REFLECTWORLD: SEEING TWO WORLDS THROUGH GLASS VIA MULTI-VIEW GEOMETRY-ROUTED DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image reflection separation aims to recover the transmission and reflection layers from mixed observations captured through glass. Recent advances in large-scale pretrained models have substantially improved reflection recovery, yet most existing methods remain limited to single-image inference, where severe layer entanglement leaves the decomposition fundamentally ambiguous. Multi-view observations offer a natural way to reduce this ambiguity, as transmission is largely consistent across viewpoints whereas reflection is more view-dependent. However, directly exploiting such cross-view cues is difficult because reflections themselves corrupt the geometric correspondences required to establish reliable associations across views. To address this, we propose **ReflectWorld**, a geometry-routed diffusion framework that employs GeoRCA to establish reliable cross-view associations from corrupted multi-view geometry, while MGR-Diff drives differential generation with the distinct cross-view variations of the two layers. In addition, we introduce **MVRefSep**, a multi-view reflection separation dataset that captures the physical characteristics of reflections across views, establishing a benchmark for evaluating both transmission and reflection recovery. Experiments demonstrate that ReflectWorld achieves state-of-the-art performance. The code and dataset will be publicly released.

1 INTRODUCTION

When imaging through glass, a camera captures both the transmitted scene behind the glass and the reflected scene on its surface, resulting in an observed image that is a mixture of two visual layers. Given a mixed observation $\boldsymbol{I}$, image reflection separation aims to recover its transmission layer $\boldsymbol{T}$ and reflection layer $\boldsymbol{R}$. Specifically, $\boldsymbol{T}$ preserves visual information from the scene behind the glass, while $\boldsymbol{R}$ retains scene content formed along another imaging path. Both layers contain meaningful visual information that can support downstream tasks such as semantic segmentation Ravi et al. (2025), geometric reconstruction Wang et al. (2025), and scene understanding Zhu et al. (2025).

In recent years, CNN-based Hu & Guo (2023); Zhao et al. (2025), Transformer-based Hu et al. (2024); Lee et al. (2026), and Mamba-based Fang et al. (2026); Yin et al. (2026) methods have advanced single-image reflection removal. However, when $\boldsymbol{T}$ and $\boldsymbol{R}$ are heavily entangled, the limited information available from a single image makes it difficult to determine the layer assignment of content. Multi-view observations provide additional cues for alleviating this ambiguity. Some methods establish correspondences across views using optical flow Yang et al. (2016), disparity Li et al. (2021) or motion aggregation Prasad et al. (2023). However, these approaches are vulnerable to reflection interference or complex textures, leading to erroneous registration and separation failure. Other studies employ NeRF Guo et al. (2022); Zhu et al. (2022) or 3D Gaussian Splatting Song et al. (2025) to reconstruct clean 3D scenes from reflection-corrupted multi-view observations. These approaches focus on reorganizing content already present in the input images, making it difficult to recover content that has been degraded by reflections.

Large-scale pretrained models provide powerful generative priors for restoring reflection-degraded content. GFRRN Chen et al. (2026) adapts pretrained models to reflection separation through efficient fine-tuning, while JLD Huang et al. (2026) leverages diffusion models for joint layer generation.

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.

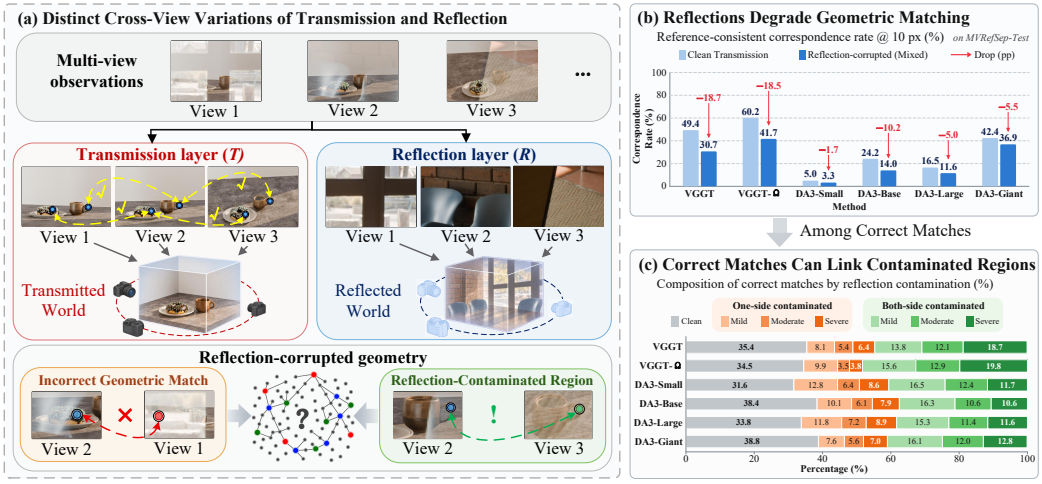


Figure 1: Distinct cross-view properties and geometric challenges in multi-view reflection separation. Reflections interfere with geometric matching, while even geometrically correct correspondences can link reflection-contaminated regions.

However, these methods rely on the appearance distribution of a single image, lacking awareness of cross-view variation patterns, causing visual content to be easily assigned to the wrong layer.

In fact, $\boldsymbol{T}$ and $\boldsymbol{R}$ exhibit distinct cross-view properties, as shown in Fig. 1(a). $\boldsymbol{T}$ originates from the scene behind the glass, whose spatial structure remains stable across viewpoints, resulting in strong cross-view consistency. In contrast, $\boldsymbol{R}$ arises from a different imaging path, changing dynamically with the viewing position, exhibiting stronger view dependence. These distinct cross-view behaviors provide discriminative cues for separating the two layers. However, exploiting these cues is nontrivial because reflections corrupt the geometric relationships established across views. On the one hand, reflections lead to incorrect geometric matches, as shown in Fig. 1(b); on the other hand, even geometrically correct correspondences may still link regions contaminated by reflections, as shown in Fig. 1(c). This leads to a challenge: how can reflection-corrupted multi-view geometry be turned into reliable separation evidence that drives differential generation of $\boldsymbol{T}$ and $\boldsymbol{R}$? To address this, we propose ReflectWorld, a geometry-routed diffusion framework that tackles two problems.

(1) How can reflection-corrupted multi-view geometry be turned into reliable separation evidence? Feed-forward 3D models Wang et al. (2025); Lin et al. (2025) can establish spatial relationships across views, but reflection interference makes the resulting geometric candidates unreliable. Geometry can provide only a coarse search space, but cannot determine which cross-view content is reliable to participate in generation. To address this, we design Geometry-Routed Cross-View Association (GeoRCA), leveraging multi-view geometry to constrain the cross-view candidate space, establishing associations within this space to identify reliable cross-view consistency. In this way, uncertain geometric correspondences are converted into separation evidence in the diffusion space: 3D geometry determines *"where to search"*, while GeoRCA determines *"what to trust"*.

(2) How can the cross-view variation patterns of $\boldsymbol{T}$ and $\boldsymbol{R}$ be leveraged to drive differential layer generation? Existing diffusion-based methods Huang et al. (2026); Wang et al. (2026b) do not explicitly distinguish how content behaves across views, i.e., whether it is cross-view-consistent or view-dependent, easily leading to content misassignment between the two generation branches. To address this, we introduce Multi-View Geometry-Routed Diffusion (MGR-Diff), incorporating explicit layer discrimination into the encoding and decoding stages of the diffusion network. During encoding, multi-view geometric priors are injected into diffusion to endow the generative representation with structural awareness. During decoding, cross-view consistency evidence modulates the two branches in opposite directions, encouraging cross-view-consistent content to be assigned to $\boldsymbol{T}$ while preserving view-dependent content in $\boldsymbol{R}$, achieving layer disentanglement.

Existing reflection datasets Zhu et al. (2024); Cai et al. (2025) focus on single-image settings. The few multi-view reflection datasets Guo et al. (2022); Zhu et al. (2022) are small-scale and focus on

reflection removal, providing only Ground-Truth $\boldsymbol{T}$ (GT-$\boldsymbol{T}$) without Ground-Truth $\boldsymbol{R}$ (GT-$\boldsymbol{R}$). This leaves multi-view reflection separation insufficiently supported by dedicated data. Building such data is challenging because obtaining GT-$\boldsymbol{T}$ and GT-$\boldsymbol{R}$ across multiple real-world viewpoints is difficult, while 2D transformations applied to individual images cannot reproduce cross-view variations under real camera motion. To address these limitations, we introduce the Multi-View Reflection Separation (MVRefSep) dataset, which follows the physical principles of reflection under multi-view imaging and covers diverse environments, establishing a benchmark for multi-view reflection separation.

In summary, our main contributions are as follows:

- We propose ReflectWorld, the first diffusion-based multi-view framework for joint transmission and reflection recovery, enabling effective layer separation by integrating multi-view geometry with generative priors.
- We introduce GeoRCA, which converts reflection-corrupted multi-view geometry into reliable cross-view separation evidence, improving the robustness of cross-view association under reflection interference.
- We develop MGR-Diff, which exploits the distinct cross-view behaviors of transmission and reflection to drive differential generation, leading to better disentanglement of the two layers.
- We construct MVRefSep, a multi-view reflection separation benchmark with simulated and real-world data, enabling quantitative evaluation of both-layer recovery and real-world generalization.

2 Related Work

Single-image reflection removal. Early methods relied on hand-crafted priors, such as relative smoothness Li & Brown (2014) and ghosting cues Shih et al. (2015), but were limited to simple imaging conditions. Deep learning methods later learned the decomposition of $\boldsymbol{T}$ and $\boldsymbol{R}$ directly from data, using CNN Hu & Guo (2023); Zhao et al. (2025), Transformers Fang et al. (2025); Lee et al. (2026), or Mamba architectures Yin et al. (2026); Fang et al. (2026). More recently, pretrained models have been introduced to provide stronger priors: GFRRN Chen et al. (2026) adapts pretrained representations to reflection separation, while DAI Hu et al. (2026), GenSIRR Li et al. (2026), PRISM Shin & Kim (2026), and JLD Huang et al. (2026) exploit diffusion priors for transmission restoration or joint layer generation. However, when $\boldsymbol{T}$ and $\boldsymbol{R}$ are heavily entangled, a single image is often insufficient to resolve the ambiguity between the two layers.

Multi-image reflection removal. Multi-image observations provide additional cues for reflection separation. Early methods estimate layer motion using optical flow Yang et al. (2016), disparity Li et al. (2021) or motion aggregation Prasad et al. (2023). However, their reliance on pixel-level correspondences makes them vulnerable to reflection interference and complex textures. S2R Wang et al. (2026b) exploits consecutive video frames for dereflection without explicitly modeling cross-frame geometry, while dual-view diffusion methods Xu et al. (2026) leverage cross-view interaction but neglect layer-dependent variations. Moreover, both focus on transmission recovery rather than joint recovery of the transmission and reflection layers. With the development of neural rendering, NeRFReN Guo et al. (2022) and TNeRF Zhu et al. (2022) reconstruct clean scenes from reflection-corrupted observations using NeRF, while RefUnlock Song et al. (2025) and GlassGS Cui et al. (2026) extend this reconstruction paradigm to 3D Gaussian Splatting. However, these methods rely on per-scene optimization and target reflection-free reconstruction rather than feed-forward recovery of both layers.

Diffusion models for low-level vision. Diffusion models have been widely applied to low-level vision tasks due to their strong natural image priors, including super-resolution Yang et al. (2026), dehazing Liu et al. (2025), and low-light enhancement Lan et al. (2025). GDP Fei et al. (2023) uses pretrained diffusion models for general image restoration. DiffIR Xia et al. (2023) improves efficiency through diffusion in a compact prior space, while UCDIR Zhang et al. (2023) enhances conditional guidance for restoration. SIR-DIFF Mao et al. (2025) extends diffusion restoration to multiple views, but directly performs global interaction among tokens from all views without distinguishing their cross-view relationships, limiting its ability to disentangle different visual layers.

