Diffusion-based Dual-view Reflection Removal

Tianyi Xu[1], Zifeng Wang[1], Boyang Lv[1], Shuchen Weng[2,1], and Boxin Shi[1†]

[1]Peking University [2]Beijing Academy of Artificial Intelligence
photon@stu.pku.edu.cn shiboxin@pku.edu.cn

Abstract. Reflection removal aims to recover clean images from photographs degraded by glass, a crucial problem in mobile photography. Single-image methods suffer from the fundamental ill-posedness of decomposition and often rely on specialized hardware to enhance robustness. We propose a user-friendly dual-view diffusion framework that effectively leverages complementary viewpoint information for ubiquitous dual-camera devices. Our approach introduces frame-aware spatial attention to capture implicit cross-view correlations and integrates camera pose as conditional embeddings for geometric guidance. We contribute DualRef, a dataset for reflection removal comprising semi-synthetic pairs and real-world dual-camera sequences with ground truth. Experiments demonstrate over 10% PSNR improvement on real-world data, significantly outperforming state-of-the-art methods.

Keywords: Reflection removal · Diffusion model · Dual-view

1 Introduction

Reflection removal aims to recover clean transmission layers from images degraded by unwanted reflections through glass surfaces [64, 65]. This problem is ubiquitous in everyday photography scenarios [5, 52, 63]—from museum exhibits and vehicle windows to architectural facades—significantly impacting image quality in mobile photography applications. Among various capture setups, single-image reflection removal (SIRR) is the most user-friendly, requiring minimal operational constraints. However, SIRR is inherently ill-posed: decomposing a single mixture image $\mathbf{M}$ into transmission $\mathbf{T}$ and reflection $\mathbf{R}$ [64, 65] components admits infinite valid solutions, highlighting the difficulty of achieving robust reflection removal from a single image [11, 17, 27, 29, 30, 33, 39, 87, 90, 92]. As shown in the view 1 input image in Figure 1, even human observers may find it challenging to distinguish between reflection and transmission, especially with structurally complex or high-intensity reflections.

To mitigate this ill-posedness, researchers have explored various auxiliary constraints with varying levels of operational complexity. Hardware-based methods utilize polarization filters [44], near-infrared imaging [22, 23], flash/no-flash

† Corresponding author. [1] State Key Laboratory of Multimedia Information Processing, National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

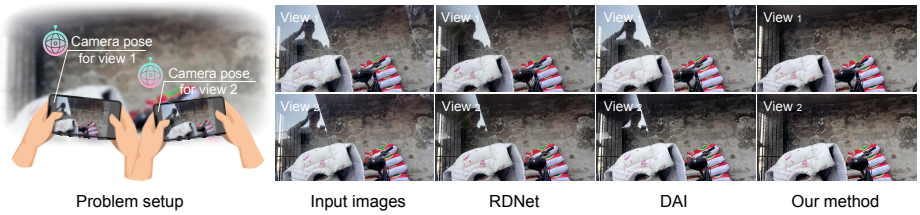


Fig. 1: Dual-view reflection removal with camera pose guidance. **Left:** Problem setup showing sequential capture of two viewpoints with camera poses tracked via device gyroscope. **Right:** Comparison of reflection removal results on challenging real-world data with overexposed reflections. The upper and bottom rows display views 1 and 2, respectively. While state-of-the-art single-image methods, such as RDNet [90] and DAI [27], fail to remove intense reflections when processing each view independently, our dual-view diffusion approach jointly processes both views to achieve effective reflection removal while preserving background fidelity.

pairs [21, 38], or panoramic capture [24, 25] to provide reliable physical cues, albeit at the cost of specialized equipment and complex user operations, which limits their practical deployment. Recent approaches incorporate language guidance [26, 91] to offer greater flexibility while still requiring user input to describe desired outputs. Another branch of research lies in multi-view reflection removal, which leverages image sequence input [1, 46, 58] based on deliberate camera movements, yet requiring capturing and processing multiple frames, significantly increasing both computational complexity and user effort. Modern smartphones equipped with dual-camera systems and Live Photo capabilities [2] can capture image pairs seamlessly, often without the user's awareness. Hence, we simplify the multi-frame setting to a user-friendly dual-view reflection removal.

However, transitioning to a dual-view setup drastically reduces the redundancy in observations across multi-frame sequences. To compensate, a model must fully exploit cross-view correspondences. Previous multi-view methods typically build upon explicit optical flow estimation [46] or 3D reconstruction [86]. Yet, these explicit techniques are highly susceptible to severe reflection interference and frequently fail in a constrained dual-view setting. Furthermore, because the physical transmission and virtual reflection layers reside at different depths, both exhibit distinct depth-induced parallax shifts. A robust framework must learn to disentangle these overlapping motion patterns while adapting to the varying viewpoint baselines inherent to mobile capture systems.

To address these challenges, we present an end-to-end dual-view diffusion framework that effectively separates reflections from transmissions without relying on fragile explicit correspondence matching or rigid 3D constraints. First, we propose a frame-aware spatial attention mechanism that implicitly captures cross-view correlations. Unlike naive joint attention, which indiscriminately mixes features and amplifies mutual noise across views, our design enforces a directed attention pattern: noisy transmission queries attend exclusively to clean mixture conditions. This elegantly bypasses error-prone optical flow and strictly prevents

cross-view noise leakage, all while seamlessly reusing powerful pre-trained generative priors. Second, inspired by how human observers shift their viewpoint to distinguish physical depth from reflections, we introduce a camera pose encoding. Rather than enforcing computationally expensive and brittle 3D hard constraints, we inject the relative camera pose as a soft conditional embedding. This lightweight global prior guides the network to disentangle depth-induced parallax motion, while exhibiting strong robustness against the inherent drift and noise of ubiquitous smartphone IMU/gyroscope sensors, as shown in Figure 1.

To enable training and evaluation of our approach, we construct DualRef, a comprehensive dataset addressing the scarcity of dual-view reflection removal data. DualRef comprises semi-synthetic image pairs for training and testing, along with real-captured dual-view videos with reflections, which feature ground truth captured with a dual-camera dolly system and synchronized pose tracking, providing both training data and real-world evaluation benchmarks.

Our principal contributions are:

- A diffusion-based reflection removal method designed specifically for the dual-view setting by proposing:
  - a frame-aware attention mechanism for effective implicit cross-view feature correlation that prevents cross-view noise leakage and
  - a robust camera pose integration scheme through learned embeddings that disentangles depth-induced parallax motion.
- DualRef dataset comprising semi-synthetic image pairs and real-world evaluation sequences with ground truth.

Extensive experiments demonstrate that our dual-view approach achieves remarkable results, surpassing single-image methods by over 10% in PSNR and comparable to complex multi-frame sequence methods.

2 Related work

Reflection removal. Deep learning approaches have substantially advanced SIRR [11, 12, 17, 27–30, 33, 39, 60, 64, 65, 73, 74, 82, 87, 89, 90]. Recently, diffusion-based methods [13, 26, 27, 33, 70, 87] that leverage powerful generative priors to handle ambiguous decomposition problems have emerged, with DAI [27] achieving single-step generation and L-DiffER [26] incorporating language guidance. However, because SIRR is fundamentally ill-posed, these methods still struggle with complex real-world reflections. Researchers also leveraged auxiliary information to address this ambiguity at the cost of requiring specialized hardware: polarization [34,37,43,44], NIR imaging [22,23], flash/no-flash pairs [21,35,38,71], dual-pixel cues [51, 85, 86], and wide field-of-view capture [19, 24, 25, 48].

Multi-view strategies exploit geometric cues from viewpoint variation. Early methods [16, 18, 81] demonstrated layer separation through motion analysis. Burst photography [9,50] demands less capture effort but provides limited parallax. Image sequence methods using deliberate camera motion [1, 46, 58] increase parallax but introduce computational complexity during inference. Dual-view



Fig. 2: Overview of the DualRef dataset. **Left:** Examples from DualRef-R showing semi-synthetic dual-view pairs with realistic and often structurally complex or high-intensity reflections generated through video-based synthesis and physics-based rendering. **Center:** Representative examples from DualRef-V demonstrating challenging real-world sequences with diverse scenes, strong reflection intensities, and complex reflection attributes. **Right:** Distribution of frame displacement between views in DualRef-R, centered around 50 pixels but spanning from near-zero (burst-like) to large displacements, providing diverse parallax patterns for enhanced generalization.

setups offer a balance; however, existing methods still face challenges: optical flow accumulates alignment errors, while 3D reconstruction [52, 86] requires additional depth sensing. By contrast, our method implicitly builds cross-frame correspondence via the proposed frame-aware spatial attention mechanism.

Diffusion models for image restoration. Diffusion models [10, 20, 59] excel at learning complex data distributions through progressive denoising, demonstrating remarkable capabilities in addressing under-constrained inverse problems [31, 55]. Applications span diverse restoration tasks: motion deblurring [77], resolution enhancement [53, 68, 72], colorization [6, 7, 75, 76], content completion [42, 80], and quality improvement [69, 79]. Their ability to preserve fine details while resolving structural ambiguities makes them suited for reflection removal. Within a diffusion framework, we propose a novel cross-view attention mechanism that prevents cross-view noise leakage while effectively reusing powerful generative priors.

