estimates and degraded geometric cues in regions where transmitted radiance is dominant.

EnvGS [18] models environment radiance using a dedicated set of Gaussians and performs ray-traced reflection queries, supported by a monocular normal prior to stabilize geometry estimation from limited viewpoints. While highly effective for opaque materials, it lacks any mechanism to account for light interactions through transparent surfaces, leading background content seen through glass to be incorrectly attributed to reflections.

TSGS [12] targets transparent objects using first-surface rasterization combined with monocular normal [19], delighting [19], and segmentation priors [16]. While effective for thin, object-centric transparency, its formulation assumes a single transparent interface and does not explicitly model transmitted radiance. Consequently, scenes containing multiple depth layers (e.g., glass–background–interior structures) often exhibit blurred transmission and incomplete geometry reconstruction. The segmentation module also produces noisy or missing transparency masks, an issue we further analyze in the next section.

Overall, existing baselines either (i) emphasize geometric fidelity, (ii) specialize in reflective appearance modeling, or (iii) operate under object-centric transparency assumptions. None provide a unified framework capable of addressing the inherently ill-posed nature of scene-scale transparency, where accurate reconstruction requires jointly modeling interface geometry, background transmission, and reflection. Our study bridges this gap by introducing a decomposed Gaussian representation and transparency-aware radiance transport, which together provide a more physically consistent formulation for scene-scale transparency.

C. Additional Implementation Details

In this section, we provide comprehensive implementation details for reproducibility, including initialization process, multi-stage optimization, and the specific loss formulations designed to ensure physical plausibility and geometric consistency. Our code will be made publicly available.

Initialization. To establish a reliable geometric foundation, we initialize the interface Gaussian ($\mathcal{G}_{\text{intr}}$) and transmission Gaussian ($\mathcal{G}_{\text{trans}}$) primitives using the sparse point cloud derived from Structure-from-Motion (SfM) [6]. Meanwhile, we follow EnvGS [18] for initializing reflectance component ($\mathcal{G}_{\text{refl}}$), where we partition the scene into N^3 sub-grids by partitioning the bounding box. Then we randomly sample K primitives within the grid where we set $N = 32$ and $K = 5$.

Optimization Schedule. To ensure stable training, we adopt a multi-stage optimization strategy that progressively recovers scene geometry before refining complex radiance effects. We begin with a 5k-iteration warm-up stage in which only the interface component is optimized, allowing the primary surface geometry to converge. Afterward, the transmission and reflection components are introduced for joint optimization. Finally, between 40k and 60k iterations, we freeze the interface Gaussians and update only the transmission and reflection Gaussians to refine their radiance behavior without destabilizing the established geometry.

Regularization. For geometric regularization, we adopt a scale-and-shift–invariant depth loss together with a normal consistency loss, following the formulation of MonoSDF [20]. Let z denote the rendered depth from the

Figure 2. **Qualitative ablation study on representation components.** Visual comparison between the full model and ablated variants.

interface component and $\hat{z}$ the monocular depth prior predicted by the encoder [13]. We align z and $\hat{z}$ by solving for the optimal scale w and shift q in closed form, and define the depth loss as:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N}\sum_{i=1}^{N}\left(w\, z_i + q - \hat{z}_i\right)^2, \tag{1}$$

where N denotes the number of pixels in the image. This formulation removes the global scale ambiguity of monocular predictions while preserving their relative depth structure. The scale–shift parameters (w, q) are recomputed for each image using the closed-form least-squares solution [20].

For the normal loss, we follow the thresholded normal supervision strategy introduced in TSGS [12] to mitigate the influence of noisy monocular priors. Given the rendered normal $\mathbf{n}$ and the normal prior from [13] $\hat{\mathbf{n}}$, we apply a cosine-similarity mask from 10k iterations:

$$\mathbf{M}_{\text{prior}}(u) = \left[\langle \mathbf{n}(u), \hat{\mathbf{n}}(u)\rangle \geq \tau_{\text{n}}\right],$$

and compute the masked normal loss as:

$$\mathcal{L}_{\text{normal}} = \sum_{u} \mathbf{M}(u)\big(1 - \langle \mathbf{n}(u), \hat{\mathbf{n}}(u)\rangle\big).$$

We set $\tau_{\text{n}} = 0.3$ for all experiments.