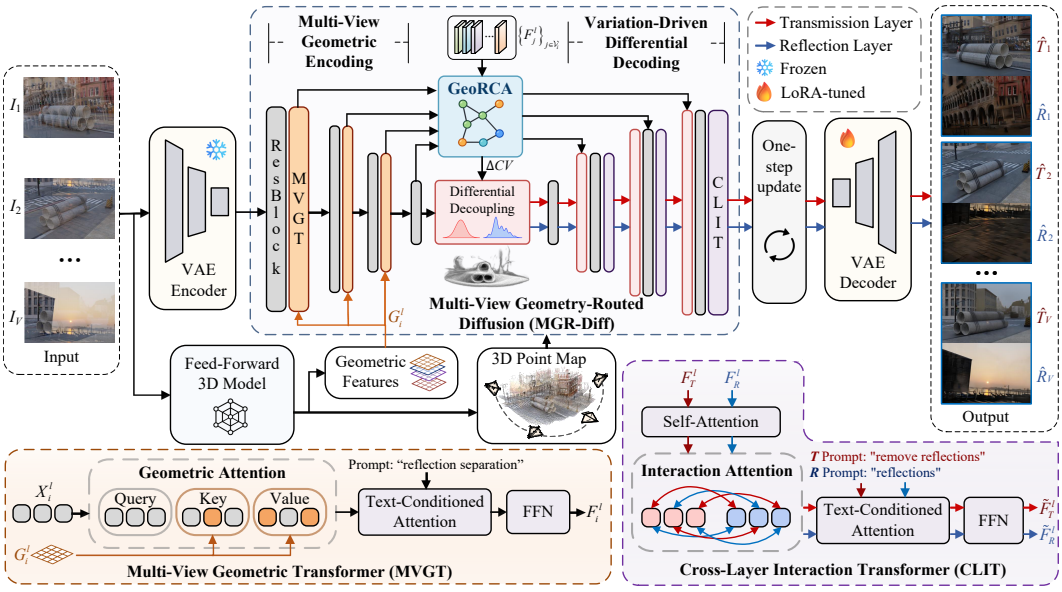


Figure 2: Overview of ReflectWorld. Multi-view geometry is integrated into the diffusion process, with GeoRCA establishing cross-view associations, while MGR-Diff exploits distinct cross-view variations to drive differential generation of transmission and reflection layers.

3 METHOD

3.1 OVERVIEW

ReflectWorld consists of a Feed-Forward 3D Model and MGR-Diff, as illustrated in Fig. 2. Given a set of V mixed observations $\mathcal{I} = \{I_i\}_{i=1}^{V}$, a frozen VAE encoder maps the observations into the latent space, while the feed-forward 3D model jointly extracts multi-view geometric priors. MGR-Diff then leverages these priors to establish reliable cross-view associations and guide layer separation. A one-step diffusion update predicts the two layer-specific latents, which are mapped back to the image space by the VAE decoder to obtain the transmission $\widehat{T}_i$ and reflection $\widehat{R}_i$ for each input view.

3.2 MULTI-VIEW GEOMETRY-ROUTED DIFFUSION

MGR-Diff exploits the distinct cross-view variation patterns of transmission and reflection for differential generation of the two layers. As shown in Fig. 2, MGR-Diff consists of two main stages: Multi-View Geometric Encoding and Variation-Driven Differential Decoding.

Multi-View Geometric Encoding. This stage injects geometric priors into diffusion features to provide geometry-aware representations for subsequent cross-view association. Specifically, we employ a Multi-View Geometric Transformer (MVGT). For the i-th view, let X_i^l and G_i^l denote the diffusion and geometric features at the l-th encoder scale, respectively. MVGT uses X_i^l as the query and conditions the keys and values on G_i^l:

$$Q_{g,i}^l = W_Q^g X_i^l, \quad K_{g,i}^l = W_K^g \Phi(X_i^l, G_i^l), \quad V_{g,i}^l = W_V^g \Phi(X_i^l, G_i^l). \tag{1}$$

where $\Phi(\cdot)$ denotes feature fusion. Geometric Attention is then applied to obtain the geometry-aware feature $A_{g,i}^l = \mathrm{Attn}(Q_{g,i}^l, K_{g,i}^l, V_{g,i}^l)$, which carries the injected geometric cues and is propagated to subsequent layers.

Variation-Driven Differential Decoding. This stage constructs different update directions to facilitate the separation of the two layers. At decoder scale l, GeoRCA establishes cross-view content associations and produces a cross-view consistency correction $\Delta_{\mathrm{cv},i}^l$. Differential Decoupling then

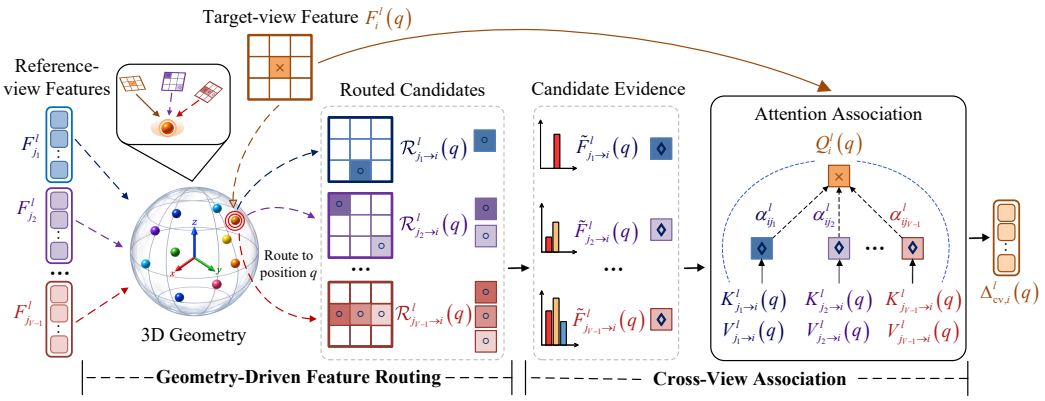


Figure 3: Geometry-Routed Cross-View Association (GeoRCA). Given a target feature $F_i^l(q)$, Geometry-Driven Feature Routing uses 3D geometry to route reference-view features to the target location q, forming a geometry-constrained candidate space $\mathcal{R}_{j\to i}^l(q)$. Cross-View Association further evaluates the routed candidate evidence $\widetilde{F}_{j\to i}^l(q)$ against the target feature, selectively associating reliable cross-view content to produce the consistency correction $\Delta_{\mathrm{cv},i}^l(q)$.

modulates the two generation branches with opposite signs:

$$F_{T,i}^l = F_i^l + \Delta_{\mathrm{cv},i}^l, \qquad F_{R,i}^l = F_i^l - \Delta_{\mathrm{cv},i}^l. \tag{2}$$

The transmission branch follows the cross-view association direction to reinforce stable content shared across views, whereas the reflection branch suppresses cross-view shared responses in the opposite direction, thereby retaining more view-dependent components.

In addition, we introduce the Cross-Layer Interaction Transformer (CLIT) to enable bidirectional information exchange between the two branches. At scale l, the transmission and reflection features are organized into a unified token sequence:

$$\mathbf{H}_{\mathrm{TR},i}^l \in \mathbb{R}^{B\times 2\times C_l\times H_l\times W_l} \xrightarrow{\text{reshape}} \mathbf{X}_{\mathrm{TR},i}^l \in \mathbb{R}^{B\times(2H_lW_l)\times C_l}. \tag{3}$$

Interaction Attention is then applied to the joint sequence, allowing the two branches to exchange information while maintaining their respective differential generation directions. The resulting features are subsequently processed by their respective Text-Conditioned Attention and Feed-Forward Network (FFN) for refinement.

3.3 Geometry-Routed Cross-View Association

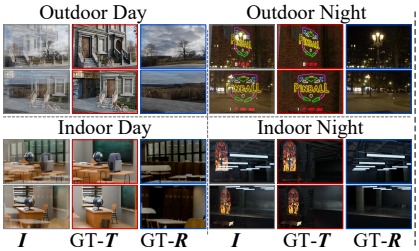
GeoRCA transforms cross-view geometric relationships into content associations in diffusion space. As illustrated in Fig. 3, GeoRCA consists of Geometry-Driven Feature Routing and Cross-View Association.

Geometry-Driven Feature Routing. This stage is responsible for determining which cross-view features can be routed to each target location. Given a target view i, the remaining $V-1$ views are treated as reference views. For the j-th reference view, let $P_j(p)$ denote the 3D point corresponding to position p, and let $c_j(p)$ denote its geometric confidence. At the l-th diffusion scale, the target-view and reference-view features are denoted by $F_i^l, F_j^l \in \mathbb{R}^{C_l\times H_l\times W_l}$, respectively. For the 3D point $P_j(p)$ corresponding to position p in reference view j, we project it onto the target image plane using the camera parameters of target view i:

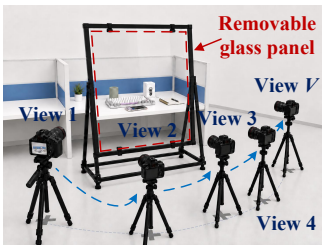
$$p_{j\to i} = \pi(K_i[R_i \mid t_i]P_j(p)), \tag{4}$$

where $\pi(\cdot)$ denotes the camera projection from a 3D point to the 2D image plane. The corresponding reference diffusion feature $F_j^l(p)$ is then routed to the target feature plane according to this geometric relationship. For a target location q, all features from reference view j reaching this location form the routed candidates:

$$\mathcal{R}_{j\to i}^l(q) = \left\{F_j^l(p) \mid p_{j\to i} = q\right\}. \tag{5}$$



(a) MVRefSep-Sim samples



(b) MVRefSep-Real acquisition setup

Dataset	Scenes	Coverage	GT-*T*/*R*
RFFR	6	Indoor	0/0
TNeRF	8	Indoor	4/0
Flash-Splat	7	Indoor/Outdoor	0/0
RealX3D-Reflection	8	Indoor	8/0
MVRefSep (Ours)	**50**	**Indoor/Outdoor; Day/Night**	**50/40**

(c) Dataset comparison

Figure 4: MVRefSep dataset and its comparison with existing datasets.

Therefore, each target location receives a set of candidate features from reference views, with the candidate space constrained by the cross-view 3D geometry.

Cross-View Association. This stage forms reliable candidate evidence from the routed candidates and establishes cross-view content associations. For a target location q, multiple locations in a reference view may be routed to it through geometric projection. These routed features are aggregated using geometric confidence to obtain the candidate evidence for each reference view:

$$\widetilde{F}^l_{j\to i}(q) = \frac{\sum_{p\in S_{j\to i}(q)} c_j(p) F^l_j(p)}{\sum_{p\in S_{j\to i}(q)} c_j(p) + \epsilon}, \tag{6}$$

where $S_{j\to i}(q)$ denotes the set of feature locations in reference view j that are routed to target location q, and ϵ ensures numerical stability. Attention Association is then performed using the current target-view feature at location q as the query, while the candidate evidence from reference view j is mapped to the key and value:

$$Q^l_i(q) = W_Q F^l_i(q), \quad K^l_{j\to i}(q) = W_K \widetilde{F}^l_{j\to i}(q), \quad V^l_{j\to i}(q) = W_V \widetilde{F}^l_{j\to i}(q). \tag{7}$$

Thus, cross-view interaction is restricted to the candidate evidence obtained through geometry-driven routing. To further distinguish the reliability of different candidate correspondences, geometric confidence is introduced as a soft prior into the association score:

$$s^l_{ij}(q) = \big(Q^l_i(q)^\top K^l_{j\to i}(q)\big)/\sqrt{d} + \log(\bar{c}_{j\to i}(q) + \epsilon)\,, \tag{8}$$

where $\bar{c}_{j\to i}(q)$ denotes the aggregated geometric confidence of reference view j at target location q, and d is the feature dimension. The association weight is then obtained as $\alpha^l_{ij}(q) = \mathrm{softmax}_j\big(s^l_{ij}(q)\big)$. Finally, by considering the candidate evidence from all reference views, we obtain the cross-view consistency correction:

$$\Delta^l_{\mathrm{cv},i}(q) = W_O\big[\sum_{j\neq i} \alpha^l_{ij}(q) V^l_{j\to i}(q)\big]. \tag{9}$$

The resulting correction $\Delta^l_{\mathrm{cv},i}(q)$ aggregates cross-view evidence that is geometrically relevant to the target location.