3 Dataset

Existing reflection removal datasets [4, 13, 39, 65, 83, 89] primarily focus on single-view images. Furthermore, current dual-view datasets [46] are not fully open-source, and multi-view datasets [9] lack sufficient scale, significantly limiting the development of multi-view approaches. To bridge this gap, we propose DualRef, a comprehensive dual-view reflection removal dataset comprising: (1) **DualRef-R**: 16,000 semi-synthetic pairs with diverse motion patterns and reflection intensities for training and evaluation; and (2) **DualRef-V**: 195 real-world sequences captured using smartphones for evaluation. DualRef compre-

hensively covers diverse reflection scenarios featuring various intensities, blurs, and geometric distortions.

3.1 Semi-synthetic data generation

We develop comprehensive synthesis methods that faithfully model real-world reflection physics. Our DualRef-R dataset comprises 16,000 semi-synthetic pairs generated through two complementary approaches, with example images displayed in Figure 2 (left panel). The distribution of frame displacement (right panel) spans from near-zero to significant shifts, providing diverse parallax patterns for enhanced generalization.

Video-based synthesis: We leverage dense video sequences with camera trajectories, sampling frame pairs with controlled separation to ensure sufficient baseline while maintaining scene overlap. The synthesis process models realistic reflection-transmission mixing using the standard optical model with gamma correction, Gaussian blur, and ghosting masks. We collect 4,000 high-quality, pure-reflection images as reflection sources. The reflection layer exhibits view-dependent motion through homography warping with stochastic perturbation.

Physics-based rendering: We generate 4,000 dual-view pairs with precise ground truth through physically-based rendering. Our setup models realistic glass reflection scenarios (50% indoor-to-outdoor, 30% outdoor-to-indoor, 20% uniform environments) with realistic optical properties (IOR = 1.43) and 2-4cm baseline distances plus random perturbations to simulate handheld capture.

3.2 Real-world evaluation data

Evaluating real-world performance is critical, but is hindered by the lack of dual-view and video datasets with reflection ground truth, mainly because ensuring trajectory reproducibility is difficult. To address this, we introduce DualRef-V, comprising 195 challenging real-world sequences (1200×650 resolution, 150 frames at 30fps, with synchronized 10Hz IMU & Gyroscope data). To obtain precise ground truth, we develop a trajectory-aligned capture system (Figure 3).

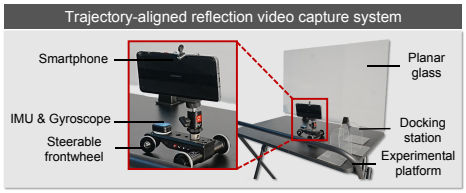


Fig. 3: Trajectory-aligned video capture system for DualRef-V. A programmable dolly ensures reproducible camera trajectories. It captures paired mixture and transmission-only ground truth by repeating paths with and without a removable glass panel, while an IMU & Gyroscope module records motion data.

Our system has a programmable camera dolly equipped with a steerable wheel and high-precision IMU & Gyroscope, starting from a fixed docking station to ensure exact trajectory reproducibility. The collection protocol involves executing a programmed path to record the mixture sequence through a removable glass panel, and subsequently repeating the identical trajectory without the glass to capture the transmission-only ground truth. Finally, frame and pose data

adjustments are applied to align the pairs and compensate for minor trajectory deviations.

The center panel of Figure 2 showcases representative examples from DUAL-REF-V, demonstrating diverse scenes, strong reflection intensities, and complex reflection attributes. These challenging real-world sequences span diverse environments, including outdoor natural settings, vehicles, and architectural scenes with varying lighting conditions and reflection complexity. This setup enables the accurate quantitative evaluation of dual-view methods on real data with ground truth—a capability previously unavailable—and can be extended to support the evaluation of multi-view reflection removal methods, as well as Live Photo settings, through its continuous trajectory recording.

4 Method

As is shown in Figure 4, given a pair of mixture images $\mathbf{M}_1, \mathbf{M}_2$ and their corresponding camera poses $\mathbf{p}_1, \mathbf{p}_2$, our goal is to recover the clean transmission layers $\mathbf{T}_1, \mathbf{T}_2$ governed by the physical model $\mathbf{M}_i = \mathbf{T}_i + \mathbf{R}_i$. Building upon the Stable Diffusion v2.1 [53] architecture, we map the images into the latent space using a frozen VAE. Our modified U-Net then processes a unified joint representation, constructed by concatenating the noisy transmission latents and the clean mixture conditions from both views, and denoises both views simultaneously. We introduce our proposed frame-aware attention mechanism in Section 4.1 and camera pose integration scheme in Section 4.2.

4.1 Frame-aware spatial attention

Transitioning from multi-frame sequences to a dual-view setup inherently reduces the available redundancy in observations. To separate reflections from transmissions effectively, it is crucial to maximize the utility of cross-view motion correlations (i.e., parallax). Previous multi-view reflection removal methods [46, 81] typically rely on explicit optical flow to align frames. However, flow estimation is highly susceptible to errors when corrupted by severe reflection interference.

Within a diffusion framework, a natural alternative to explicit flow estimation is leveraging self-attention mechanisms to implicitly capture cross-view correspondences. Yet, directly adapting existing multi-view attention designs presents significant drawbacks. Naively extending standard self-attention to joint cross-view attention [14, 56] introduces an unrestricted bidirectional flow. In our dual-view setup, this inherently allows noisy latents to inappropriately attend to each other across views, mutually amplifying noise and disrupting the denoising process. Besides, adopting ResNet 3D layers [45] alters the inherent feature dimensions, which prevents the reuse of powerful pre-trained generative priors.

To overcome these limitations, we propose a frame-aware spatial attention mechanism that enables robust cross-view reasoning while strictly preventing noise propagation. As shown in Figure 4 (bottom right), our frame-aware spatial attention is applied to the 16×16 and 32×32 upsampling layers of the U-Net [54].

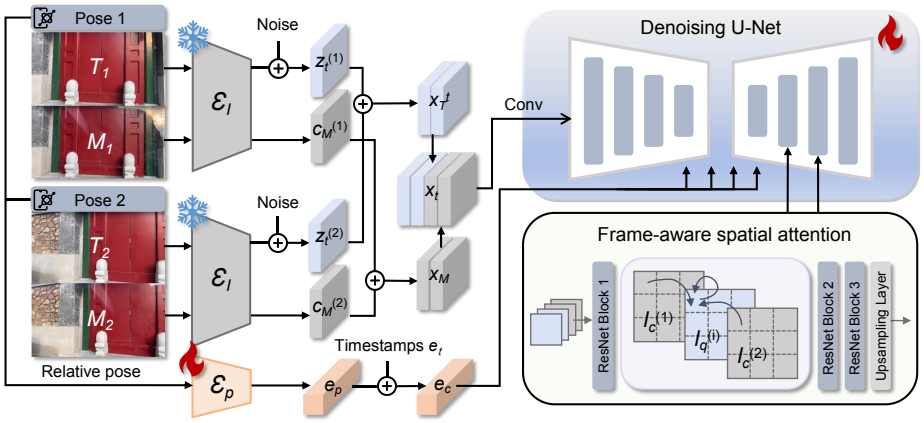


Fig. 4: Training pipeline of our dual-view diffusion framework. **Left:** Mixture and transmission images are encoded via frozen VAE. Transmission latents are noised while mixture latents remain clean, then concatenated channel-wise to form a 16-channel input to the modified U-Net [54]. Camera poses are converted to relative poses and encoded via an MLP, then combined with a timestep embedding for conditioning. **Bottom right:** Frame-aware spatial attention at 16×16 and 32×32 layers enables each noisy patch to attend to all mixture patches from both views, achieving cross-view exchange while reusing pretrained weights.