D. Qualitative Ablation Studies

In this section, we present additional qualitative ablation results to further elucidate the distinct roles of each component within the GLINT framework.

First, we examine the impact of our decomposed representation in Fig. 2. Removing the transmission branch ($\mathcal{G}_{\text{trans}}$) leads to significant entanglement between the interface and background content, resulting in geometric inconsistencies and a washed-out appearance in regions observed behind glass. Excluding the reflection branch ($\mathcal{G}_{\text{refl}}$) diminishes specular cues, leading to a loss of realistic surface gloss, particularly on glass or highly reflective surfaces.

Next, we visualize the ablation of geometry regularization losses in Fig. 3. While our framework remains relatively robust, removing specific losses introduces characteristic degradations that highlight their individual contributions. Excluding the depth loss ($\mathcal{L}_{\text{depth}}$) causes inaccuracies in interface geometry placement, manifesting as deviations in absolute depth, especially for transparent surfaces. Removing the normal loss ($\mathcal{L}_{\text{normal}}$) results in less consistent surface orientation, which appears as locally unstable shading and noisy normal transitions. When all geometric priors are removed ($\mathcal{L}_{\text{geo}}$), these artifacts accumulate, leading to noticeable distortions in depth and normal maps. This indicates that the geometric constraints operate complementarily to maintain structural fidelity.

Finally, we analyze the effect of the transparency bootstrapping loss ($\mathcal{L}_{\text{trans}}$) in Fig. 4. Omitting this loss yields noisier and spatially inconsistent transparency estimates. This is because the loss utilizes 3D geometric cues derived from depth separation to guide the optimization toward sharp and physically consistent transparency boundaries.