4 MVRefSep: A Multi-View Reflection Separation Dataset

We introduce the MVRefSep dataset for the multi-view reflection separation task, as shown in Fig. 4. It consists of 50 scenes, including 40 simulated scenes (MVRefSep-Sim) and 10 real-world scenes (MVRefSep-Real). MVRefSep-Sim is constructed in Blender, where reflective scenes are rendered from multiple viewpoints using physically based rendering. It covers four categories, i.e., outdoor daytime, outdoor nighttime, indoor daytime, and indoor nighttime, with 10 scenes in each category and 50 views per scene. For each MVRefSep-Sim scene, we provide the camera poses of all views, together with GT-$\boldsymbol{T}$ and GT-$\boldsymbol{R}$, enabling quantitative evaluation of both layers. In addition, we capture 10 real-world scenes in controlled environments using a Canon EOS 90D camera, obtaining the corresponding GT-$\boldsymbol{T}$ with the glass removed to assess generalization under real-world scenarios. Compared with existing multi-view reflection datasets Guo et al. (2022); Zhu et al. (2022); Xie et al. (2024); Liu et al. (2026), MVRefSep offers more comprehensive data support in terms of scene scale, environmental diversity, and layer-level ground truth.

Table 1: Quantitative comparison on MVRefSep-Test. Values ranked **1st**, 2nd, and 3rd are highlighted accordingly.

Method	Outdoor Day				Outdoor Night				Indoor Day				Indoor Night				Average			
	Transmission		Reflection		Transmission		Reflection		Transmission		Reflection		Transmission		Reflection		Transmission		Reflection	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MIRM (TIP'21)	12.53	0.570	18.60	0.649	15.56	0.604	18.27	0.577	11.07	0.585	15.67	0.590	13.36	0.634	17.68	0.580	13.13	0.598	17.56	0.599
NeRFReN (CVPR'22)	11.94	0.215	12.49	0.039	13.74	0.253	15.84	0.101	12.63	0.393	10.84	0.111	12.18	0.349	14.91	0.113	12.62	0.302	13.52	0.091
Ref-Unlock (arXiv'25)	19.08	0.752	14.81	0.388	19.72	0.703	18.95	0.571	18.80	0.803	13.53	0.478	19.17	0.775	17.29	0.545	19.19	0.758	16.14	0.495
DSRNet (ICCV'23)	19.33	0.823	17.63	0.619	20.22	0.781	20.55	0.704	18.63	0.844	17.12	0.655	19.57	0.847	18.12	0.651	19.44	0.824	18.35	0.657
DSIT (NIPS'24)	20.77	0.838	14.92	0.380	20.65	0.791	18.78	0.582	19.87	0.860	13.23	0.475	20.73	0.858	16.85	0.483	20.50	0.837	15.95	0.480
RDRNet (CVPR'24)	19.81	0.824	14.58	0.333	20.54	0.801	18.03	0.516	19.78	0.851	13.09	0.434	20.73	0.851	16.55	0.432	20.21	0.832	15.56	0.429
RDNet (CVPR'25)	20.72	0.840	14.33	0.267	21.08	0.810	18.20	0.474	20.20	0.866	12.89	0.388	21.12	0.858	16.16	0.360	20.78	0.843	15.40	0.372
IRLR (TIP'26)	19.36	0.817	18.72	0.705	20.01	0.776	20.60	0.679	18.83	0.843	17.30	0.667	19.35	0.838	18.41	0.646	19.39	0.818	18.76	0.674
DMDNet (AAAI'26)	20.99	0.844	15.26	0.371	21.96	0.815	19.09	0.583	20.80	0.877	13.93	0.488	21.69	0.862	16.61	0.425	21.36	0.850	16.22	0.467
ReflexSplit (CVPR'26)	21.25	0.845	14.93	0.346	20.93	0.790	18.74	0.540	19.57	0.856	13.05	0.424	21.12	0.859	16.73	0.445	20.72	0.838	15.86	0.439
GFRRN (CVPR'26)	21.02	0.848	14.63	0.336	21.84	0.814	19.01	0.601	20.58	0.878	13.14	0.461	21.50	**0.863**	16.66	0.461	21.24	0.851	15.86	0.465
JLD (CVPR'26)	22.09	0.795	19.86	0.656	21.52	0.727	21.37	0.715	20.79	0.834	18.60	0.721	20.94	0.790	19.84	0.690	21.34	0.787	19.92	0.696
ReflectWorld (Ours)	**22.99**	**0.856**	**23.25**	**0.790**	**23.61**	**0.833**	**23.30**	**0.734**	**22.21**	**0.882**	**21.93**	**0.797**	**23.19**	0.862	**21.78**	**0.763**	**23.00**	**0.858**	**22.56**	**0.771**

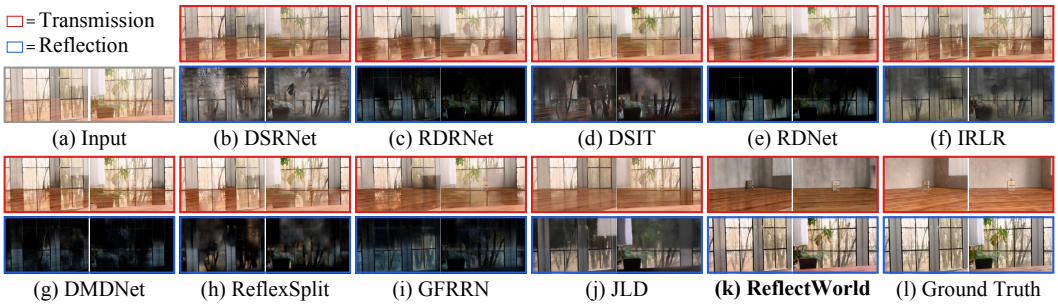


Figure 5: Qualitative comparison on MVRefSep-Test. ReflectWorld shows better recovery of both transmission and reflection layers.

5 Experiments

5.1 Experimental Setup

We adopt the one-step SD-Turbo Sauer et al. (2024) as the diffusion model and VGGT-Ω Wang et al. (2026a) as the feed-forward 3D model. For multi-view training, each sample consists of 1 target view and 4 reference views. During training, all input images are resized to 512×512, and the diffusion timestep is fixed at $t = 199$. The LoRA Hu et al. (2021) rank for the VAE decoder is set to 4. We train the model for 40K steps using the AdamW optimizer Loshchilov & Hutter (2017) with a batch size of 1 and a learning rate of 2×10^{-5}. All experiments are conducted on a single NVIDIA RTX 6000 Ada GPU. See the Appendix for more details.

5.2 Datasets

We use 8 scenes from MVRefSep-Sim as MVRefSep-Train and the remaining 32 scenes as MVRefSep-Test. Multi-view training data are constructed from DL3DV-1K Ling et al. (2024) and MVRefSep-Train, supplemented with single-image reflection datasets for basic layer decomposition supervision. DL3DV-1K, MVRefSep-Train, VOC2012 Everingham et al. (2010), Flickr30k Young

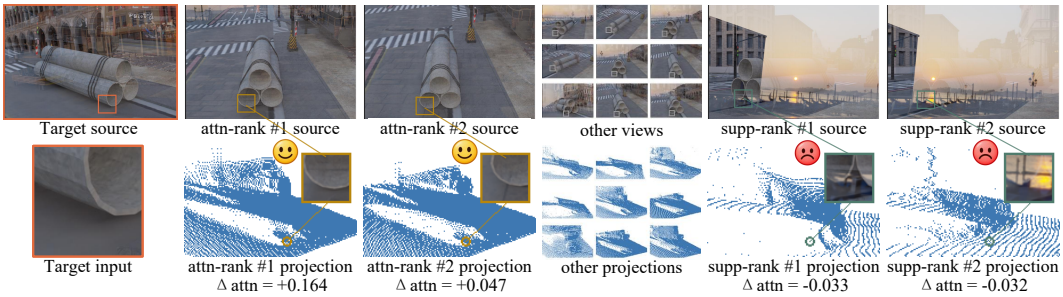


Figure 6: Geometry-driven routing and cross-view association in GeoRCA. Candidates more consistent with the target across views receive increased association weights, while inconsistent or reflection-contaminated candidates are suppressed. Δattn denotes the change in association weight.

et al. (2014), Real Zhang et al. (2018b), and Nature Li et al. (2020) are sampled with probabilities of $[0.5, 0.1, 0.1, 0.2, 0.05, 0.05]$, respectively.

5.3 Performance Evaluation

We compare ReflectWorld with 12 methods, including three multi-image methods, MIRM Li et al. (2021), NeRFReN Guo et al. (2022), and Ref-Unlock Song et al. (2025), and nine single-image methods, DSRNet Hu & Guo (2023), DSIT Hu et al. (2024), RDRNet Zhu et al. (2024), RDNet Zhao et al. (2025), IRLR Song et al. (2026), DMDNet Fang et al. (2026), ReflexSplit Lee et al. (2026), GFRRN Chen et al. (2026), and JLD Huang et al. (2026). As shown in Table 1, ReflectWorld achieves leading performance on MVRefSep-Test for both transmission and reflection recovery, with a particularly pronounced advantage in reflection recovery. The improvements across all four scene categories further demonstrate its robustness under diverse illumination and scene conditions.

Table 2: Component ablation.

Setting	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
w/o GeoRCA	20.73	0.844	21.64	0.759
w/o Geometric Encoding	22.28	0.852	22.39	0.765
w/o $\boldsymbol{T}/\boldsymbol{R}$ Interaction	22.98	0.853	22.18	0.768
Full	**23.00**	**0.858**	**22.56**	**0.771**

Figure 5 presents a qualitative comparison on MVRefSep-Test. ReflectWorld better recovers both $\boldsymbol{T}$ and $\boldsymbol{R}$ than the compared methods. Additional evaluations of cross-view consistency and real-world generalization are provided in the Appendix.

5.4 Ablation Studies

All ablation experiments are conducted on MVRefSep-Test.

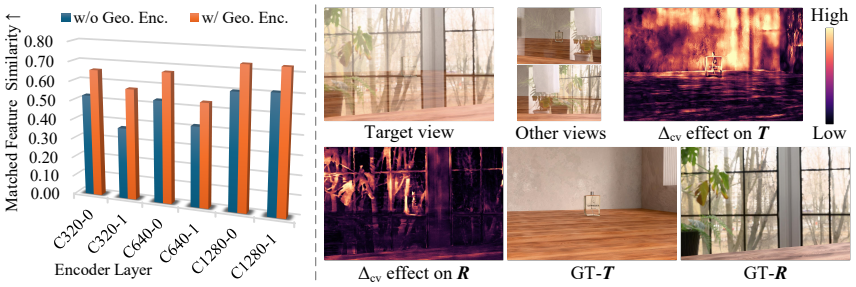
Table 3: Effect of reference-view count.

Refs.	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
0	20.56	0.835	20.38	0.662
4	22.61	0.859	22.48	0.763
8	22.70	**0.860**	22.51	0.765
12	22.73	**0.860**	22.51	0.765
16	**22.75**	**0.860**	**22.53**	**0.766**

Component ablation. We individually remove GeoRCA, Geometric Encoding, and $\boldsymbol{T}/\boldsymbol{R}$ Interaction. As shown in Table 2, removing any component degrades the performance of the full model.

Reference-view analysis. As shown in Table 3, adding reference views improves both transmission and reflection recovery, but the gains gradually saturate as more views are introduced. To investigate whether the additional reflection contamination introduced by more views limits their marginal benefit, we progressively replace mixed reference views with clean GT-$\boldsymbol{T}$. As shown in Table 4, both layers consistently improve as the GT-$\boldsymbol{T}$ ratio increases, suggesting that reference quality is a key factor in exploiting multi-view information.

Geometric projection. Figure 6 visualizes geometric projections from reference views to a target view. While the projections provide plausible cross-view candidates, projection errors and reflection contamination can still make them unreliable, rendering geometry a coarse search space rather than reliable evidence.