At these layers, the hidden features have been processed into 4 components corresponding to 2 views, each containing noisy and clean paths. Given the feature tensor $\mathbf{l} \in \mathbb{R}^{B\times 4C'\times H\times W}$ where C' is the feature dimension at this layer, we first separate it into query and condition components:

$$\mathbf{l}_q = \mathbf{l}[:, :2C', :, :], \quad \mathbf{l}_c = \mathbf{l}[:, 2C' :, :, :], \tag{1}$$

where $\mathbf{l}_q \in \mathbb{R}^{B\times 2C'\times H\times W}$ contains features from noisy transmission paths of both views, and $\mathbf{l}_c \in \mathbb{R}^{B\times 2C'\times H\times W}$ contains features from clean mixture paths. We then reshape these into a sequence format for attention computation:

$$\mathbf{l}_q \to \text{reshape} \to \mathbf{s}_q \in \mathbb{R}^{B\times 2HW\times C'}, \mathbf{l}_c \to \text{reshape} \to \mathbf{s}_c \in \mathbb{R}^{B\times 2HW\times C'}. \tag{2}$$

We concatenate query and condition sequences along the sequence dimension and apply the pretrained projection matrices $\mathbf{P}_q$, $\mathbf{P}_k$, $\mathbf{P}_v$ from SD v2.1 [53]:

$$\mathbf{s}_{\text{concat}} = [\mathbf{s}_q; \mathbf{s}_c] \in \mathbb{R}^{B\times 4HW\times C'}, \tag{3}$$

$$[\mathbf{Q}_{\text{all}}, \mathbf{K}_{\text{all}}, \mathbf{V}_{\text{all}}] = [\mathbf{P}_q, \mathbf{P}_k, \mathbf{P}_v] \cdot \mathbf{s}_{\text{concat}}, \tag{4}$$

where each of $\mathbf{Q}_{\text{all}}, \mathbf{K}_{\text{all}}, \mathbf{V}_{\text{all}} \in \mathbb{R}^{B\times 4HW\times C'}$. Then we extract queries only from positions corresponding to noisy features, while keys and values come only from positions corresponding to clean conditions:

$$\mathbf{Q} = \mathbf{Q}_{\text{all}}[:, :2HW, :], \mathbf{K} = \mathbf{K}_{\text{all}}[:, 2HW :, :], \mathbf{V} = \mathbf{V}_{\text{all}}[:, 2HW :, :]. \tag{5}$$

This creates a directed attention pattern where each element in the query sequence (from noisy transmissions) attends to all elements in the key/value sequence (from clean mixtures):

$$\text{Attn}(\mathbf{s}_q, \mathbf{s}_c) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{6}$$

The resulting attention output is added to the query features via a residual connection, reshaped back to the spatial image format, and concatenated with the unchanged condition features.

4.2 Camera pose encoding

When a human observer struggles to separate reflections from transmissions in a complex scene, they naturally shift their viewpoint. By comparing their own movement against the apparent motion of objects in the scene, they can effortlessly distinguish the real background (transmission) from virtual images (reflection) through depth-induced parallax.

Inspired by this, we introduce relative camera pose encoding into our dual-view framework. This encoding serves as a global geometric prior, informing the network of the exact camera motion and guiding spatial attention to features that adhere to the transmission parallax, while suppressing reflection features whose apparent motion violates this transformation.

Given poses $\mathbf{p}_1 = [\mathbf{t}_1, \mathbf{q}_1]^\top$ and $\mathbf{p}_2 = [\mathbf{t}_2, \mathbf{q}_2]^\top$, where $\mathbf{t}_i \in \mathbb{R}^3$ represents translation and $\mathbf{q}_i \in \mathbb{R}^4$ is the unit quaternion for rotation, we compute the relative pose:

$$\mathbf{p}_{\text{rel}} = \left[\frac{\mathbf{t}_2 - \mathbf{t}_1}{\|\mathbf{t}_2 - \mathbf{t}_1\|}, \mathbf{q}_2 \otimes \mathbf{q}_1^{-1}\right]. \tag{7}$$

We normalize the translation vector to a unit direction because IMU-derived translations suffer from magnitude drift, whereas rotations and translation directions remain highly reliable.

To integrate this 7-dimensional relative pose, we process it through a lightweight 2-layer MLP (hidden dimension 256) to produce the pose embedding $\mathbf{e}_{\text{p}} = \text{MLP}(\mathbf{p}_{\text{rel}})$. This is then fused with the time embedding $\mathbf{e}_{\text{t}}$ to form the combined conditioning signal $\mathbf{e}_{\text{c}} = \mathbf{e}_{\text{t}} + \gamma\mathbf{e}_{\text{p}}$. We empirically set the scaling factor $\gamma = 0.15$ to prevent the pose information from overwhelming the temporal diffusion signal while maintaining sufficient geometric guidance.

Our approach fundamentally differs from prior burst-based methods [9] that rely on explicit 3D projections during test-time optimization. Because they enforce strict geometric constraints, these methods are computationally expensive, highly sensitive to noise, and require high-precision IMUs. In contrast, we inject camera pose as a soft conditional embedding within an end-to-end diffusion pipeline (Figure 4). This formulation provides robust geometric guidance that effectively tolerates the inherent drift and noise of common smartphone sensors.

4.3 Training strategy

We discover that training exclusively on simple reflection scenarios produces models unable to handle strong reflections, such as overexposed or structurally intense reflections. Conversely, directly training on challenging cases prevents the model from learning to leverage parallax information for understanding the physical principles underlying reflection-transmission decomposition. We therefore adopt two-stage curriculum learning that progressively increases difficulty.

Stage 1: Foundation training. We begin with semi-synthetic data at 768×768 resolution with moderate reflection intensity $\alpha_r \sim \mathcal{U}(0.3, 0.7)$. Crucially, we apply stochastic augmentations to enhance model robustness: with 15% probability, we replace view 2 with view 1's content, and with another 15% probability, we replace view 1 with view 2's content. In both cases, the relative pose is set to the identity (zero) since the views become identical. Additionally, when both views are distinct (in 70% of cases), we apply pose dropout with a 25% conditional probability, removing pose information to encourage the model to rely on visual features as a fallback. These augmentations enable the model to handle both single-view and dual-view scenarios gracefully while establishing core capabilities for reflection-transmission decomposition.

Stage 2: Robustness enhancement. We increase reflection intensity $\alpha_r \sim \mathcal{N}(0.8, 0.2)$ clipped to $[0.4, 1.2]$ and introduce diverse augmentations including motion blur, chromatic aberration, and ghosting effects. This stage fine-tunes the model for challenging scenarios where reflections may dominate transmissions. The training objective is the standard DDPM loss [20] with conditional dropout:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon},m_v,m_p}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_T^t, t, m_v \odot \mathbf{c}_M, m_p \cdot \mathbf{e}_{\text{pose}})\|_2^2\right], \tag{8}$$

where $m_v \in \{0,1\}^2$ and $m_p \in \{0,1\}$ are dropout masks for views and poses.

5 Experiments

5.1 Experimental setup

Datasets. We evaluate our method quantitatively on three datasets, namely (1) our DualRef-V: 195 real-world sequences featuring challenging scenarios, including museums, storefronts, and vehicles; (2) DualRef-R test split: 2819 semi-synthetic pairs (2014 video-based synthesis, 805 physics-based rendering) with precise ground truth; (3) Niklaus dataset [46] test split: 20 real-captured dual-view pairs. Our model is trained on the DualRef-R training split comprising 13,181 pairs. During evaluation, we sample 500 dual-view pairs from 195 captured sequences to approximately form a Gaussian distribution with a mean view shift of 50 pixels and a variance of 40, in accordance with the view distance of common dual-camera systems. Furthermore, we evaluate our method on images from various sources (the Chugunov dataset [9], in-the-wild Internet data, and real-captured photos) with diverse viewpoint distances.

Baselines. We compare against state-of-the-art single-image reflection removal methods including ZF24 [92], DSIT [30], RDNet [90], DAI [27], WSeat [87], and

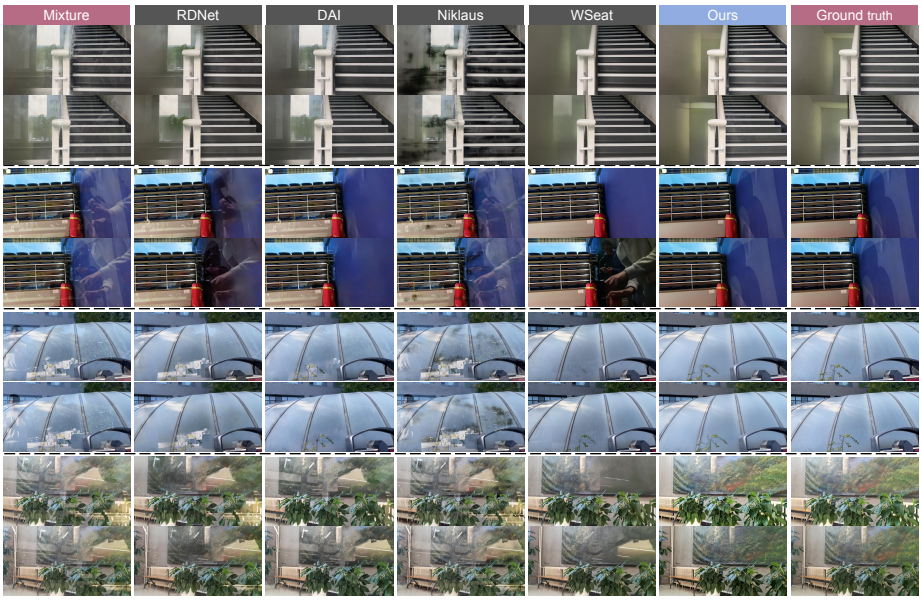


Fig. 5: We compare with RDNet [90], DAI [27], Niklaus [46] and Weat [87] on challenging real-world scenarios in our DualRef-V dataset. For each scene divided by dashed lines, the upper row shows view 1, and the bottom row shows view 2. Readers are encouraged to zoom in for best viewing.

multi-view methods TF15 [81], Niklaus [46]. For a fair comparison, we also evaluate fine-tuned versions (marked with *) of methods on our training dataset (both views are separately used during finetuning of SIRR methods). L-DiffER [26], Kee *et al.* [33], and Prasad *et al.* [50] are not open-source, yet we compare with Adobe Photoshop 2026, which is linked to Kee *et al.* [33], along with other general-purpose image-editing models [15, 47] in our supplementary materials.
Metrics. We employ five metrics: SSIM and PSNR for structural similarity and signal quality; LPIPS [88] for perceptual distance; CLIPIQA [67] and MUSIQ [32] for no-reference image quality assessment. These metrics collectively evaluate pixel accuracy and perceptual quality.