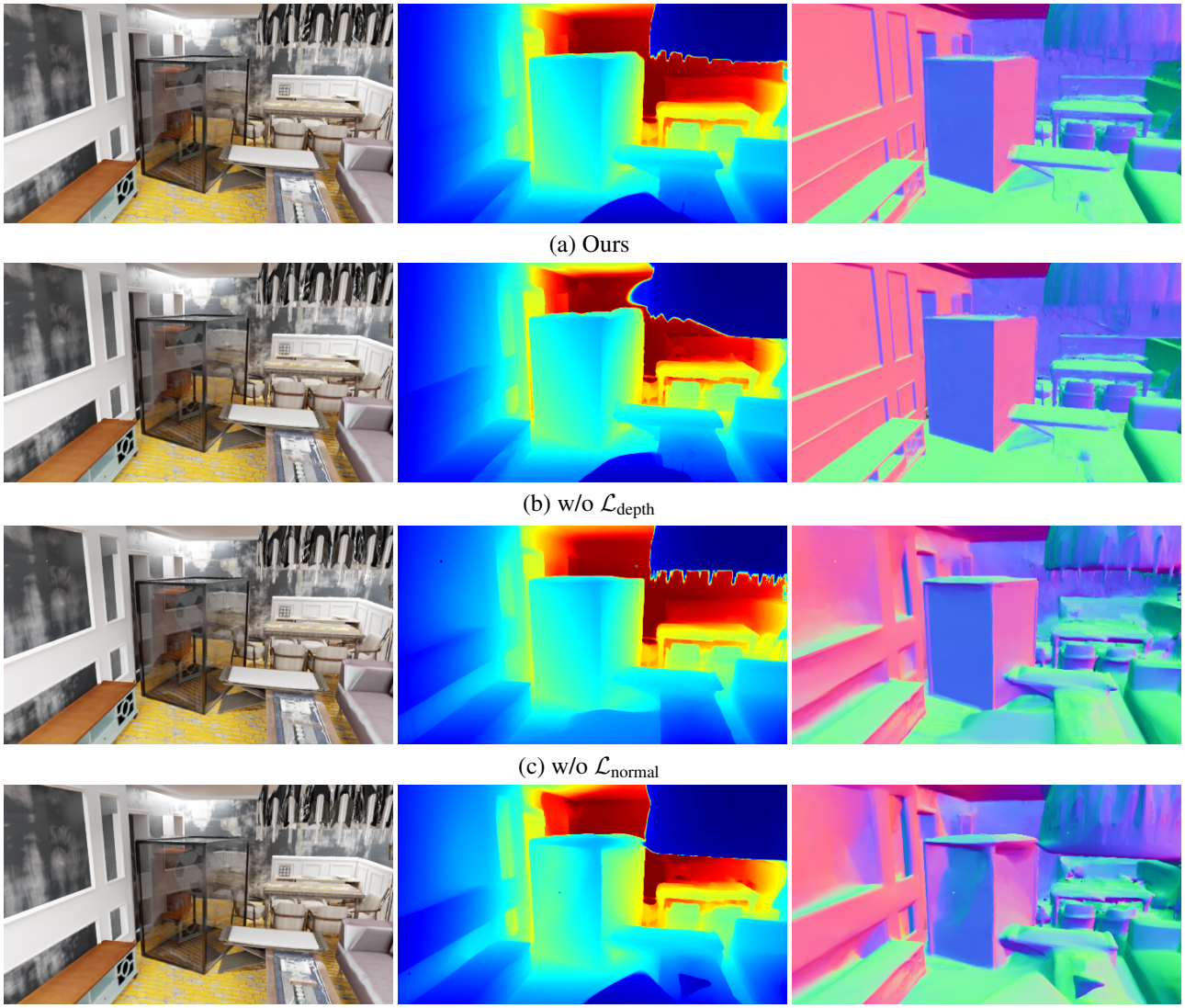
E. Comparative Analysis with Baselines

In this section, we provide a comprehensive comparative analysis against baseline methods. We focus on three key aspects: the interpretability of radiance decomposition, the accuracy of transparency localization, and a quantitative analysis of computational costs versus reconstruction quality.

Radiance decomposition. Fig. 5 presents an extended comparison of radiance decomposition between our method and EnvGS [18].

For our method (Fig. 5(a)), the decomposed Gaussian representation enables a clear and physically grounded partitioning of radiance. The interface component isolates the surface base color, while the reflection component captures specular highlights. Crucially, the transmission component reconstructs the background scene visible specifically through transparent surfaces, revealing the correct spatial structure behind the glass. These components form a coherent explanation of the observed radiance, demonstrating that our explicit decomposition naturally disentangles overlapping radiance sources in transparent regions.

In contrast, EnvGS (Fig. 5(b)) focuses solely on modeling view-dependent reflections via environment Gaussians and lacks a dedicated transmission component. Consequently, radiance originating from behind transparent sur-



(a) Ours

(b) w/o $\mathcal{L}_{\text{depth}}$

(c) w/o $\mathcal{L}_{\text{normal}}$

(d) w/o $\mathcal{L}_{\text{geo}}$

Figure 3. **Effect of geometric losses.** Ablating geometric supervision leads to degraded geometry reconstruction. Removing $\mathcal{L}_{\text{depth}}$ produces inaccurate interface depth, removing $\mathcal{L}_{\text{normal}}$ results in unstable surface orientation, and removing all geometric losses ($\mathcal{L}_{\text{geo}}$) severely distorts both depth and normals.

faces is incorrectly entangled within its diffuse or reflection modeling, leading to mixed or incomplete visual explanations. Furthermore, its reflection strength is modulated by a single scalar weight s, lacking the physically accurate Fresnel-driven angular dependence inherent to our formulation.

To further validate the necessity of our decomposition, we conducted an experiment training EnvGS [18] with the geometric normal priors from DiffusionRenderer [13], identical to our setup. As illustrated in Fig. 6, enforcing geometric consistency on EnvGS paradoxically leads to corrupted appearance rendering. This degradation occurs because EnvGS lacks a dedicated transmission component and fundamentally conflates transmitted and reflected radiance into a single surface interaction. Typically, EnvGS implicitly minimizes photometric error by distorting the geometry to effectively baking background textures onto incorrect depths. However, when the surface geometry is enforced via stronger priors, this compensatory mechanism is blocked. Consequently, the model is forced to approximate the superposition of multiple radiance layers onto a single interface, leading to an averaging effect that manifests as severe blurring.