(a) Analysis of Geometric Encoding (b) Visualization of cross-view consistency correction Δ_{cv}

Figure 7: Analysis of Geometric Encoding and cross-view consistency correction Δ_{cv}. (a) Geometric Encoding consistently improves cross-view feature similarity across encoder layers. (b) Δ_{cv} exhibits distinct spatial responses on the transmission and reflection branches.

Table 4: Recovery with GT-T reference replacement.

GT-T ratio	4 References				8 References			
	T-PSNR	T-SSIM	R-PSNR	R-SSIM	T-PSNR	T-SSIM	R-PSNR	R-SSIM
0%	22.63	0.859	22.49	0.763	22.72	0.860	22.52	0.765
25%	22.76	0.862	22.58	0.765	22.83	0.862	22.60	0.766
50%	22.84	0.863	22.65	0.767	22.88	0.863	22.66	0.767
75%	22.91	0.864	22.73	0.769	22.92	0.864	22.74	0.769
100%	**23.02**	**0.866**	**22.93**	**0.775**	**23.00**	**0.866**	**22.95**	**0.775**

What GeoRCA selects. As shown in Fig. 6, GeoRCA tends to assign higher weights to reference views with smaller viewpoint changes and cleaner corresponding content, while suppressing views with stronger reflection contamination or larger viewpoint changes. This enables GeoRCA to identify reliable cross-view consistency evidence.

Table 5: Differential modulation directions.

Setting	T-PSNR	T-SSIM	R-PSNR	R-SSIM
Same direction $(+/+)$	21.60	0.846	22.12	0.766
Reversed $(-/+)$	21.60	0.847	21.45	0.750
T only $(+/0)$	22.32	0.850	21.73	0.757
R only $(0/-)$	22.40	0.850	22.04	0.768
Full $(+/-)$	**23.00**	**0.858**	**22.56**	**0.771**

Analysis of Geometric Encoding. As shown in Fig. 7(a), enabling Geometric Attention in MVGT consistently improves cross-view feature consistency across all encoder layers. This indicates that geometric priors are effectively injected into the diffusion features, promoting more consistent representations across views.

Effect of cross-view consistency correction Δ_{cv}. As shown in Fig. 7(b), Δ_{cv} exhibits distinct spatial responses on the T/R branches, indicating that it modulates different regions rather than acting as a uniform global bias. Stronger responses appear in cross-view-stable regions for T and view-dependent regions for R, supporting the differential decoding of the two layers.

Differential modulation directions. The results in Table 5 show that the full $(+/-)$ modulation achieves the best performance. Using the same or reversed directions leads to clear degradation, while modulating only one layer also yields inferior results. This demonstrates that applying the cross-view correction in opposite directions to T and R is important for differential layer generation.

6 Conclusion

In this work, we introduced ReflectWorld, a geometry-routed diffusion framework for multi-view reflection separation. ReflectWorld turns reflection-corrupted multi-view geometry into reliable separation evidence through GeoRCA, while MGR-Diff exploits the distinct cross-view behaviors of transmission and reflection to guide differential layer generation. We further introduced MVRef-Sep, a multi-view benchmark for evaluating both transmission and reflection recovery. Extensive experiments demonstrate state-of-the-art performance. Future work will explore tighter integration of geometric reasoning and layer generation to reduce reliance on external 3D models.

AI USE STATEMENT

Generative AI tools were used to assist with language translation and polishing, literature retrieval and discovery, and software coding. All AI-assisted content, code, and retrieved references were reviewed and verified by the authors. Generative AI tools were not used for research ideation. The authors take full responsibility for the final content of this work, including text, claims, and code produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, hyperparameters, dataset construction and splits, data preprocessing, and evaluation protocols in the main paper and appendix. Additional implementation details, training configurations, ablation settings, qualitative results, and detailed experimental analyses are also included to facilitate reproduction and verification of the reported findings. Anonymous source code and configuration files are provided as supplementary material.

REFERENCES

Jie Cai, Kangning Yang, Ling Ouyang, Lan Fu, Jiaming Ding, Jinglin Shen, and Zibo Meng. Openrr-5k: A large-scale benchmark for reflection removal in the wild. In *2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 14–19. IEEE, 2025.

Yu Chen, Zewei He, Xingyu Liu, Zixuan Chen, and Zhe-Ming Lu. Gfrrn: Explore the gaps in single image reflection removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5690–5699, 2026.

Mingxuan Cui, Yunrui Zhu, Wuqi Wang, Di Lin, Jianhua Zhang, Jie Zhang, Ming-Ming Cheng, Shengyong Chen, and Qing Guo. Glassgs: Geometry and concept-aware 3d gaussian splatting for reflective enclosures. In *European Conference on Computer Vision*, pp. 109–127. Springer, 2026.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Siyan Fang, Yuntao Wang, Jinpu Zhang, Ziwen Li, and Yuehuan Wang. Adaptive language-aware image reflection removal network. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 973–981, 8 2025.

Siyan Fang, Long Peng, Yuntao Wang, Ruonan Wei, and Yuehuan Wang. Depth-synergized mamba meets memory experts for all-day image reflection separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3849–3857, 2026.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9935–9946. IEEE, 2023.

Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18388–18397. IEEE, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jichen Hu, Chen Yang, Zanwei Zhou, Jiemin Fang, Qi Tian, and Wei Shen. Dereflection any image with diffusion priors and diversified data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4860–4868, 2026.

Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13092–13101. IEEE, 2023.

Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. *Advances in Neural Information Processing Systems*, 37:55228–55248, 2024.

Zheng-Hui Huang, Zhixiang Wang, Yu-Lun Liu, and Yung-Yu Chuang. Reflection separation from a single image via joint latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4569–4579, 2026.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21277–21286. IEEE, 2025.

Chia-Ming Lee, Yu-Fan Lin, Jin-Hui Jiang, Yu-Jou Hsiao, Chih-Chung Hsu, and Yu-Lun Liu. Reflexsplit: Single image reflection separation via layer fusion-separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1300–1309, June 2026.

Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E. Hopcroft. Single image reflection removal through cascaded refinement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3562–3571, 2020. doi: 10.1109/CVPR42600.2020.00362.

Mingjia Li, Jin Hu, Hainuo Wang, Qiming Hu, Jiarui Wang, and Xiaojie Guo. Rectifying latent space for generative single-image reflection removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8397–8407, 2026.

Tingtian Li, Yuk-Hee Chan, and Daniel P. K. Lun. Improved multiple-image-based reflection removal algorithm using deep neural networks. *IEEE Transactions on Image Processing*, 30:68–79, 2021. doi: 10.1109/TIP.2020.3031184.

Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2752–2759, 2014.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Chengxu Liu, Lu Qi, Jinshan Pan, Xueming Qian, and Ming-Hsuan Yang. Frequency domain-based diffusion model for unpaired image dehazing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7538–7547. IEEE, 2025.

Shuhong Liu, Chenyu Bao, Ziteng Cui, Yun Liu, Xuangeng Chu, Lin Gu, Marcos V Conde, Ryo Umagami, Tomohiro Hashimoto, Zijian Hu, et al. Realx3d: A physically-degraded 3d benchmark for multi-view visual restoration and reconstruction. *International Journal of Computer Vision*, 134(8):382, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yucheng Mao, Boyang Wang, Nilesh Kulkarni, and Jeong Joon Park. Sir-diff: Sparse image sets restoration with multi-view diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21620–21630, 2025. doi: 10.1109/CVPR52734.2025.02014.

BH Prasad, Kaushik Mitra, et al. Burst reflection removal using reflection motion aggregation cues. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 239–248, 2023.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3193–3201, 2015.

Junseong Shin and Tae Hyun Kim. Prism: Latent composition consistency for single-image reflection removal. In *European Conference on Computer Vision (ECCV)*, 2026.

Binbin Song, Jiantao Zhou, Shuning Xu, Xina Liu, Haiwei Wu, Xiaopeng Fan, and Bihan Wen. Single image reflection removal via iterative prompt learning of reflection level. *IEEE Transactions on Image Processing*, 2026.

Jiayi Song, Zihan Ye, Qingyuan Zhou, Weidong Yang, Ben Fei, Jingyi Xu, Ying He, and Wanli Ouyang. Reflections unlock: Geometry-aware reflection disentanglement in 3d gaussian splatting for photorealistic scenes rendering. *arXiv preprint arXiv:2507.06103*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Zepeng Wang, Jiagao Hu, Fuhao Li, Yuxuan Chen, Fei Wang, and Daiguo Zhou. From synthesis to removal: Physics-grounded reflection simulation and diffusion-based video dereflection. *arXiv preprint arXiv:2608.11562*, 2026b.

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.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13049–13059, 2023. doi: 10.1109/ICCV51070.2023.01204.

Mingyang Xie, Haoming Cai, Sachin Shah, Yiran Xu, Brandon Y Feng, Jia-Bin Huang, and Christopher A Metzler. Flash-splat: 3d reflection removal with flash cues and gaussian splats. In *European Conference on Computer Vision*, pp. 122–139. Springer, 2024.

Tianyi Xu, Zifeng Wang, Boyang Lv, Shuchen Weng, and Boxin Shi. Diffusion-based dual-view reflection removal. In *European Conference on Computer Vision*, pp. 326–346. Springer, 2026.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. Hdw-sr: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23462–23472, 2026.

Jiaolong Yang, Hongdong Li, Yuchao Dai, and Robby T Tan. Robust optical flow estimation of double-layer images under transparency or reflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1410–1419, 2016.

Wenbin Yin, Junkang Zhang, Sunzhe Yang, Faming Fang, and Guixu Zhang. Lightrr: A lightweight network for single image reflection removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19706–19715, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

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

Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4786–4794, 2018b. doi: 10.1109/CVPR.2018.00503.

Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems*, 36:49703–49714, 2023.

Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26430–26439. IEEE, 2025.

Chengxuan Zhu, Renjie Wan, and Boxin Shi. Neural transmitted radiance fields. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38994–39006. Curran Associates, Inc., 2022.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305. IEEE, 2025.

Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting single image reflection removal in the wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25468–25478. IEEE, 2024.

Appendix

A Additional Experimental Settings

A.1 Additional Implementation Details

For single-image samples, VGGT-Ω, MVGT, and GeoRCA, which rely on multi-view information, are disabled. For multi-view training samples constructed from DL3DV-1K Ling et al. (2024), we sample transmission and reflection sequences from two different scenes. To better approximate realistic cross-view variation of the two layers, for each transmission reference view, we select a reflection view with a similar camera-motion magnitude based on changes in camera rotation and camera-center displacement. The mixed observations are then synthesized as

$$T' = \alpha T, \qquad R' = \beta\, \mathcal{B}(R), \qquad I = T' + R' - T'R', \tag{10}$$

where $\mathcal{B}(\cdot)$ denotes Gaussian blur, and α and β control the intensities of the transmission and reflection layers, respectively.

During multi-view inference, the views of each scene are divided into two interleaved groups and processed separately to accommodate GPU memory constraints. For image-quality evaluation, all predicted and ground-truth images are first saved in PNG format and then reloaded for metric computation. PSNR and SSIM are computed in the RGB color space on the saved images.

Experiments are conducted on a Linux workstation using a single NVIDIA RTX 6000 Ada Generation GPU and an Intel Xeon w9-3575X CPU. The implementation uses Python 3.10 and PyTorch 2.10 with CUDA 12.8. Training and evaluation are performed using the same software environment.

A.2 Training Loss

We jointly optimize the transmission and reflection layers using pixel reconstruction, structural similarity, and perceptual losses. For each layer $X \in \{T, R\}$, the three losses are defined as

$$\mathcal{L}^X_{\text{pix}} = \|\widehat{X} - X\|_2^2, \qquad \mathcal{L}^X_{\text{ssim}} = 1 - \text{SSIM}(\widehat{X}, X), \qquad \mathcal{L}^X_{\text{lpips}} = \text{LPIPS}(\widehat{X}, X). \tag{11}$$

where SSIM Wang et al. (2004) and LPIPS Zhang et al. (2018a) measure structural and perceptual similarity, respectively. The layer-wise objective is

$$\mathcal{L}_X = \lambda_{\text{pix}}\mathcal{L}^X_{\text{pix}} + \lambda_{\text{ssim}}\mathcal{L}^X_{\text{ssim}} + \lambda_{\text{lpips}}\mathcal{L}^X_{\text{lpips}}, \qquad X \in \{T, R\}. \tag{12}$$

Finally, the two layers are jointly optimized as

$$\mathcal{L} = \mathcal{L}_T + \lambda_R \mathcal{L}_R, \tag{13}$$

where λ_R balances the contribution of the reflection layer. We set $\lambda_{\text{pix}} = 1$, $\lambda_{\text{ssim}} = 1$, $\lambda_{\text{lpips}} = 2$, and $\lambda_R = 0.25$ by default.