5.2 Comparison to state-of-the-arts

As shown in Table 1, our method achieves superior performance across most evaluation metrics, especially on real-world data (DualRef-V and the Niklaus [46] dataset), demonstrating the effectiveness of leveraging dual-view information through our diffusion framework.

As shown in Figure 5 and Figure 6, single-image methods [27, 30, 87, 90] often fail when confronted with overexposed or structurally complex reflections, sometimes even misidentifying reflections as transmission and enhancing them rather than removing them. Diffusion-based SIRR approaches like DAI [27] and

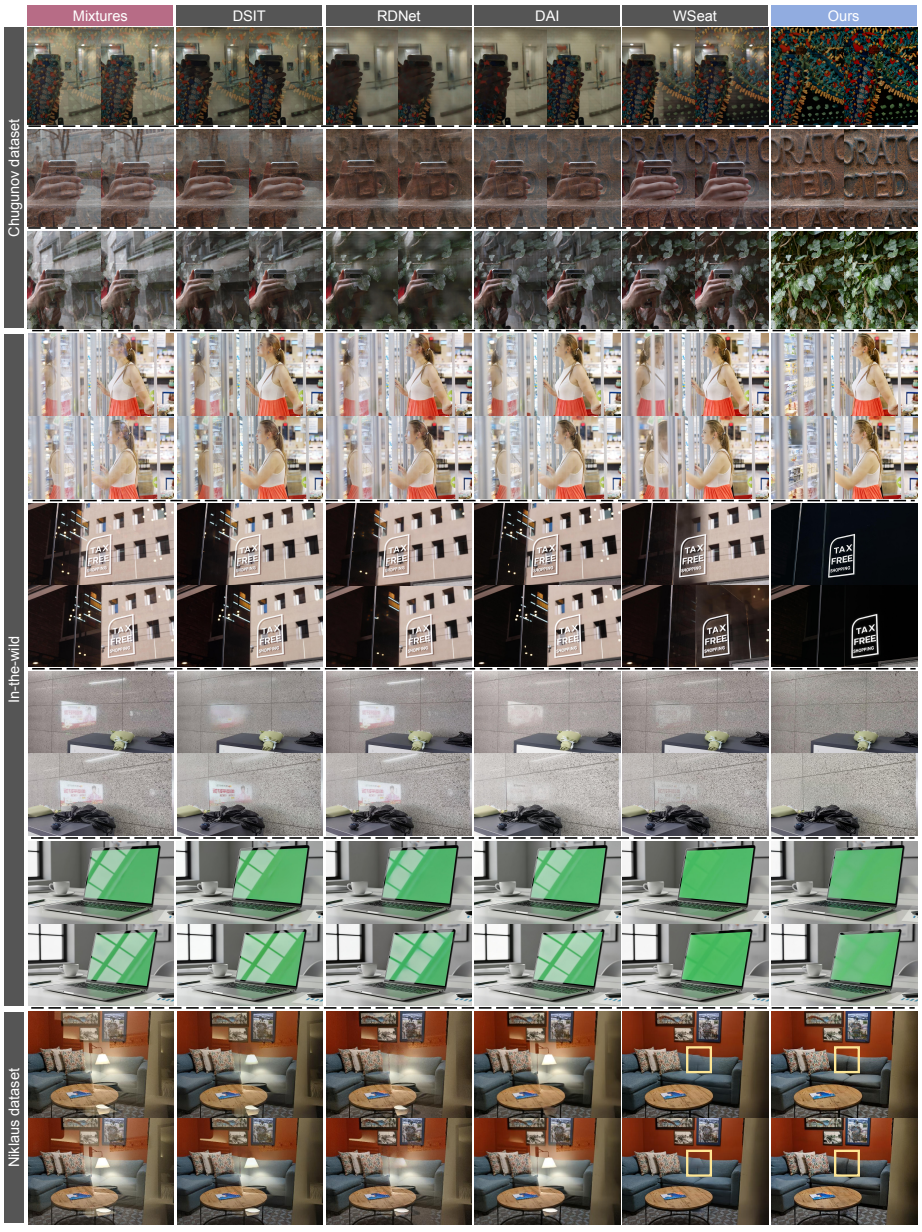


Fig. 6: Qualitative comparison on challenging real-world scenarios collected from diversified sources, including the Chugunov dataset [9], the Niklaus dataset [46], and videos downloaded from the Internet. For each scene separated by dashed lines, the upper row (or left column) and the bottom row (or right column) correspond to views 1 and 2, respectively. Our dual-view diffusion framework effectively removes complex and spatially varying reflections while faithfully preserving fine transmission details and scene structures. The images are best viewed zoomed in for a clearer comparison of local details. Please refer to our supplementary material for more results.

Table 1: Quantitative comparison on DualRef and the Niklaus dataset [46]. Our method demonstrates superior performance across both real-world and semi-synthetic scenarios. * indicates methods fine-tuned on our training data. **Bold** and underlined values denote best and second-best results.

Dataset	Metric	TF15 [81]	ZF24 [92]	ZF24*	DSIT [30]	DSIT*	RDNet [90]	DAI [27]	WSeat [87]	Niklaus* [46]	Ours
DualRef-V	SSIM↑	0.6854	0.7300	0.7317	0.7285	0.7481	0.7342	0.7354	0.7478	0.7062	**0.7591**
	PSNR↑	18.65	19.80	20.17	19.54	20.55	20.24	21.57	21.44	18.36	**24.17**
	LPIPS↓	0.2243	0.1923	0.2005	0.2072	0.1975	0.2014	0.1762	0.1692	0.2351	**0.1638**
	CLIPIQA↑	0.3762	0.4420	0.4538	0.4550	0.4527	0.4569	0.4431	0.4691	0.3759	**0.4705**
	MUSIQ↑	61.77	63.53	64.01	63.56	63.52	64.22	62.89	64.00	59.33	**64.81**
DualRef-R	SSIM↑	0.7868	0.8273	0.8359	0.8337	**0.8389**	0.8362	0.8274	0.8369	0.8169	0.8378
	PSNR↑	23.75	25.64	26.61	26.43	26.71	26.60	26.66	27.11	25.35	**28.19**
	LPIPS↓	0.1523	0.1133	0.1045	0.1003	0.0990	0.1017	0.1102	0.1448	0.1410	**0.0983**
	CLIPIQA↑	0.3425	0.3778	0.3690	0.3575	0.3705	0.3826	0.4024	0.3981	0.3618	**0.4384**
	MUSIQ↑	44.32	46.26	45.94	47.07	46.85	48.25	47.11	47.90	47.52	**48.76**
Niklaus [46]	SSIM↑	0.7575	0.7726	0.7743	**0.7999**	0.7971	0.7969	0.7842	0.7893	0.7653	0.7973
	PSNR↑	21.76	22.61	22.88	22.71	23.40	23.27	23.33	23.72	22.60	**25.09**
	LPIPS↓	0.1432	0.1209	0.1185	0.1168	0.1154	0.0934	0.1813	0.1669	0.1112	**0.0865**
	CLIPIQA↑	0.4882	0.4952	0.4976	0.4676	0.4712	**0.5989**	0.3461	0.5909	0.4946	0.5937
	MUSIQ↑	51.94	52.37	52.66	53.29	53.57	52.71	50.81	56.31	55.68	**57.27**

Table 2: Ablation study on core components. The first and second set of experiments ablate the pose encoding and SD2.1's generative prior, respectively. The third set of experiment compares different attention mechanisms.

Configuration	DualRef-V					DualRef-R				
	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Single-view	21.59	0.7236	0.1720	0.4571	62.13	26.07	0.8235	0.1280	0.4096	45.27
Dual separate	21.44	0.7245	0.1703	0.4597	63.77	26.12	0.8219	0.1248	0.4231	45.12
Dual w/o pose	23.53	0.7468	0.1694	0.4632	63.89	27.47	0.8310	0.1176	0.4270	47.26
SD2.1+RAFT	21.60	0.7369	0.1744	0.4588	63.21	26.33	0.8226	0.1265	0.4178	45.19
Cat3D-Style Attn	22.09	0.7267	0.1717	0.4588	62.40	26.52	0.8230	0.1244	0.4133	45.68
SIR-Style Attn	20.87	0.7036	0.2069	0.3678	55.32	25.99	0.7963	0.1452	0.3685	43.20
Width-concat	22.03	0.7354	0.1838	0.4234	58.79	25.83	0.8220	0.1392	0.4002	42.99
Ours	**24.17**	**0.7591**	**0.1638**	**0.4705**	**64.81**	**28.19**	**0.8378**	**0.0983**	**0.4384**	**48.76**

WindowSeat [87] tend to alter content structure. In contrast, our method robustly handles these challenging scenarios without such artifacts while maintaining user-friendliness by using dual-camera systems that require only a single shutter press. As shown in Figure 5, optical flow-based dual-view methods [46,81] struggle with strong reflections, which hinder accurate flow estimation, leading to error propagation and unsatisfactory results. In contrast, our approach circumvents these issues. Furthermore, our method generalizes well to challenging in-the-wild scenes with diverse cross-view distances, which robustly adapts to various dual-camera devices.