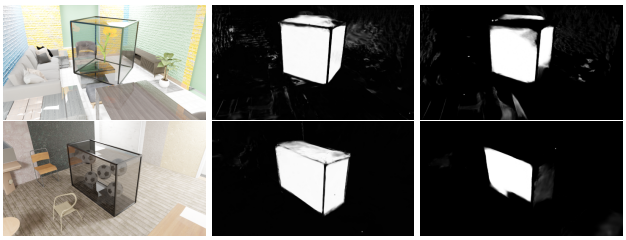


Figure 4. **Effect of the transparency loss $\mathcal{L}_{\text{trans}}$.** With the proposed transparency loss $\mathcal{L}_{\text{trans}}$, the learned transparency maps align well with the true transmitted content, while removing this loss leads to noisy or inconsistent transparency estimation.

Transparency Map Comparison. We compare our learned transparency maps with those generated by Grounded-SAM-2 (G-SAM2) [16], a state-of-the-art open-world segmentation model. As illustrated in Fig. 7, although G-SAM2 can roughly localize transparent objects, it frequently yields spatially inconsistent or fragmented masks. Common failure cases include missing sections of glass cabinets or exhibiting severe flickering across viewpoints, even when utilizing the tracking mode. These limitations arise because G-SAM2 relies on 2D image features, which are inherently ambiguous for transparent surfaces that mix reflection and transmission, lacking a unified understanding of 3D geometry.

In contrast, GLINT bootstraps transparency localization by explicitly leveraging 3D geometric cues—specifically, the depth discrepancy (Δz) between the interface and transmission components that emerges during optimization and the diffuse-albedo prior from [13]. This allows our method to generate spatially coherent and boundary-sharp transparency maps that accurately align with the physical extent of the glass.

Computational Costs. We report a runtime comparison with EnvGS and TSGS on the synthetic 3D-FRONT-T dataset (downscale $\times 2$), with results summarized in Tab. B. TSGS achieves the highest speed due to rasterization-only rendering, while EnvGS adopts a hybrid formulation limited to reflection. Our method introduces additional overhead from explicit multi-component optimization and transmission handling, resulting in lower throughput. However, this design choice directly supports transparent geometry modeling and leads to improved reconstruction quality, reflecting a deliberate and practical trade-off between efficiency and capability.

F. Discussion on Foundation Models

In this section, we discuss the motivation behind integrating foundation models into the GLINT framework and analyze the advantages of video-based priors compared to conventional image-based approaches in the context of transparent scene reconstruction.

Table 1. Comparison of computational costs and reconstruction quality on the 3D-FRONT-T dataset.

Method	FPS (↑)	Training Time (↓)	PSNR (↑)	MAE (↓)	AbsRel (↓)
TSGS [12]	**159**	**∼40 mins**	28.80	9.89	0.08
EnvGS [18]	80	∼1 h	33.71	14.37	0.13
Ours	51	∼2.5 h	**34.50**	**7.96**	**0.04**

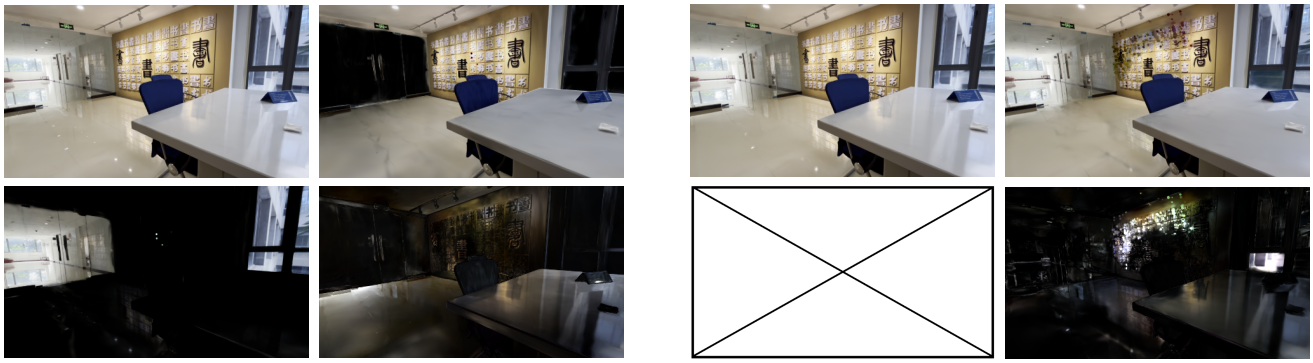
In our implementation, we utilize the inverse-rendering encoder of the video relighting model, DiffusionRenderer [13], to obtain geometric (depth, normal) and material (diffuse-albedo) priors. A key motivation for choosing a video-based foundation model over monocular estimators is its multi-view consistency. Since the model leverages the architecture of Stable Video Diffusion [1], it processes multiple frames in a single feed-forward pass, producing geometric cues that are coherent across viewpoints. This is particularly crucial for transparent surfaces, where per-frame estimation often flickers or yields inconsistent depth due to view-dependent reflections. To the best of our knowledge, GLINT is the first approach to adopt video relighting priors for 3D Gaussian splatting reconstruction, demonstrating that material-aware priors effectively aid in distinguishing between interface and transmitted radiance.