A.3 Computational Complexity

ReflectWorld contains 2.571B parameters in total, of which 1.003B are trainable. For 512×512 RGB inputs, the full framework requires 9.654 TFLOPs per view and takes an average of $0.392\,\text{s}$ to process each view on an NVIDIA RTX 6000 Ada Generation GPU.

B Additional Details of MVRefSep

B.1 MVRefSep-Sim

All images in MVRefSep-Sim are rendered at a resolution of 1200×800. Each scene contains 50 viewpoints distributed over five different height levels, with 10 cameras placed at each height. At the same height, cameras are evenly spaced along an arc-shaped trajectory facing the glass. Since through-glass imaging requires the camera to remain on the same side of the glass, camera placement

is terminated when the trajectory approaches the glass plane, resulting in a truncated orbital sampling scheme constrained by the glass plane.

For each viewpoint, we render the mixed observation $\boldsymbol{I}$, the Ground-Truth Transmission (GT-$\boldsymbol{T}$), and the Ground-Truth Reflection (GT-$\boldsymbol{R}$) under the same camera configuration, and record the corresponding camera poses. This multi-height and multi-view design introduces viewpoint variations in both horizontal and vertical directions, supporting multi-view reflection separation and cross-view consistency evaluation.

B.2 MVRefSep-Real

We adopt a viewpoint-wise paired acquisition protocol with removable glass. The camera is mounted on a tripod, while the glass panel is installed on a fixed holder. At each viewpoint, the tripod is first fixed at the desired camera position. Without moving the camera, we separately capture the with-glass mixed observation $\boldsymbol{I}$ and the corresponding without-glass GT-$\boldsymbol{T}$. The tripod is moved to the next camera position only after the paired acquisition at the current viewpoint is completed, and the same procedure is then repeated. The resulting MVRefSep-Real images are provided at a resolution of 1200×800.

During the acquisition of each scene, the camera focal length, shutter speed, aperture, ISO, white balance, and focus mode are kept fixed. Each image pair is captured consecutively within a short time interval to minimize additional differences caused by scene changes or illumination variations. This viewpoint-wise paired acquisition avoids pose discrepancies that may arise from separately repeating the same camera trajectory, thereby providing more reliable paired GT-$\boldsymbol{T}$.

C Extended Evaluation

We compare ReflectWorld with three multi-image methods, MIRM Li et al. (2021), NeRFReN Guo et al. (2022), and Ref-Unlock Song et al. (2025), as well as nine single-image methods, DSRNet Hu & Guo (2023), DSIT Hu et al. (2024), RDRNet Zhu et al. (2024), RDNet Zhao et al. (2025), IRLR Song et al. (2026), DMDNet Fang et al. (2026), ReflexSplit Lee et al. (2026), GFRRN Chen et al. (2026), and JLD Huang et al. (2026).

C.1 Cross-View Consistency Evaluation

We evaluate the cross-view consistency of the recovered transmission and reflection layers through novel-view synthesis. We use Depth Anything 3 (DA3-GIANT-1.1) Lin et al. (2025) for this evaluation. For each 50-view scene in MVRefSep-Test, we sparsely sample 7 views at a stride of 8 as context inputs, while the remaining 43 views are used as held-out targets for evaluating synthesis quality. As shown in Table 6, ReflectWorld achieves the best average performance for both transmission and reflection layers, indicating that the recovered $\boldsymbol{T}$ and $\boldsymbol{R}$ exhibit stronger cross-view consistency and better support novel-view synthesis.

Figures 8 and 9 provide qualitative novel-view synthesis comparisons. In both scenes, ReflectWorld produces cleaner transmission synthesis and more coherent reflection synthesis, indicating stronger cross-view consistency of the recovered layers.

C.2 Real-World Evaluation

We evaluate ReflectWorld on MVRefSep-Real to assess its generalization to real-world reflection-corrupted scenes. For each scene, we sparsely sample context views at a stride of 8 and use the remaining views as held-out targets for novel-view synthesis. For depth consistency, we use Depth Anything 3 (DA3-GIANT-1.1) and report depth Absolute Relative Error (AbsRel, lower is better) between the depth maps estimated from the recovered transmission and GT. As shown in Table 7, ReflectWorld achieves the best performance across transmission recovery, novel-view synthesis, and depth consistency, demonstrating real-world generalization in both image quality and cross-view consistency.

Figure 10 presents a real-world example. Several competing methods exhibit cross-layer leakage. In particular, JLD incorrectly places the white label on the document envelope in the reflection layer

Table 6: Quantitative comparison of **novel-view synthesis** using recovered transmission and reflection layers on MVRefSep-Test. Values ranked **1st**, 2nd, and 3rd are highlighted accordingly.

Method	Outdoor Day				Outdoor Night				Indoor Day				Indoor Night				Average			
	Transmission		Reflection		Transmission		Reflection		Transmission		Reflection		Transmission		Reflection		Transmission		Reflection	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MIRM (TIP'21)	12.04	0.440	16.75	0.591	14.04	0.453	18.25	0.595	10.42	0.468	14.10	0.530	12.64	0.514	16.72	0.571	12.28	0.469	16.45	0.572
NeRFReN (CVPR'22)	10.91	0.259	13.32	0.236	12.53	0.302	16.84	0.405	11.12	0.368	11.27	0.244	11.87	0.376	15.62	0.416	11.61	0.326	14.26	0.325
Ref-Unlock (arXiv'25)	16.82	0.520	14.37	0.387	15.59	0.468	18.56	0.603	15.30	0.545	12.94	0.443	16.14	0.576	16.76	0.560	15.96	0.527	15.66	0.498
DSRNet (ICCV'23)	16.67	0.507	16.51	0.567	15.73	0.465	19.27	0.634	15.43	0.540	14.90	0.566	16.26	0.571	17.41	0.608	16.02	0.521	17.02	0.594
DSIT (NIPS'24)	17.49	0.517	14.48	0.406	16.01	0.467	18.24	0.579	15.79	0.543	12.66	0.454	16.91	0.580	16.11	0.472	16.55	0.527	15.37	0.478
RDRNet (CVPR'24)	16.89	0.509	14.01	0.349	16.21	0.475	17.84	0.546	15.53	0.542	12.44	0.403	16.73	0.579	15.99	0.470	16.34	0.526	15.07	0.442
RDNet (CVPR'25)	17.29	0.512	13.82	0.310	16.26	0.477	17.77	0.527	**16.22**	0.548	12.35	0.389	16.97	0.577	15.82	0.443	16.68	0.529	14.94	0.417
IRLR (TIP'26)	16.72	0.506	16.91	0.623	16.07	0.478	18.88	0.606	15.53	0.538	15.04	0.581	16.31	0.576	17.05	0.598	16.16	0.525	16.97	0.602
DMDNet (AAAI'26)	17.37	0.517	14.10	0.369	16.37	0.482	18.11	0.571	16.21	0.549	12.62	0.423	17.05	0.576	15.82	0.444	16.75	0.531	15.16	0.452
ReflexSplit (CVPR'26)	17.70	0.521	14.24	0.357	15.92	0.465	17.89	0.545	15.78	0.541	12.52	0.426	16.84	0.576	15.86	0.438	16.56	0.526	15.13	0.441
GFRRN (CVPR'26)	17.50	0.517	13.96	0.340	16.08	0.474	17.77	0.532	16.21	0.551	12.50	0.421	16.93	0.579	15.91	0.446	16.68	0.530	15.03	0.435
JLD (CVPR'26)	17.60	0.504	18.61	0.622	15.88	0.453	19.69	0.640	16.00	0.543	15.77	0.598	16.69	0.564	18.28	0.642	16.54	0.516	18.09	0.625
ReflectWorld (Ours)	**17.82**	**0.524**	**19.68**	**0.666**	**16.71**	**0.491**	**19.92**	**0.648**	16.09	**0.556**	**16.31**	**0.603**	**17.51**	**0.589**	**19.18**	**0.655**	**17.03**	**0.540**	**18.77**	**0.643**

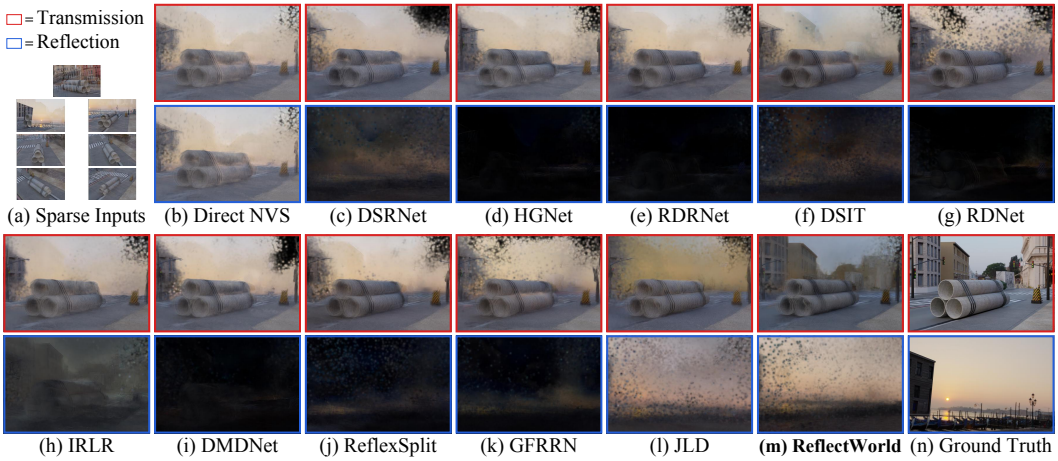


Figure 8: Qualitative novel-view synthesis comparison on an urban street scene. ReflectWorld produces less reflection contamination in the transmission layer and more complete reflection content in the reflection layer, yielding novel views that more closely match the ground truth.

instead of preserving it in the transmission layer. In contrast, ReflectWorld preserves this detail in the transmission layer while avoiding its leakage into the reflection estimate, demonstrating more accurate layer assignment.

D Additional Ablations

D.1 GeoRCA Routing and Candidate Selectivity

Routing displacement. To quantify how much geometry-driven routing changes the spatial locations used for cross-view association, we measure the displacement between each geometry-routed reference location and its same-coordinate counterpart on MVRefSep-Test. A geometry-routed reference location is displaced by 218.18 pixels on average across all query locations. Figure 11(a) further

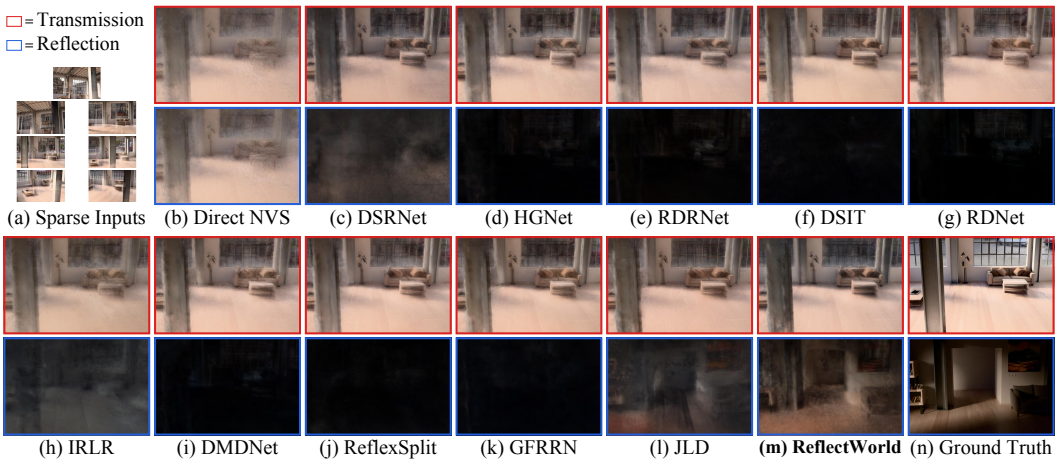


Figure 9: Qualitative novel-view synthesis comparison on an industrial-style loft scene. ReflectWorld more faithfully recovers the dim indoor scene in the reflection layer while maintaining cleaner transmission synthesis, yielding novel views that better match the ground truth.