5.3 Ablation study

As shown in Table 2, Figure 7 and Figure 8, we conduct comprehensive ablation studies using models trained on the DualRef-R train split and evaluated on both DualRef-R test split and DualRef-V. Our analysis is twofold: verifying the core components and validating our specific architectural design.

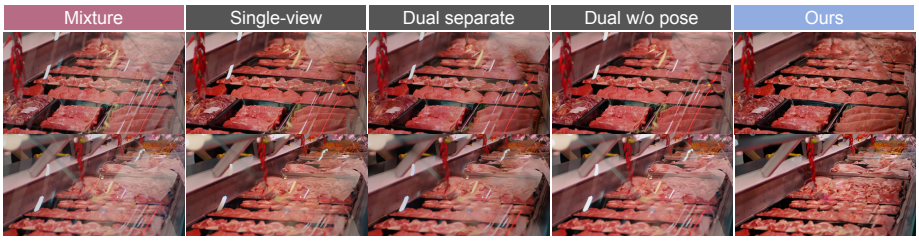


Fig. 7: Visual results for our ablation study. The upper row and the bottom row display views 1 and 2, respectively. Single-view denotes directly fine-tuning InstructPix2Pix [3] on DualRef-R dataset (considering dual-view pairs as two separate single-views). Dual separate denotes the input of two identical images. Dual w/o pose denotes removing the pose encoding from our method.



Fig. 8: Analysis of explicit flow on the Burst dataset [9]. We compare against the SD2.1+RAFT and TF15 [81] baseline with different number of input frames. Explicit flow becomes unreliable in reflection-corrupted regions and needs more frames to reduce ambiguity, whereas our method produces robust results from only two views.

Analysis of pose encoding. The Single-view baseline (fine-tuned InstructPix2Pix without cross-view attention) and Dual separate (identical inputs with zero relative pose) both achieve strong baseline performance, proving that our base diffusion pipeline is inherently well-suited for reflection removal. Upgrading to Dual w/o pose brings clear quantitative gains, confirming that the network effectively leverages dual-view parallax to suppress reflections. Finally, our full model (**Dual w/ pose**) yields the highest performance across all metrics. As shown in Figure 7, Dual w/o pose fails to suppress intense reflections when they also demonstrate distinct parallax, as they become easily confused with the physical transmission. However, by informing the model of the actual viewpoint motion via our pose encoding integration scheme, the reflections are robustly separated through the correlation between camera trajectory and scene shift. This clearly validates our design choice. Readers are referred to the supplementary material for evaluations of other pose integration methods.

Analysis of explicit flow. To isolate the impact of SD2.1's pre-trained generative prior, we built an SD2.1+RAFT baseline using identical training settings, using RAFT [61] to warp view 2 to view 1 before feeding the channel-wise concatenated latents into the U-Net, which consistently underperforms ours, as shown in Table 2. While explicit flow assumes a unique motion per pixel, reflection-degraded regions exhibit dual motions (transmission and reflection),

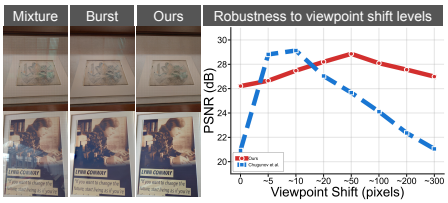


Fig. 9: Left: Comparison with state-of-the-art burst method [9] on their dataset. Our method achieves comparable or better quality while running 60× faster. **Right:** Performance across different viewpoint shift levels on DualRef-R test set, demonstrating our method's robustness compared to burst approaches.

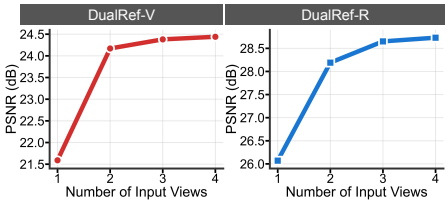


Fig. 10: Performance analysis with varying numbers of input views. We observe substantial improvements from single-view to dual-view, whereas additional views offer diminishing returns with minimal marginal gains. Our dual-view setup achieves an optimal balance between quality gains and capture complexity.

creating fundamental correspondence ambiguity. Our frame-aware attention bypasses this by implicitly aggregating evidence from multiple plausible matches without explicit commitment. Furthermore, flow-based methods like TF15 [81] require multiple frames to resolve such ambiguity and are particularly error-prone in the dual-view regime (Figure 8), where ours remains robust.

Analysis of cross-view attention mechanisms. We compare our frame-aware spatial attention against recent multi-view diffusion paradigms. The Cat3D-style baseline [14, 56] (naive channel-wise concatenation) suffers from cross-view feature ambiguity and noise leakage, visually manifesting as incomplete reflection removal and suboptimal quantitative metrics. Even worse, the SIR-style approach [45] (incorporating 3D ResNet blocks) aggressively alters the network topology, thereby destroying the delicate pre-trained 2D generative priors. Width-concat [78] converges slower and yields inferior results (Table 2). Unlike [78], inputs are inherently pixel-aligned in our task, where channel-wise fusion better exploits this direct correspondence. In contrast, our zero-overhead, directed attention perfectly preserves image fidelity while maximizing cross-view disambiguation. We also conduct ablation studies on the Niklaus dataset [46] with Dual w/o pose mode since pose data is not provided in this dataset. Please see our supplementary materials.

5.4 Discussions

Robustness to different levels of viewpoint shift. Figure 9 compares our method with the state-of-the-art burst approach [9]. The visual comparison (left) shows that our method achieves comparable or superior quality on their dataset while being significantly more efficient—requiring only 3 seconds per scene on an RTX 6000, compared to 3 minutes for burst methods, which require scene-specific optimization. The quantitative analysis (right) evaluates performance across varying viewpoint shifts on DualRef-R test set. While burst methods

perform well at minimal displacements but degrade rapidly with larger shifts or single-view inputs, our approach maintains robust performance across all displacement levels, validating its robustness to different dual cameras with varying view distances. This robustness stems from our diverse training data and the synergy between frame-aware spatial attention and pose encoding.

Effectiveness of dual-view setup. We investigate the optimal number of views for reflection removal by training models with different numbers of input views on our dataset. Figure 10 reveals an interesting pattern: performance improves substantially from single-view to dual-view, but additional views beyond two provide diminishing returns. The marginal improvement from 2 to 4 views is minimal, while the capture complexity and inference time increase linearly. This validates our dual-view design as a reasonable balance—it achieves most of the quality gains possible from multi-view information while maintaining practical usability. Users need only a single shutter press with dual-camera systems, which is common in modern mobile devices, avoiding the operational burden of multi-frame sequences while still achieving significant improvements in reflection removal over single-image methods.

6 Conclusion

We present a dual-view diffusion framework for reflection removal that addresses the fundamental ill-posedness of single-image decomposition through frame-aware spatial attention and camera pose encoding. With dual-camera systems common in current devices, our approach provides a reasonable balance between capture simplicity and removal quality—requiring only a single shutter press with ubiquitous dual-camera systems on mobile device while achieving substantial improvements over single-image methods. We introduce DualRef, the first comprehensive dual-view dataset with trajectory-aligned dual-view real-world sequences. Extensive experiments demonstrate superior performance across diverse scenarios, validating dual-view reflection removal as a practical paradigm for mobile photography.

Limitations. Our approach relies on effective parallax. In scenarios where viewpoint shift yields no or negligible parallax and insufficient geometric cues to disambiguate layer depths, the task reverts to an ill-posed single-image problem, and our performance naturally degrades to match single-image diffusion baselines. As a diffusion-based method, our model may also exhibit color drift in some challenging cases, although the dual-view cue can partially mitigate this issue by providing additional cross-view evidence. We further discuss our failure modes in our supplementary materials.

Acknowledgement. This work is supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012).