Comparison with Image-based Priors. Recent approaches [12, 17, 18] typically combine various image-based foundation models, such as monocular depth [2, 8], normal estimators [10, 19], and segmentation modules (e.g., Grounded-SAM [11, 16]). For instance, TSGS [12] relies on Grounded-SAM to mask transparent objects and uses StableDelight and StableNormal [19] to make the texture of the transaprent surface dinstinctive. While empirically beneficial for object-centric scenes with clear boundaries, this mask strategy often struggles with scene-scale transparency—such as large glass facades or windows—where semantic boundaries are ambiguous and binary segmentation fails to capture the continuous transition of radiance as in Fig. 7.

G. Limitations and Future Work

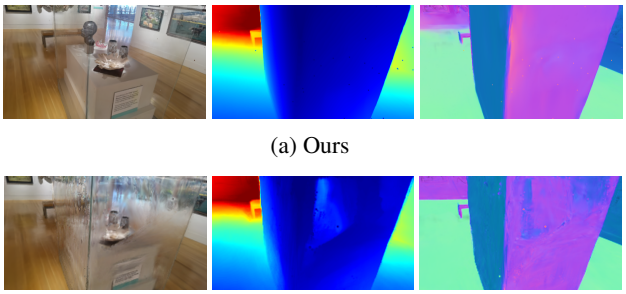
While GLINT achieves state-of-the-art performance in reconstructing scene-scale transparent geometry and appearance, there are still room for improvement in various perspectives. In this section, we detail the discussion on the current boundary conditions of our framework and suggest potential directions to address them.

Decomposition ambiguity under sparse observations. Our method implicitly relies on multi-view consistency to



(a) Ours (b) EnvGS [18]

Figure 5. **Radiance decomposition comparison.** Each 2×2 grid is arranged in clockwise order as rendered RGB, base color, reflection and transmission. EnvGS does not provide a transmission component, so the corresponding slot is left blank.



(a) Ours

(b) EnvGS [18] w/ DiffusionRenderer [13]

Figure 6. **Comparison with EnvGS [18] trained with a DiffusionRenderer [13] prior.** With the DiffusionRenderer normal priors, EnvGS still fails to produce consistent depth and normals for transparent regions. In particular, the transmission surface remains incorrectly reconstructed, showing blurry rendering quality.

disentangle the intertwined radiance contributions from interface reflection and background transmission. Consequently, in scenarios with sparse viewpoints or limited parallax, where a surface is observed from a stationary angle, the problem becomes inherently ill-posed. In such cases, the optimization may struggle to uniquely assign radiance to either the reflection ($\mathcal{G}_{\text{refl}}$) or transmission ($\mathcal{G}_{\text{trans}}$) component. We believe that incorporating high-level semantic understanding could resolve this ambiguity. Future work could leverage vision-language models (VLMs) or unprojecting semantic features to enforce physically and semantically plausible decomposition, ensuring that regions identified as glass (e.g., windows vs. mirrors) adhere to their expected optical behaviors even under constrained observation.

Recursive light transport. To maintain rendering efficiency and optimization stability, our current radiance transport formulation focuses on primary transmission and reflection events (i.e., up to first-order interactions at the interface). While this approximation is sufficient for most architectural glass and display cases, it may not fully capture complex multi-bounce phenomena found in nested transparent structures, such as a glass vase inside a glass cabinet or a mirror room. Extending our hybrid rendering pipeline to support recursive ray-tracing or multi-pass rendering with physically-based rendering formulation would allow for the simulation of higher-order approximation of light transport, albeit at the cost of increased computational overhead.