Table 7: Quantitative comparison of transmission recovery, Novel-View Synthesis (NVS), and depth consistency on MVRefSep-Real. Higher values are better for all metrics except AbsRel, where ↓ indicates lower is better.

Method	PSNR	SSIM	NVS-PSNR	NVS-SSIM	AbsRel↓
DSRNet (ICCV'23)	19.79	0.893	16.07	0.690	0.0067
DSIT (NIPS'24)	23.13	0.915	16.78	0.694	0.0069
RDRNet (CVPR'24)	22.40	0.899	16.54	0.691	0.0069
RDNet (CVPR'25)	22.84	0.907	16.61	0.690	0.0068
IRLR (TIP'26)	21.35	0.894	16.53	0.691	0.0065
DMDNet (AAAI'26)	22.65	0.909	16.61	0.689	0.0075
ReflexSplit (CVPR'26)	23.66	0.914	16.76	0.694	0.0058
GFRRN (CVPR'26)	24.31	0.922	16.86	0.697	0.0061
JLD (CVPR'26)	25.08	0.898	16.99	0.686	0.0056
ReflectWorld (Ours)	**25.57**	**0.924**	**17.17**	**0.702**	**0.0054**

shows the distribution of scene-level mean displacements, indicating substantial spatial relocation across scenes. These substantial displacements show that geometry-driven routing substantially changes the spatial locations from which cross-view evidence is retrieved, rather than relying on same-coordinate interaction across views.

Association selectivity. To assess whether GeoRCA selectively emphasizes reliable routed candidates, we analyze the resulting association weights on MVRefSep-Test. When all other 49 views are potential references, validity filtering retains 22.64 candidates per query on average. If the valid candidates were weighted uniformly, each would receive an average weight of 0.044. As shown in Fig. 11(b), GeoRCA instead assigns an average top-1 association weight of 0.166, which is 3.75 times larger. Figure 11(c) further shows that the top-ranked candidate exhibits a transmission-consistency win rate above 50% across all analyzed encoder layers, with the preference being more pronounced at higher spatial resolutions.

D.2 Additional Analysis of Geometric Encoding

Geometry incorporation strategy. To determine how geometric information should be incorporated into the diffusion network, we compare different geometric fusion strategies in Table 8. Concatenation performs better than elementwise addition, as it preserves the appearance and geometric features separately rather than directly mixing them. Moreover, extending the concatenated

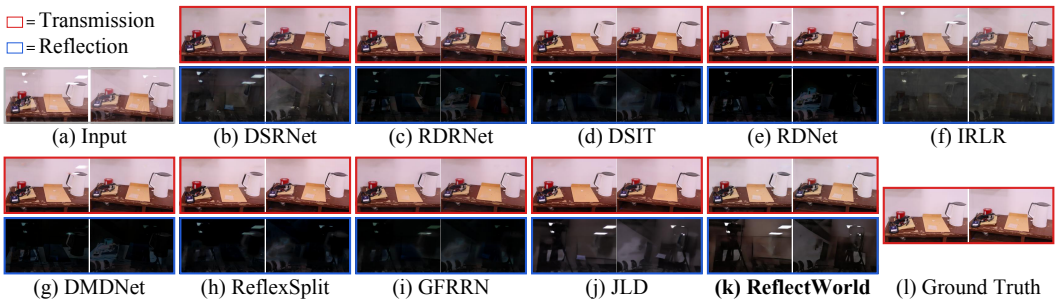


Figure 10: Qualitative comparison on MVRefSep-Real for a desk scene with document envelopes and a kettle. JLD incorrectly transfers the white label on the document envelope from the transmission layer to the reflection estimate, whereas ReflectWorld preserves it in the transmission layer and achieves cleaner layer separation.

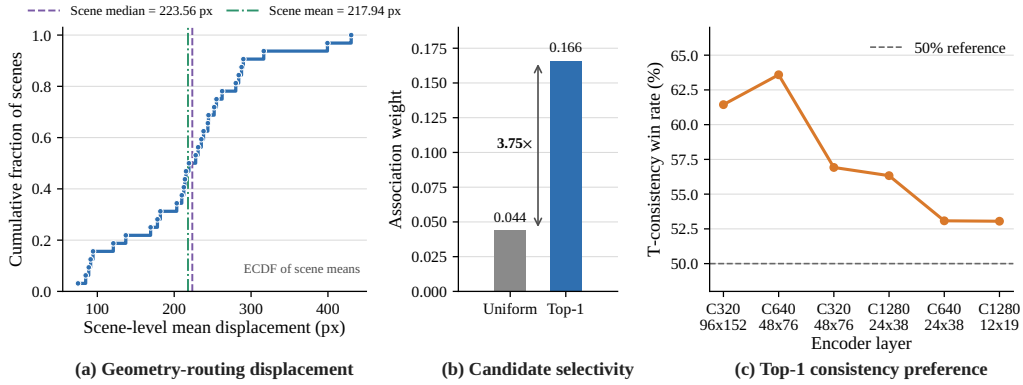


Figure 11: GeoRCA routing and candidate selectivity on MVRefSep-Test. (a) ECDF of scene-level mean geometry-routing displacement. The ECDF reports the fraction of scenes whose mean displacement does not exceed a given value. The substantial displacements show that geometry-driven routing markedly relocates the reference locations used for cross-view association rather than relying on same-coordinate interaction across views. (b) GeoRCA assigns substantially higher weights to selected candidates than uniform weighting, demonstrating selective cross-view association. (c) Top-ranked candidates consistently achieve transmission-consistency win rates above 50%, with a stronger preference at higher spatial resolutions. These results demonstrate that GeoRCA performs both nontrivial geometric routing and selective association of reliable cross-view candidates.

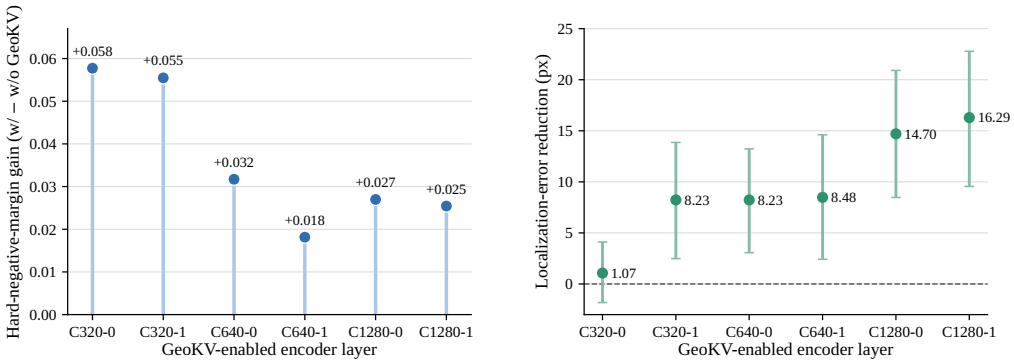
geometric encoding from the encoder to the decoder brings no additional benefit. We therefore adopt encoder-only concatenation in the final model.

Correspondence discrimination. To assess whether geometric conditioning improves the ability to distinguish true correspondences from nearby confusing locations, we analyze the feature margin between true correspondences and hard negatives. As shown in Fig. 12(a), geometric conditioning consistently enlarges the margin between true correspondences and nearby hard negatives (20–80 pixels from the true correspondence), with all scene-level bootstrap 95% confidence intervals above zero. This indicates that geometric conditioning improves correspondence discrimination.

Correspondence localization. To evaluate whether geometric conditioning improves the spatial localization of cross-view correspondences, we measure correspondence localization error across encoder layers. As shown in Fig. 12(b), geometric conditioning reduces the localization error at nearly all encoder layers, with larger improvements at deeper layers. The only exception is C320-0, whose confidence interval includes zero. This is because C320-0 lies at an early encoder stage, where the injected geometric information has undergone limited propagation. As features progress

Table 8: Ablation of geometric incorporation strategies on MVRefSep-Test.

Setting	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
Add	22.36	0.853	22.18	0.770
Concat (Encoder + Decoder)	22.75	0.856	22.22	0.769
Concat (Encoder)	**23.00**	**0.858**	**22.56**	**0.771**



(a) Hard-negative-margin gain **(b) Localization-error reduction**

Figure 12: Geometric-conditioning correspondence diagnostics across encoder layers. (a) Hard-negative-margin gains are positive at all layers. (b) Geometric conditioning reduces correspondence-localization error at nearly all layers, with larger reductions at deeper layers. These results show that geometric conditioning improves correspondence discrimination and localization.

through deeper encoder layers, geometric information is more fully integrated with appearance representations, leading to clearer improvements in correspondence localization.

Effect of reflection interference. To examine whether the benefit of geometric conditioning changes with the strength of reflection interference, we group correspondence queries according to their local reflection levels. As shown in Fig. 13(a), the gain in matched-feature similarity consistently increases from low- to high-reflection regions across all encoder layers. Figure 13(b) shows a similar overall tendency for the hard-negative-margin gain. These results suggest that geometric conditioning becomes particularly beneficial under stronger reflection interference, as appearance-based correspondence becomes more ambiguous, making geometric cross-view cues increasingly useful as a complement.

D.3 Additional Analysis of Differential Decoupling

Controlled reflection-motion stress test. To examine whether the benefit of differential decoupling increases with stronger cross-view reflection variation, we vary the view-dependent component of the reflection layer. Specifically, for each scene, we compute the scene-wise mean reflection $\bar{R}$ across all views and construct

$$R_\alpha = \mathrm{clip}\big(\bar{R} + \alpha(R - \bar{R})\big), \tag{14}$$

where $R - \bar{R}$ represents the view-dependent deviation from the mean reflection, and α controls the strength of this variation. The corresponding mixed observation is adjusted as

$$I_\alpha = \mathrm{clip}\big(I + R_\alpha - R\big), \tag{15}$$

while the transmission layer remains unchanged. Thus, $\alpha = 1$ recovers the original input, whereas $\alpha < 1$ suppresses and $\alpha > 1$ amplifies cross-view reflection variation.

As shown in Fig. 14(a), when reflection variation is weak, differential decoupling provides little benefit in $\boldsymbol{R}$-PSNR and can even cause degradation. As α increases, the $\boldsymbol{R}$-PSNR gain steadily increases and becomes positive, while the $\boldsymbol{T}$-PSNR gain remains positive. Consequently, the mean $\boldsymbol{T}/\boldsymbol{R}$ PSNR gain increases with the strength of cross-view reflection variation. A similar trend is

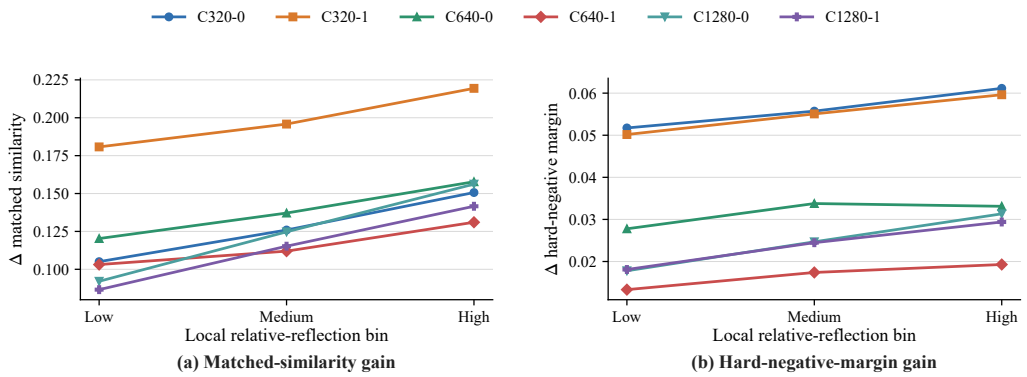


Figure 13: Gains from geometric conditioning under different levels of local reflection interference. (a) Matched-similarity gains consistently increase from low- to high-reflection regions. (b) Hard-negative-margin gains show the same overall tendency. This indicates that geometric conditioning becomes more beneficial as reflection interference increases.