References

1. Alayrac, J.B., Carreira, J., Zisserman, A.: The visual centrifuge: Model-free layered video representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2457–2466 (2019)
2. Apple Inc.: Live photos. https://developer.apple.com/design/human-interface-guidelines/live-photos (2024), accessed: 2025-11-11
3. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
4. Cai, J., Yang, K., Ouyang, L., Fu, L., Ding, J., Shen, J., Meng, Z.: OpenRR-5K: A large-scale benchmark for reflection removal in the wild. arXiv preprint arXiv:2506.05482 (2025)
5. Chang, Y., Jung, C., Sun, J.: Joint reflection removal and depth estimation from a single image. IEEE Transactions on Cybernetics **51**(12), 5836–5849 (2020)
6. Chang, Z., Weng, S., Li, Y., Li, S., Shi, B.: L-CoDer: Language-based colorization with color-object decoupling transformer. In: European Conference on Computer Vision. pp. 360–375. Springer (2022)
7. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-CoIns: Language-based colorization with instance awareness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19221–19230 (2023)
8. Chaturvedi, S., Ren, M., Hold-Geoffroy, Y., Liu, J., Dorsey, J., Shu, Z.: Synthlight: Portrait relighting with diffusion model by learning to re-render synthetic faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 369–379 (2025)
9. Chugunov, I., Shustin, D., Yan, R., Lei, C., Heide, F.: Neural spline fields for burst image fusion and layer separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25763–25773 (2024)
10. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
11. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5017–5026 (2021)
12. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3238–3247 (2017)
13. Fang, S., Wang, Y., Zhang, J., Li, Z., Wang, Y.: Adaptive language-aware image reflection removal network. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 973–981 (2025). https://doi.org/10.24963/ijcai.2025/109
14. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3D: Create anything in 3D with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
15. Google DeepMind: Gemini 3 pro image. https://deepmind.google/models/gemini-image/pro/ (2026), accessed: 2026-02-27
16. Guo, X., Cao, X., Ma, Y.: Robust separation of reflection from multiple images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2187–2194 (2014)
17. Guo, Y., Lao, Z., Song, X., Zhou, Y., Yu, H.: Sirr-lmm: Single-image reflection removal via large multimodal model (2026), https://arxiv.org/abs/2601.07209

18. Han, B.J., Sim, J.Y.: Reflection removal using low-rank matrix completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5438–5446 (2017)
19. Han, B.J., Sim, J.Y.: Zero-shot learning for reflection removal of single 360-degree image. In: Proceedings of the European Conference on Computer Vision. pp. 533–548 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
21. Hong, Y., Chang, Y., Liang, J., Ma, L., Huang, T., Shi, B.: Light flickering guided reflection removal. International Journal of Computer Vision **132**(9), 3933–3953 (2024)
22. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with NIR and RGB image feature fusion. IEEE Transactions on Multimedia **25**, 7101–7112 (2022)
23. Hong, Y., Lyu, Y., Li, S., Shi, B.: Near-infrared image guided reflection removal. In: IEEE International Conference on Multimedia and Expo. pp. 1–6. IEEE (2020)
24. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Panoramic image reflection removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7762–7771 (2021)
25. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Par2Net: End-to-end panoramic image reflection removal. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12192–12205 (2023)
26. Hong, Y., Zhong, H., Weng, S., Liang, J., Shi, B.: L-DiffER: Single image reflection removal with language-based diffusion model. In: Proceedings of the European Conference on Computer Vision. pp. 58–76 (2024)
27. Hu, J., Yang, C., Zhou, Z., Fang, J., Yang, X., Tian, Q., Shen, W.: Dereflection any image with diffusion priors and diversified data. arXiv preprint arXiv:2503.17347 (2025)
28. Hu, Q., Guo, X.: Trash or Treasure? An interactive dual-stream strategy for single image reflection separation. Advances in Neural Information Processing Systems **34**, 24683–24694 (2021)
29. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13138–13147 (2023)
30. Hu, Q., Wang, H., Guo, X.: Single image reflection separation via dual-stream interactive transformers. Advances in Neural Information Processing Systems **37**, 55228–55248 (2024)
31. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022)
32. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
33. Kee, E., Pikielny, A., Blackburn-Matzen, K., Levoy, M.: Removing reflections from RAW photos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 161–171. Computer Vision Foundation / IEEE (2025)
34. Kong, N., Tai, Y.W., Shin, S.Y.: A physically-based approach to reflection separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9–16. IEEE (2012)
35. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14811–14820 (2021)

36. Lei, C., Huang, X., Qi, C., Zhao, Y., Sun, W., Yan, Q., Chen, Q.: A categorized reflection removal dataset with diverse real-world scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3040–3048 (2022)
37. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1750–1758 (2020)
38. Lei, C., Jiang, X., Chen, Q.: Robust reflection removal with flash-only cues in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15530–15545 (2023)
39. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3565–3574 (2020)
40. Li, X., Wang, Z., Huang, Z., Rehg, J.: Cue3D: Quantifying the role of image cues in single-image 3D generation. Advances in Neural Information Processing Systems **38**, 105199–105232 (2026)
41. Liang, Z., Yu, H., Li, B., Zhang, D., Cao, Z., Gong, T., Liu, J., Cui, S., Wang, F.: 3DReflecNet: A large-scale dataset for 3D reconstruction of reflective, transparent, and low-texture objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7244–7255 (2026)
42. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
43. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. Advances in Neural Information Processing Systems **32** (2019)
44. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Physics-guided reflection separation from a pair of unpolarized and polarized images. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 2151–2165 (2022)
45. Mao, Y., Wang, B., Kulkarni, N., Park, J.J.: SIR-DIFF: Sparse image sets restoration with multi-view diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21620–21630 (2025)
46. Niklaus, S., Zhang, X.C., Barron, J.T., Wadhwa, N., Garg, R., Liu, F., Xue, T.: Learned dual-view reflection removal. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3713–3722 (2021)
47. OpenAI: Openai chatgpt 5.2. `https://chatgpt.com//` (2025), accessed: 2026-02-27
48. Park, J., Kim, H., Park, E., Sim, J.Y.: Fully-automatic reflection removal for 360-degree images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1609–1617 (2024)
49. Poly Haven: Poly haven. `https://polyhaven.com/` (2026), accessed: 2026-06-22
50. Prasad, B., Mitra, K., 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)
51. Punnappurath, A., Brown, M.S.: Reflection removal using a dual-pixel sensor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1556–1565 (2019)
52. Qiu, J., Jiang, P.T., Zhu, Y., Yin, Z.X., Cheng, M.M., Ren, B.: Looking through the glass: Neural surface reconstruction against high specular reflections. In: Proceed-

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20823–20833 (2023)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
54. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
55. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
56. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: MVDream: Multi-view diffusion for 3D generation. arXiv preprint arXiv:2308.16512 (2023)
57. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3193–3201 (2015)
58. Simon, C., Kyu Park, I.: Reflection removal for in-vehicle black box videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4231–4239 (2015)
59. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
60. Song, Z., Zhang, Z., Zhang, K., Luo, W., Fan, Z., Ren, W., Lu, J.: Robust single image reflection removal against adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24688–24698 (2023)
61. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: Proceedings of the European Conference on Computer Vision. pp. 402–419 (2020)
62. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3922–3930 (2017)
63. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. International Journal of Computer Vision **129**(2), 385–399 (2021)
64. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: CoRRN: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(12), 2969–2982 (2019)
65. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1424–1441 (2022)
66. Wang, C., Xu, D., Wan, R., He, B., Shi, B., Duan, L.Y.: Background scene recovery from an image looking through colored glass. IEEE Transactions on Multimedia **25**, 2876–2887 (2022)
67. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
68. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
69. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)

70. Wang, T., Lu, W., Zhang, K., Luo, W., Kim, T.K., Lu, T., Li, H., Yang, M.H.: PromptRR: Diffusion models as prompt generators for single image reflection removal. arXiv preprint arXiv:2402.02374 (2024)
71. Wang, T., Xie, M., Cai, H., Shah, S., Metzler, C.A.: Flash-Split: 2D reflection removal with flash cues and latent diffusion separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5688–5698 (2025)
72. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: SinSR: Diffusion-based image super-resolution in a single step. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25796–25805 (2024)
73. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8178–8187 (2019)
74. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3771–3779 (2019)
75. Weng, S., Wu, H., Chang, Z., Tang, J., Li, S., Shi, B.: L-CoDe: Language-based colorization using color-object decoupled conditions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2677–2684 (2022)
76. Weng, S., Zhang, P., Li, Y., Li, S., Shi, B., et al.: L-CAD: Language-based colorization with any-level descriptions using diffusion priors. Advances in Neural Information Processing Systems **36**, 77174–77186 (2023)
77. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16293–16303 (2022)
78. Winter, D., Shul, A., Cohen, M., Berman, D., Pritch, Y., Rav-Acha, A., Hoshen, Y.: ObjectMate: A recurrence prior for object insertion and subject-driven generation. arXiv preprint arXiv:2412.08645 (2024)
79. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: DiffIR: Efficient diffusion model for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13095–13105 (2023)
80. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: SmartBrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
81. Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. ACM Transactions on Graphics **34**(4), 1–11 (2015)
82. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: Proceedings of the European Conference on Computer Vision. pp. 654–669 (2018)
83. Yang, K., Ouyang, L., Sun, H., Cai, J., Fu, L., Ding, J., Ho, C.M., Meng, Z.: OpenRR-1K: A scalable dataset for real-world reflection removal. arXiv preprint arXiv:2506.08299 (2025)
84. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: MANIQA: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
85. Yu, K., Han, M., Pan, L., Liu, L., Liu, M., Liang, W.: DCRR++: Unsupervised reflection removal and novel view synthesis via dual-pixel guided 3d gaussian splatting. International Journal of Computer Vision **134**(6), 284 (2026)

86. Yu, K., Pan, L., Liu, L., Liang, W.: Enhanced dual-pixel image reflection removal via gaussian splatting. In: Proceedings of the ACM International Conference on Multimedia. pp. 7766–7775 (2025)
87. Zakarin, D., Wandel, T., Obukhov, A., Dai, D.: Reflection removal through efficient adaptation of diffusion transformers. arXiv preprint arXiv:2512.05000 (2025)
88. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
89. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4786–4794 (2018)
90. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26430–26439 (2025)
91. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24913–24922 (2024)
92. Zhu, Y., Fu, X., Jiang, P.T., Zhang, H., Sun, Q., Chen, J., Zha, Z.J., Li, B.: Revisiting single image reflection removal in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25468–25478 (2024)

Diffusion-based Dual-view Reflection Removal

Supplementary Material

Tianyi Xu[1] Zifeng Wang[1] Boyang Lv[1] Shuchen Weng[2,1] Boxin Shi[1†]

[1]Peking University [2]Beijing Academy of Artificial Intelligence

photon@stu.pku.edu.cn shiboxin@pku.edu.cn

A Dataset details

A.1 Survey of existing datasets

The development of dual-view and sequence-based reflection removal methods is primarily hindered by the lack of a publicly accessible dataset of dual-view images or video clips containing reflections and their transmission ground truth, which is why we collected our DualRef dataset. We conduct a thorough survey of existing reflection-removal datasets, as shown in Table A1.