H. Additional Qualitative Results

We present additional qualitative examples to complement the results in the main paper. Figures 8 and 9 provide further visualizations on representative scenes from both the 3D-FRONT-T benchmark and the real-world DL3DV-10K dataset [14].

We also include the video results for the visualization of the continuous frames. We kindly refer the readers to the attached supplementary videos.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 5

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025. 5

[3] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *arXiv preprint arXiv:2406.06521*, 2024. 2, 7, 8

[4] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl,

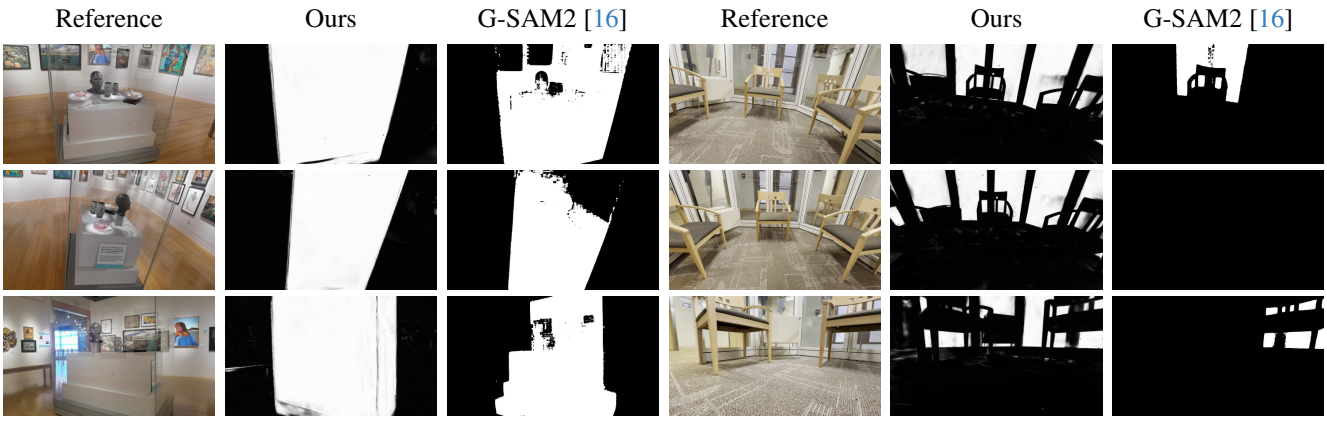


Figure 7. Comparison of transparency masks across two scenes. Each scene is shown with three viewpoints (rows), including ground-truth reference RGB (left), our predicted transparency maps (middle), and G-SAM2 [16] masks (right). GT images provide visual context, highlighting that our method consistently isolates transparent surfaces, while Grounded-SAM2 often produces noisy, incomplete or completely missing masks.