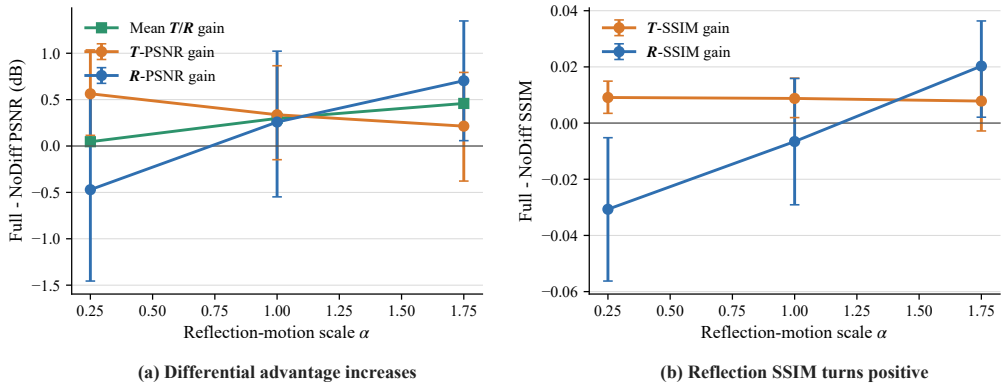


Figure 14: Controlled cross-view reflection-variation test. (a) The $\boldsymbol{R}$-PSNR gain increases with reflection variation, while the $\boldsymbol{T}$-PSNR gain remains positive, increasing the mean $\boldsymbol{T}/\boldsymbol{R}$ PSNR gain. (b) The $\boldsymbol{R}$-SSIM gain changes from negative to positive, whereas the $\boldsymbol{T}$-SSIM gain remains relatively stable. These trends show that differential decoupling becomes more effective as reflection exhibits stronger cross-view variation. Error bars denote scene-clustered 95% confidence intervals.

observed in Fig. 14(b): the $\boldsymbol{R}$-SSIM gain changes from negative to positive as α increases, whereas the $\boldsymbol{T}$-SSIM gain remains relatively stable and positive.

D.4 Training and Design Choices

We analyze the training configurations and design choices of ReflectWorld to determine the default settings used in our experiments.

Training reference count. As shown in Table 9, increasing the number of training reference views from 0 to 4 consistently improves transmission recovery. Further increasing the number of references to 5 or 6 no longer benefits transmission, although some reflection metrics continue to improve. This indicates that using more views does not uniformly benefit the two layers. We therefore use 4 reference views by default, which provides a favorable balance between transmission and reflection recovery while avoiding additional reference-view overhead.

Feed-forward 3D backbone. Table 10 compares different feed-forward 3D backbones for extracting multi-view geometric cues. All backbones can be integrated into ReflectWorld and achieve reasonable separation performance, indicating that the framework is not restricted to a specific 3D foundation model. Within the VGGT family, VGGT-Ω Wang et al. (2026a) overall outperforms

Table 9: Ablation of training reference-view count.

Refs.	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
0	21.38	0.847	22.25	0.748
1	21.81	0.848	21.51	0.752
2	21.82	0.851	22.26	0.774
3	21.95	0.852	22.57	0.766
4	**23.00**	**0.858**	22.56	0.771
5	22.48	0.856	22.55	0.788
6	22.09	0.855	**22.80**	**0.793**

Table 10: Ablation of feed-forward 3D backbone choice.

Backbone	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
VGGT	22.24	0.853	22.23	0.773
VGGT-Ω	**23.00**	**0.858**	**22.56**	0.771
DA3-Small	22.22	0.847	22.14	**0.775**
DA3-Base	21.57	0.848	22.06	0.751
DA3-Large	22.18	0.852	22.51	0.773
DA3-Giant	21.91	0.851	22.20	0.757

Table 11: Ablation of the one-step diffusion time step.

t	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
99	22.98	0.850	21.68	0.762
149	22.83	0.850	21.53	0.759
199	**23.00**	**0.858**	**22.56**	**0.771**
249	22.86	0.853	21.64	0.758
299	22.68	0.850	21.90	0.764

Table 12: Ablation of VAE LoRA placement and rank.

Setting	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
Decoder, $r = 1$	22.20	0.852	22.09	0.757
Decoder, $r = 2$	22.39	0.854	**22.57**	**0.787**
Decoder, $r = 4$	**23.00**	**0.858**	22.56	0.771
Decoder, $r = 6$	22.67	0.855	22.36	0.769
Decoder, $r = 8$	22.23	0.853	22.31	0.767
Encoder+decoder $r = 4$	22.07	0.851	21.77	0.723

VGGT Wang et al. (2025), suggesting that stronger multi-view geometric representations can provide more effective guidance for our geometry-routed separation. In contrast, the Depth Anything 3 (DA3) Lin et al. (2025) variants do not exhibit a clear scaling trend with model size. This is because increasing DA3 capacity mainly strengthens its unified depth-ray representation, whereas Reflect-World benefits more from geometric cues that can be converted into cross-view associations. VGGT, by modeling richer multi-view geometric relations, provides more diverse geometric cues that are better aligned with this association-and-routing mechanism. Therefore, its stronger geometric representation can be more effectively translated into gains in the downstream separation task. Among the evaluated backbones, VGGT-Ω achieves the best $\boldsymbol{T}$-PSNR, $\boldsymbol{T}$-SSIM, and $\boldsymbol{R}$-PSNR, and is therefore adopted as our default backbone.

Diffusion time step. As reported in Table 11, the effect of the one-step diffusion time step is non-monotonic. Although $t = 99$ already provides competitive transmission recovery, its reflection performance remains lower, whereas larger time steps such as $t = 249$ and 299 also reduce the overall recovery quality. In contrast, $t = 199$ achieves the best result on all four metrics, indicating a favorable balance between transmission and reflection recovery. We therefore set the default diffusion time step to $t = 199$.

VAE adaptation. Table 12 studies both the placement and rank of VAE LoRA adaptation. Decoder-only adaptation with rank $r = 2$ slightly favors reflection recovery, whereas $r = 4$ achieves the strongest transmission performance while retaining nearly the best $\boldsymbol{R}$-PSNR. Increasing the rank further to $r = 6$ or 8 brings no additional benefit. Moreover, jointly adapting the encoder and decoder degrades performance, particularly $\boldsymbol{R}$-SSIM, suggesting that modifying the pretrained latent encoding is unfavorable for the subsequent diffusion process. We therefore freeze the encoder and apply rank-4 LoRA only to the VAE decoder.

Training data composition. Our default training mixture consists of DL3DV-1K Ling et al. (2024), MVRefSep-Train, VOC2012 Everingham et al. (2010), Flickr30k Young et al. (2014), Real Zhang et al. (2018b), and Nature Li et al. (2020), sampled with probabilities of 0.50, 0.10, 0.10, 0.20, 0.05, and 0.05, respectively. Table 13 shows that these training sources provide complementary supervision. Removing Flickr30k mainly degrades transmission recovery, whereas removing Real+Nature has a larger impact on reflection recovery. We further vary the proportion of MVRefSep-Train from the default 0.10 to 0.20 and 0.05, while adjusting DL3DV-1K from 0.50 to 0.40 and 0.55, respectively.

Table 13: Ablation of training-data composition.

Setting	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
w/o VOC2012	22.81	0.856	22.33	**0.784**
w/o Flickr30k	21.30	0.849	21.99	0.778
w/o Real+Nature	22.32	0.850	20.99	0.703
More MVRefSep-Train	22.82	0.853	22.03	0.760
Less MVRefSep-Train	22.76	0.857	22.46	0.781
Default	**23.00**	**0.858**	**22.56**	0.771

Table 14: Ablation of reflection-layer loss weight.

λ_R	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
0.10	22.89	0.854	21.77	0.760
0.125	22.83	0.851	21.30	0.759
0.25	**23.00**	**0.858**	**22.56**	0.771
0.50	22.13	0.853	22.35	**0.778**

Table 15: Ablation of individual loss terms.

Setting	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
w/o $\mathcal{L}_{\text{pix}}$	22.60	**0.858**	22.30	0.769
w/o $\mathcal{L}_{\text{ssim}}$	21.74	0.848	21.70	0.760
w/o $\mathcal{L}_{\text{lpips}}$	22.46	0.846	21.42	0.759
$\mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{lpips}}$	**23.00**	**0.858**	**22.56**	**0.771**

Table 16: Ablation of the LPIPS loss weight.

λ_{lpips}	$\boldsymbol{T}$-PSNR	$\boldsymbol{T}$-SSIM	$\boldsymbol{R}$-PSNR	$\boldsymbol{R}$-SSIM
1	22.41	0.856	22.45	**0.776**
2	**23.00**	**0.858**	**22.56**	0.771
3	22.33	0.852	21.95	0.759

Neither direction yields a consistent improvement across the two layers. The default data mixture therefore provides the most balanced overall performance.

Loss design. Tables 14, 15, and 16 analyze the loss formulation and weighting strategy. The reflection-layer weight λ_R controls the balance between transmission and reflection supervision, with the default setting providing the most balanced two-layer recovery. The loss-term ablation further demonstrates the complementary roles of pixel, perceptual, and structural supervision: removing $\mathcal{L}_{\text{ssim}}$ mainly degrades transmission recovery, whereas removing $\mathcal{L}_{\text{lpips}}$ has a larger impact on reflection recovery. We further observe that an intermediate λ_{lpips} provides a better overall balance than either a smaller or larger weight. Accordingly, we use $\lambda_R = 0.25$, the full $\mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{lpips}}$ objective, and $\lambda_{\text{lpips}} = 2$ as the default loss configuration.

E Additional Qualitative Comparisons

We provide additional qualitative comparisons on both MVRefSep-Test and MVRefSep-Real, covering diverse scenes and reflection conditions.

Additional results on MVRefSep-Test. Figures 15–22 present additional qualitative comparisons on MVRefSep-Test. ReflectWorld achieves cleaner transmission recovery and more complete reflection recovery than competing methods across diverse scenes, demonstrating superior two-layer separation quality.

Additional results on MVRefSep-Real. Figures 23–25 present additional qualitative comparisons on MVRefSep-Real, further illustrating the real-world generalization of ReflectWorld.

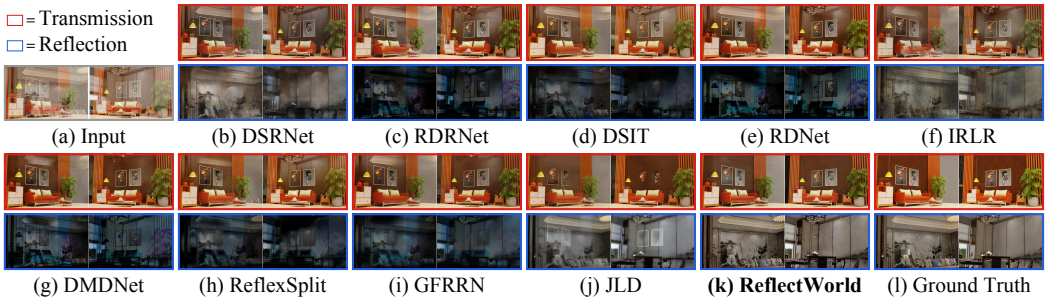


Figure 15: Qualitative comparison on MVRefSep-Test for a living-room scene. ReflectWorld produces cleaner transmission recovery and more complete reflection recovery, with less cross-layer leakage than competing methods.