As evident from Table A1, the research community on reflection removal faces a critical shortage of multi-view datasets suitable for training and evaluating dual-view methods. Among the 15 surveyed datasets, 12 (80%) are designed exclusively for single-image methods, which fundamentally limits their applicability to multi-view approaches that leverage parallax information. While two dual-view datasets exist [9, 46], they present significant limitations: the Niklaus [46] dataset remains inaccessible to the research community, and the Burst [9] dataset, though publicly available, contains only 11 test sequences without corresponding ground truth annotations. This scarcity necessitated the creation of our DualRef dataset, enabling the systematic development and evaluation of dual-view reflection removal methods.

A.2 Details of dataset synthesis methods

We generate dual-view training pairs using two complementary approaches: video-based synthesis leveraging existing datasets and physics-based rendering for precise ground-truth generation.

Video-based synthesis from TartanAir. Our video-based synthesis pipeline leverages the dense video sequences and precise camera trajectories of the TartanAir dataset to generate realistic dual-view reflection pairs. The synthesis process involves three key components:

† Corresponding author. [1] State Key Laboratory of Multimedia Information Processing, National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

Pure reflection collection. We collected 4,000 high-quality pure reflection images using a custom portable apparatus. The setup consists of an adjustable camera mount, a planar glass sheet, and a black-light-absorbing backing. By blocking background illumination with the absorbing material, captured images inherently contain only reflection components without transmission interference. Unlike synthesis methods that rely on software-side simulation of optical effects from independent photographs [33], our captured pure reflections natively preserve complex physical phenomena—such as Fresnel attenuation, true illuminant colors, and depth-dependent defocus—as they are already optically encoded by the physical camera sensor.

Frame pair sampling. For each TartanAir sequence, we sample frame pairs with controlled separation following a Gaussian distribution: $d \sim \mathcal{N}(5.0, 2.5^2)$ clipped to $[0, 10]$ frames. This ensures a sufficient baseline for parallax while maintaining scene overlap, which is essential for correspondence.

Reflection-transmission mixing. We use Tartanair frames as transmission and the collected pure reflection images as the mixture to synthesize training pairs. While recent rigorous physics-based pipelines [33] reconstruct linear scene-referred irradiance using RAW DNG formats and un-white-balanced color spaces, large-scale video datasets with dense poses (e.g., TartanAir) are predominantly available only in 8-bit sRGB formats. To approximate the additive property of light—which dictates that photons linearly accumulate on the sensor—we perform mixing in an approximate linear radiance space:

$$M_i = \alpha_t \cdot T_i^\gamma + \alpha_r \cdot \mathcal{B}_\sigma(R_i^\gamma) \odot G \tag{A1}$$

where $\gamma \sim \mathcal{U}(1.6, 2.4)$ serves as a gamma expansion to reverse the display-referred compression before the linear addition of transmission and reflection. After superposition, the mixture is recompressed to simulate the final image. Although this gamma-based linearization is an approximation compared to the exact RAW-based radiometric calibration in [33], it provides a physically grounded compromise that enables the synthesis of dual-view motion correlations. $\mathcal{B}_\sigma$ represents Gaussian blur with $\sigma \sim \mathcal{U}(0, 3)$ pixels modeling defocus blur, G is a ghosting mask simulating partial reflections, and α_t, α_r are transmission and reflection mixing weights sampled from stratified distributions to ensure diverse intensity levels.

The reflection component exhibits view-dependent motion correlated with camera movement:

$$R_2 = \mathcal{W}(R_1, H_{12} + \xi) \tag{A2}$$

where $\mathcal{W}$ is a warping operator, H_{12} represents the homography induced by camera motion between frames, and $\xi \sim \mathcal{N}(0, \sigma_\xi^2 I)$ adds a stochastic perturbation to simulate natural handheld capture variations.

To enhance realism, we implement three key augmentations. First, we employ stratified intensity sampling, where reflection intensities are sampled from four categories: weak (25%), medium (40%), strong (25%), and overexposed (10%) to ensure balanced representation of real-world scenarios. Second, we apply correlated transformations including rotation ($\pm 15°$) and scaling (0.9-1.15×) with

Table A1: Comprehensive survey of reflection removal datasets. We systematically categorize existing datasets by their capture type (Single/Dual/Video), data source (Synthetic/Real), and availability. The table shows the evolution of dataset scale and quality over the years, with our DualRef dataset (marked with †) being the first real-captured reflection video dataset with ground truth. Checkmark (✓) indicates publicly available datasets, while cross (×) indicates datasets that are not publicly accessible.

Dataset	Year	Type	Usage	# Images	Avg. Resolution	Source	Available
Single-View Datasets							
CEIL [12]	2017	Single	Train/Test	7,643/850	224×224	Synthetic	✓
RID [89]	2017	Single	Train	3,250	224×320	Synthetic	✓
SIR^2 [62]	2017	Single	Test	454	540×400	Real	✓
Real [89]	2018	Single	Train/Test	89/20	1152×930	Real	✓
Nature [39]	2020	Single	Train/Test	200/20	598×398	Real	✓
CDR [36]	2021	Single	Train/Test	1,063	1542×1638	Real	×
SIR^2+ [65]	2022	Single	Train	1,700	540×400	Real	✓
CID [66]	2023	Single	Train/Test	4,800/1,200	224×288	Synthetic	×
RRW [92]	2024	Single	Train	14,952	2580×1460	Real	✓
DRR [27]	2025	Single	Train/Test	23,303/400	3840×2160	Real	×
OpenRR-1K [83]	2025	Single	Train/Test	800/100	1024×768	Real	✓
CRLAV [13]	2025	Single	Test	600	512×385	Real	✓
Multi-View Datasets							
Niklaus [46]	2020	Dual	Train/Test	1,000/∼70	N/A	Synth./Real	×
Burst [9]	2024	Video	Test	11	1536×2040	Real	✓
DualRef-R (Ours)†	**2025**	**Dual**	**Train/Test**	**13181/2819**	**768×768**	**Synthetic**	✓*
DualRef-V (Ours)†	**2025**	**Video**	**Test**	**195**	**1200×650**	**Real**	✓*

*Available upon paper acceptance.

30% probability, loosely following camera motion with correlation factor 0.5-0.6. Third, 35% of samples include realistic partial occlusions with varying patterns (vertical edges: 40%, horizontal edges: 30%, rectangular: 20%, irregular: 10%) covering 10-40% of the image area, simulating common real-world obstruction scenarios.

Physics-based rendering in Blender. While our video-based split leverages optical capture and linear approximation to maximize motion diversity, we construct a strictly physically accurate rendering pipeline in Blender's Cycles engine that adheres to exact optical transport (ray tracing with precise IOR, Fresnel equations, and linear emission) to generate 4,000 dual-view pairs with precise ground truth decomposition. Our setup models real-world glass reflection scenarios with careful attention to optical properties.

We carefully configure scene components to achieve realistic glass-reflection effects. Following prior works [8,40,41], we utilize 600 outdoor HDR environment maps from Poly Haven [49] that provide diverse natural illumination conditions as world lighting for background scenes. The reflection sources consist of 200 indoor HDR images mapped onto large emissive planes positioned behind the glass, with emission strength randomly sampled from [3.0, 5.0] to create varying reflection intensities. The glass surfaces are modeled as thin planes (0.01m thickness using Solidify modifier) with Glass BSDF material, IOR= 1.43, and optional roughness $\in [0, 0.05]$ for frosted glass effects.

For dual viewpoint generation, we simulate natural handheld capture by setting baseline distances of 2-4cm between camera positions. The second camera incorporates random perturbations, including rotation ($\pm 5^\circ$ for pitch/yaw, $\pm 10^\circ$ for roll) and translation (up to 30cm displacement). Camera poses are stored as 7D vectors comprising 3D position and 4D quaternion rotation for accurate geometric representation.

For each scene configuration, we render three outputs: (1) mixed image with glass visible, (2) pure transmission with glass hidden, and (3) pure reflection with background HDR disabled. This triple-output approach enables precise decomposition for supervised training while maintaining physical accuracy in the synthesized data.