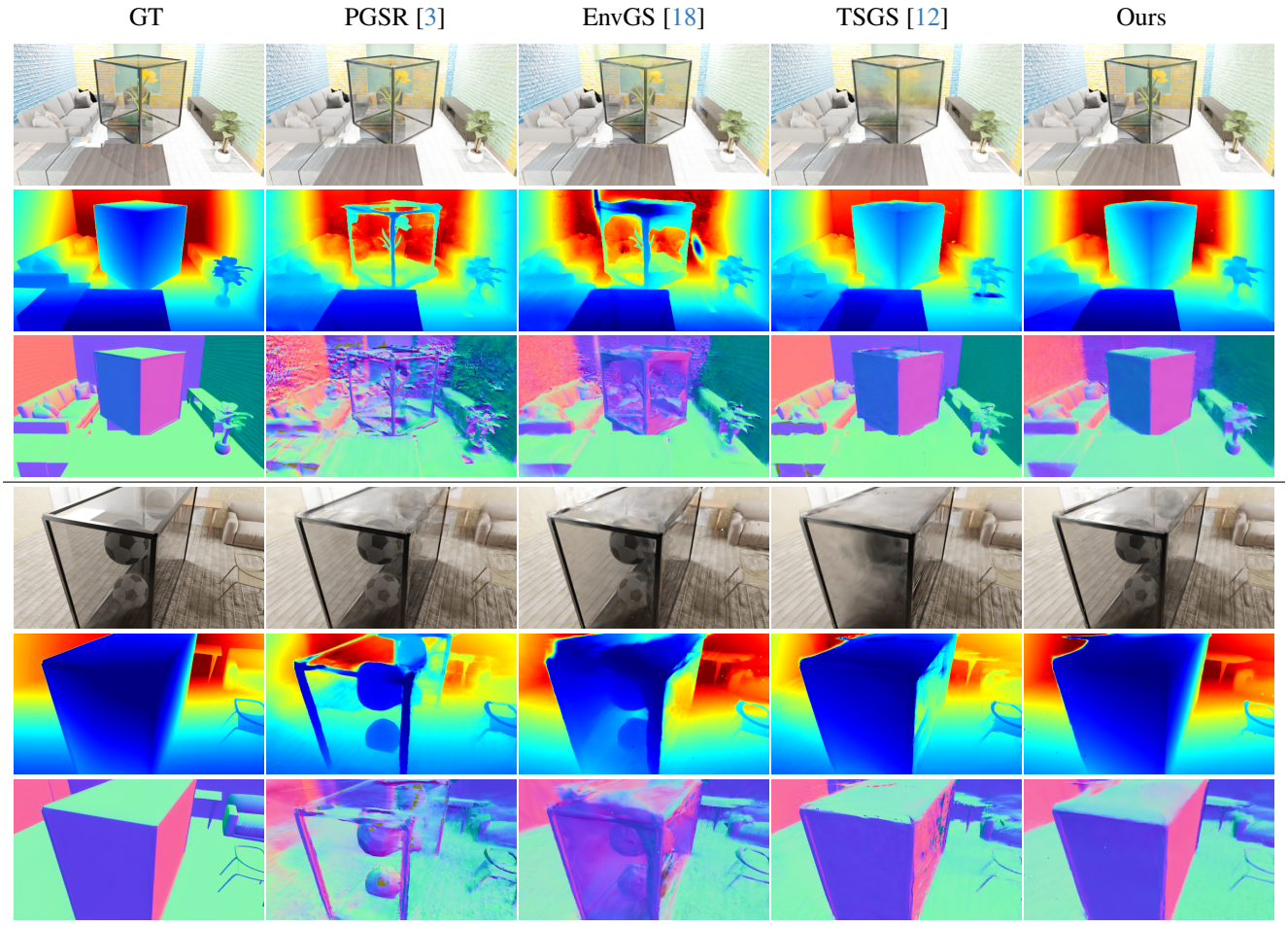


Figure 8. **Qualitative comparison on synthetic scenes.** Each column shows results from GT, PGSR [3], EnvGS [18], TSGS [12], and Ours. For each scene, rows correspond to RGB (top), depth (middle), and normal (bottom) maps.

Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 2

[5] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming

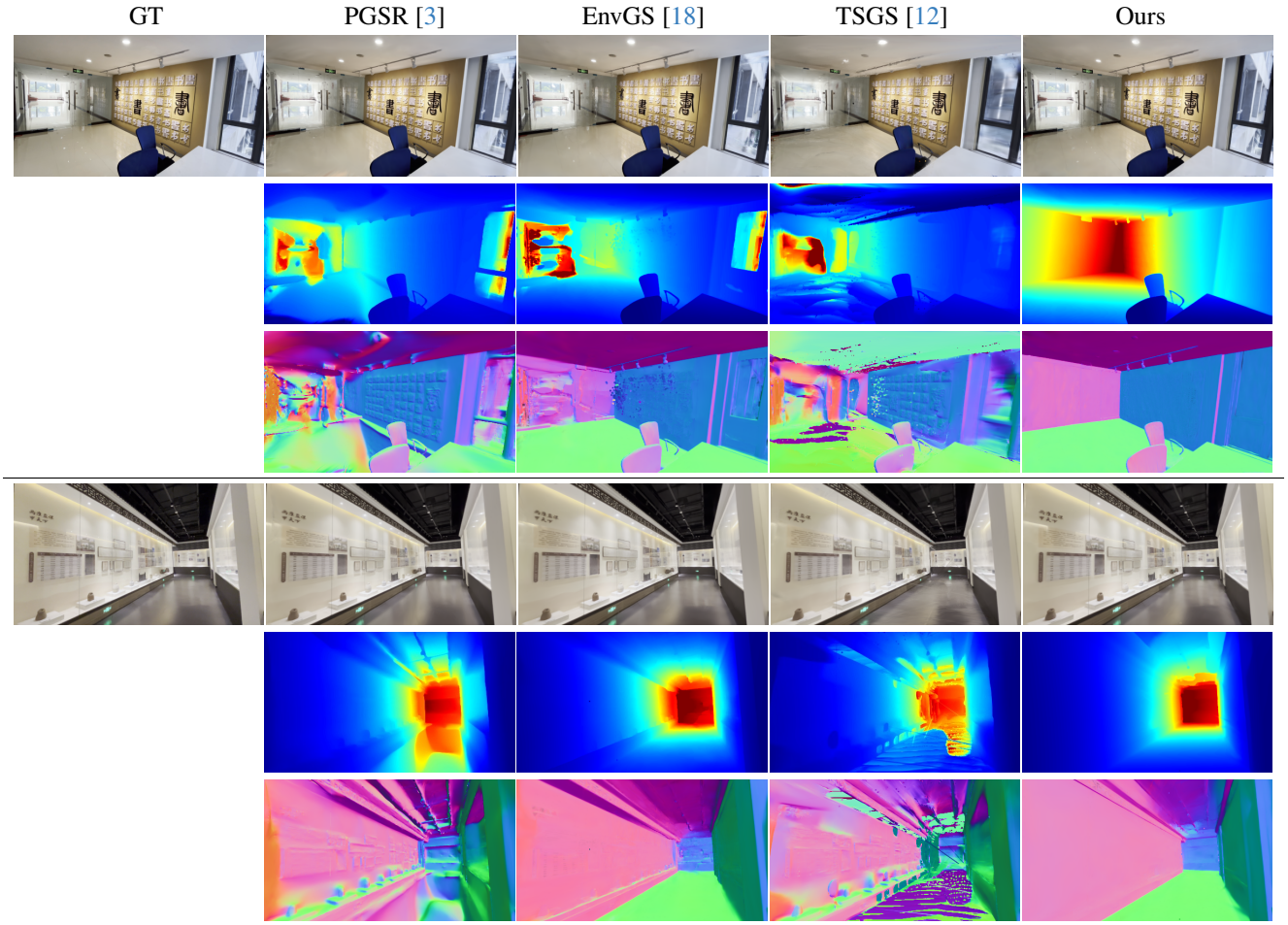


Figure 9. **Additional qualitative comparisons on the DL3DV-10K dataset.** Each column shows results from GT, PGSR [3], EnvGS [18], TSGS [12], and Ours. For each scene, rows correspond to RGB (top), depth (middle), and normal (bottom) predictions.

Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, pages 10933–10942, 2021. 1

[6] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *CVPR*, pages 20796–20805, 2024. 2

[7] Roland Hess. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press, 2010. 1

[8] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *PAMI*, 2024. 5

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

[10] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 5

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5

[12] Mingwei Li, Pu Pang, Hehe Fan, Hua Huang, and Yi Yang. Tsgs: Improving gaussian splatting for transparent surface reconstruction via normal and de-lighting priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7220–7229, 2025. 2, 3, 5, 7, 8

[13] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *CVPR*, pages 26069–26080, 2025. 3, 4, 5, 6

[14] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024. 1, 6

[15] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. In *CVPR*, pages 792–802, 2023. 1

[16] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 5, 7

[17] 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 *CVPR*, pages 21547–21557, 2025. 5

[18] 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 *CVPR*, pages 5742–5751, 2025. 2, 3, 4, 5, 6, 7, 8

[19] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM TOG*, 2024. 2, 5

[20] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 35:25018–25032, 2022. 2, 3

[21] Youjia Zhang, Anpei Chen, Yumin Wan, Zikai Song, Junqing Yu, Yawei Luo, and Wei Yang. Ref-gs: Directional factorization for 2d gaussian splatting. In *CVPR*, pages 26483–26492, 2025. 2