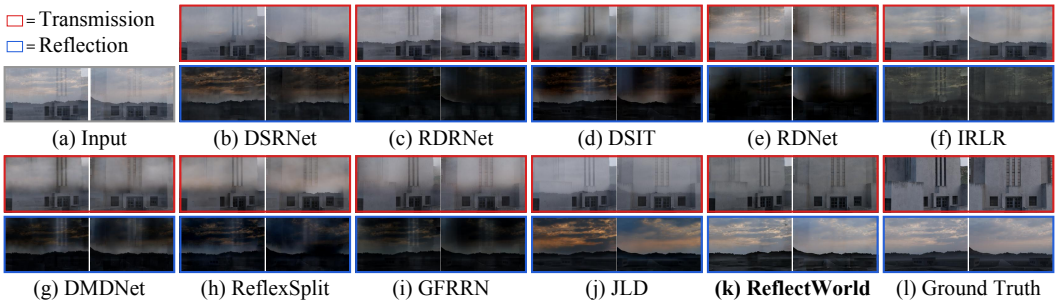


Figure 16: Qualitative comparison on MVRefSep-Test for an outdoor industrial-building scene under a cloudy sky. ReflectWorld better preserves the transmitted scene while recovering more complete reflected content, yielding a clearer separation of the two layers.

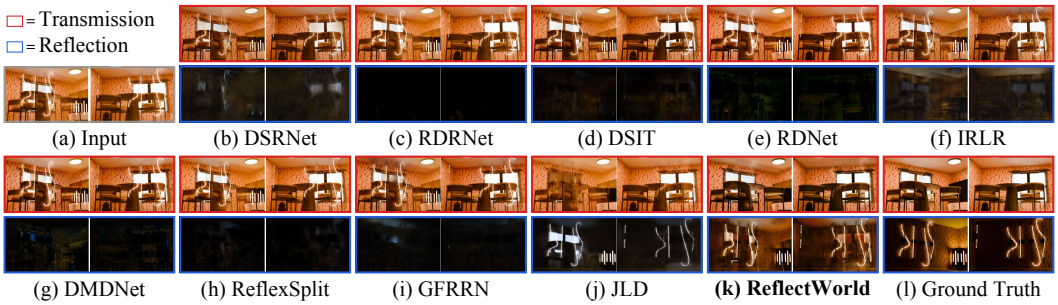


Figure 17: Qualitative comparison on MVRefSep-Test for a wood-paneled dining-room scene. ReflectWorld better restores the furniture structures in the transmission layer while recovering the distinctive bright-light pattern in the reflection layer, yielding a more faithful separation of the two layers.

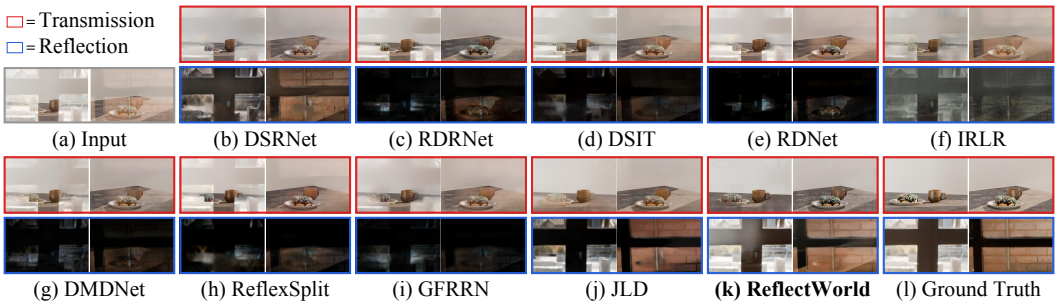


Figure 18: Qualitative comparison on MVRefSep-Test for a breakfast scene. ReflectWorld better preserves the donut and surrounding tabletop details in the transmission layer while recovering more complete structures in the reflection layer, yielding a cleaner separation of the two layers.

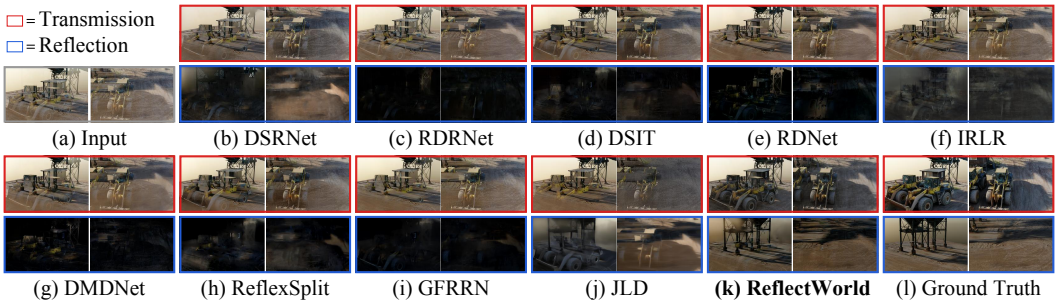


Figure 19: Qualitative comparison on MVRefSep-Test for an outdoor construction-site scene. ReflectWorld suppresses reflection contamination more effectively in the transmission layer while recovering clearer and more complete reflected content.

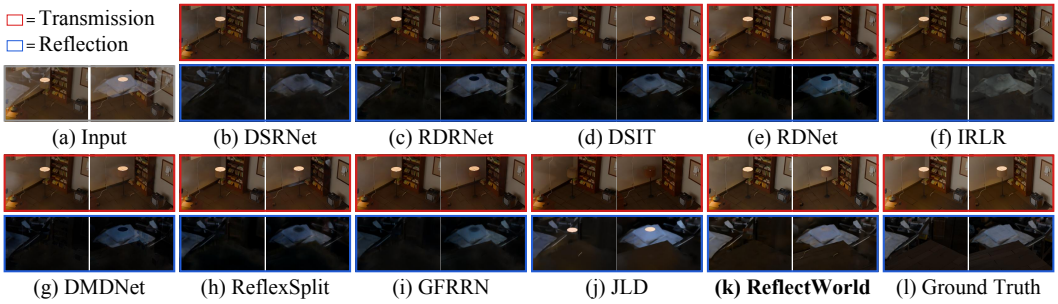


Figure 20: Qualitative comparison on MVRefSep-Test for a dimly lit office scene. Under challenging low-light conditions, ReflectWorld preserves the light in the transmission layer without erroneous leakage into the reflection layer, while recovering clear book structures in the reflection layer.

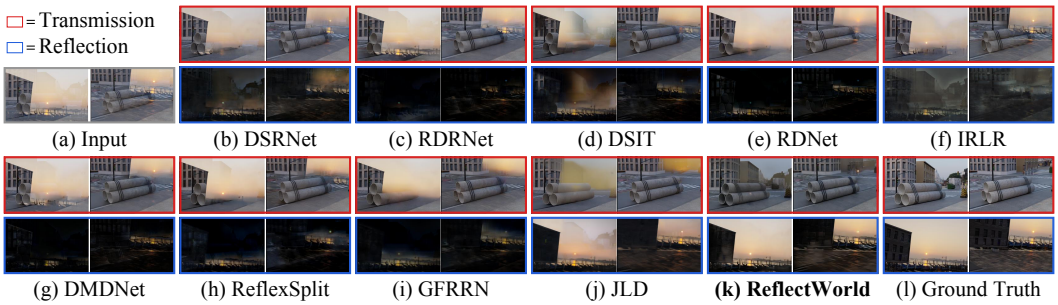


Figure 21: Qualitative comparison on MVRefSep-Test for an urban street scene. ReflectWorld produces a cleaner transmission layer with reduced reflection residuals while preserving more coherent content in the reflection layer.

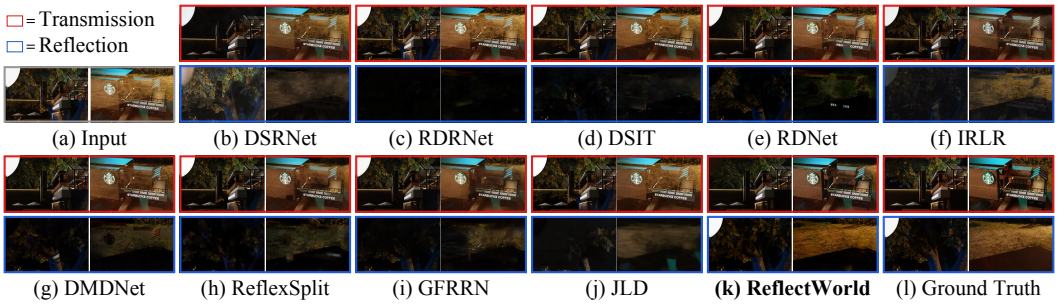


Figure 22: Qualitative comparison on MVRefSep-Test for an outdoor cafe storefront scene. ReflectWorld better suppresses cross-layer interference and recovers both transmission and reflection with more complete scene structures.

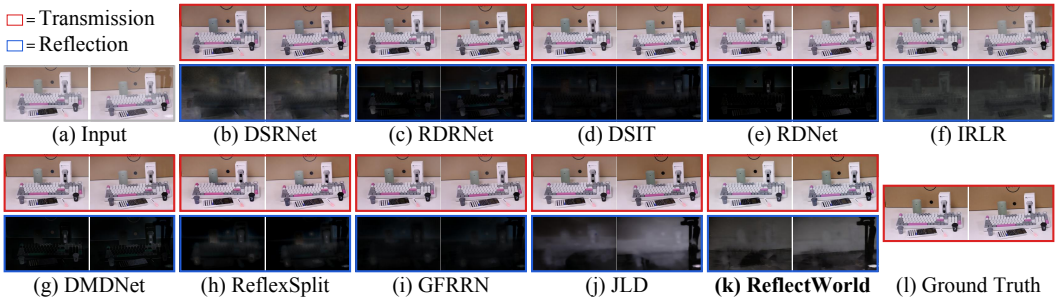


Figure 23: Qualitative comparison on MVRefSep-Real for a desktop scene with a pink keyboard and accessories. ReflectWorld produces a cleaner and more coherent reflection layer with fewer artifacts than competing methods.

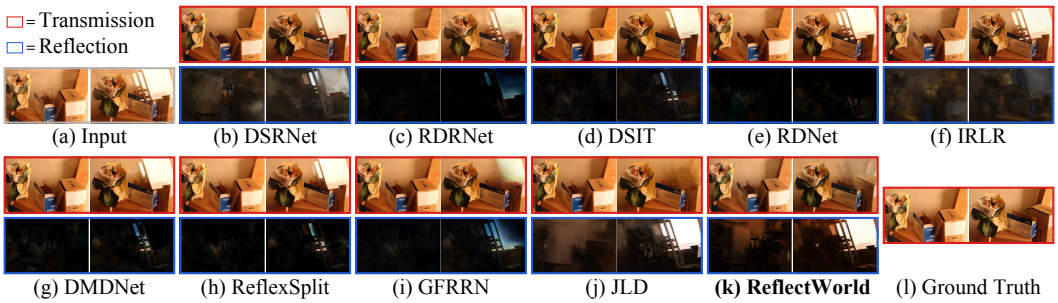


Figure 24: Qualitative comparison on MVRefSep-Real for an indoor scene with flowers and cardboard boxes. ReflectWorld better preserves the transmitted scene while recovering more complete reflection content.

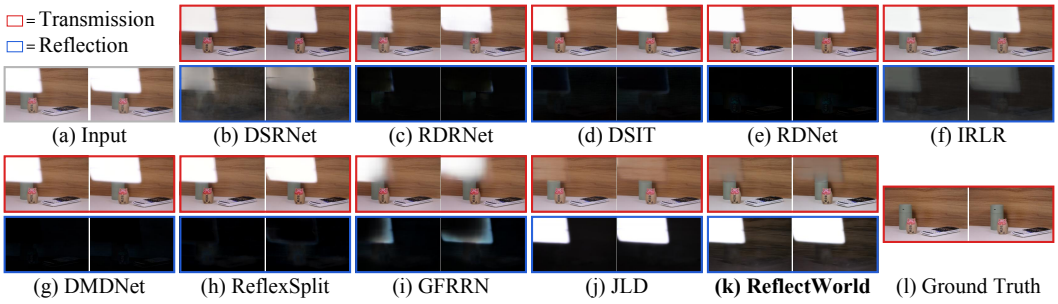


Figure 25: Qualitative comparison on MVRefSep-Real for a desk scene with a green can and stone figurines. ReflectWorld recovers a cleaner transmission layer with improved structural preservation while maintaining a clearer separation from reflected content.