Our current synthetic setup models a single glass pane. It does not explicitly simulate dual-pane windows with ghosting offsets [57] and incorporating such configurations is a promising direction for extending DualRef-R.

A.3 Details for real capture method.

Capturing high-quality, perfectly aligned video sequences, both with and without reflections, in the real world is extremely challenging. To construct our DualRef-V dataset, we utilized a programmable, motorized camera dolly to ensure trajectory reproducibility. Beyond spatial alignment, we implemented strict protocols to address hardware variations and photometric consistency. Below, we provide detailed answers to hardware-specific and camera-setting concerns regarding our capture process.

Photometric and focal consistency between passes. When capturing paired ground-truth data (mixture image vs. pure transmission image), any variation in camera settings—such as exposure time, ISO, or focus distance—will introduce unwanted photometric or blur discrepancies, compromising the evaluation fairness. To prevent this, we developed a custom smartphone application that uses a camera API to strictly lock all capture parameters.

Specifically, before recording a sequence, we perform an initial metering and focusing pass without the glass panel. During this step, the camera calculates the optimal shutter speed and ISO for the background scene and auto-focuses on the center of the transmission layer. Once these values are determined, we manually lock the shutter speed, ISO, and focal distance. The camera then traverses the programmed trajectory twice: once with the glass panel (to capture the mixture video) and once without it (to capture the transmission ground truth). Because all settings are strictly locked, the exposure remains constant throughout the entire trajectory and identically matches between the two passes. Furthermore, locking the focal distance to the transmission layer ensures that during the glass-present pass, the camera focuses exactly on the background rather than being drawn to the reflection on the glass surface. This mimics the actual user intent in reflection-removal scenarios and ensures pixel-level consistency between the input and the ground truth.



Fig. A1: Robustness to heterogeneous mobile lenses. We test a real wide/ultra-wide pair from a Huawei Mate 40 after simple overlap cropping and color transfer. Despite differences in focal length, field of view, and ISP response, our model removes reflections without fine-tuning, supporting practical deployment on consumer dual-camera hardware.

A.4 Robustness to Heterogeneous Mobile Lenses

Although our quantitative DualRef-V evaluation uses a single moving camera to obtain precisely aligned ground truth, practical mobile deployment often uses heterogeneous camera pairs, such as a wide and ultra-wide lens, whose intrinsic parameters, fields of view, and image signal processing pipelines differ. We therefore conduct an additional real-device test on a Huawei Mate 40 using its wide/ultra-wide pair (approximately 27/17mm-equivalent focal lengths). Before inference, we perform lightweight preprocessing: center-cropping the overlapping field of view and applying CIE-Lab Reinhard/histogram color transfer to reduce exposure and color response differences. No model fine-tuning is performed.

As shown in Figure A1, our method remains effective under true heterogeneous mobile-lens captures. Furthermore, flagship phones share similar lens configurations, enabling practical cross-device generalization.

A.5 Robustness to IMU and Camera Pose Noise

To maintain a consistent and reliable benchmark for quantitative evaluation, all 195 sequences in the DualRef-V dataset were captured using the same Huawei Mate 40 smartphone. However, we acknowledge that in practical applications, IMU and gyroscope sensors can exhibit slight noise profiles, drift characteristics, and calibration differences across various device models.

To rigorously evaluate the generalization ability and robustness of our model to such hardware discrepancies, we conduct a controlled stress test on 500 sampled pairs from our DualRef-V dataset (same data used in our main evaluations in Figure 5) and the Chugunov dataset [9]. Specifically, we inject varying levels of zero-mean Gaussian noise $\mathcal{N}(0, \sigma^2)$ into the relative camera poses (both the unit translation directions and rotation quaternions). Mathematically, injecting a noise level of $\sigma = 0.05$ to unit vectors and quaternions equates to severe spatial deviations of approximately 3° to 6°. Given that the temporal interval between dual-view captures is extremely short (typically < 0.5s), modern smartphone IMUs exhibit negligible short-term rotational drift well below 1°. Therefore, our high-noise setting ($\sigma = 0.05$) serves as a highly conservative, worst-case bound

Table A2: Quantitative robustness to camera pose perturbations. We report the performance under different levels of injected pose noise (σ). Niklaus [46] does not utilize pose data, thus its performance remains constant but sub-optimal. Explicit 3D methods like Chugunov [9] suffer severe degradation under noise, whereas our soft-conditioning approach maintains high robustness. (Note: Bold indicates the best performance).

Method	Noise Level	DualRef-V (500 pairs)					Chugunov Dataset [9]				
		PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Niklaus [46]	All levels	18.36	0.7062	0.2351	0.3759	59.33	22.14	0.7812	0.1654	0.4210	52.18
Chugunov [9]	Clean ($\sigma = 0$)	23.89	0.7495	0.1481	0.4502	63.19	26.54	0.8356	0.1155	0.4858	58.62
	Low ($\sigma = 0.02$)	21.56	0.7109	0.2054	0.4109	60.22	23.47	0.7985	0.1427	0.4404	54.33
	High ($\sigma = 0.05$)	17.29	0.6655	0.2581	0.3550	56.48	20.13	0.7324	0.1988	0.3857	49.13
Ours	Clean ($\sigma = 0$)	**24.17**	**0.7591**	**0.1638**	**0.4705**	**64.81**	**27.05**	**0.8412**	**0.1102**	**0.5012**	**59.85**
	Low ($\sigma = 0.02$)	23.92	0.7554	0.1651	0.4687	64.59	26.85	0.8382	0.1120	0.4959	59.38
	High ($\sigma = 0.05$)	23.61	0.7484	0.1682	0.4612	63.91	26.42	0.8310	0.1175	0.4821	58.46

that vastly exceeds typical real-world sensor inaccuracies. We compare our approach against the state-of-the-art multi-view flow-based method (Niklaus [46]) and the burst-based 3D optimization method (Chugunov [9]). We note that, for ours and Niklaus [46], we input two images, whereas for Chugunov [9] the input is 42 consecutive frames. The quantitative results are summarized in Table A2.

As demonstrated in Table A2, explicit 3D optimization methods, such as the burst-based baseline [9], suffer severe performance degradation under camera pose perturbations. These approaches rely on rigid epipolar geometry to warp and align frames; consequently, even sub-degree pose inaccuracies cause projected transmission layers to misalign, introducing structural artifacts. Conversely, the optical flow-based method (Niklaus *et al.* [46]) remains immune to IMU noise as it estimates correspondences purely from images. However, explicit flow estimation is inherently fragile under complex reflection interference (as discussed in Section 4), fundamentally limiting its baseline performance, even compared to our model, under high-noise conditions.

In contrast, our proposed dual-view diffusion framework exhibits remarkable resilience to pose perturbations, maintaining high quantitative scores even under severe noise ($\sigma = 0.05$). This robustness directly stems from two key architectural designs. First, rather than enforcing strict spatial warping matrices, we inject the relative pose as a soft conditional embedding ($\mathbf{e}_c = \mathbf{e}_t + \gamma \mathbf{e}_p$). This acts as a geometric prior that guides the frame-aware spatial attention (Section 4.1) to dynamically establish cross-view correspondences, effectively tolerating minor pose deviations. Second, the pose dropout strategy employed during our curriculum training (Section 4.3) forces the network to avoid over-reliance on perfect IMU data, enabling it to gracefully fall back on robust cross-view visual features when camera pose information is unreliable. Ultimately, this stress test confirms that our method is practical for deployment on diverse consumer mobile devices, where inherent sensor noise is inevitable.

A.6 Justification of Evaluation Pair Sampling

During the evaluation of our DualRef-V dataset, we construct testing pairs by sampling frame pairs with a relative pixel displacement forming a Gaussian distribution centered at 50 pixels ($\mathcal{N}(50, 40)$). This sampling strategy is not chosen arbitrarily but is meticulously formulated to mirror the exact physical optical properties of modern smartphone dual-camera systems and natural capture behaviors.

According to the standard stereo vision model, the pixel disparity d between two views is governed by $d = (B \cdot f_{px})/Z$, where B is the physical baseline distance between the cameras, f_{px} is the camera focal length expressed in pixels, and Z is the scene depth.

We parameterize this equation based on prevalent consumer hardware and realistic photography conditions:

1. **Hardware Baseline (B):** In modern flagship smartphones (e.g., the Apple iPhone 15 Pro or Huawei Mate series), the physical distance between the primary wide lens and the ultra-wide or telephoto lens typically ranges from $15mm$ to $25mm$ ($0.015m$ to $0.025m$).
2. **Focal Length (f_{px}):** Given our evaluation image width of $W = 1200$ and a standard wide-camera Field of View (FoV $\approx 75°$), the focal length in pixel coordinates is approximately $f_{px} \approx W/(2\tan(\text{FoV}/2)) \approx 800$ to 1000 pixels.
3. **Operating Depth (Z):** Reflection removal scenarios prominently involve shooting through glass surfaces at close-to-mid ranges, such as museum exhibits, storefronts, or vehicle windows. The target transmission layers in these environments predominantly fall within a depth range of $0.3m$ to $1.0m$.

Substituting these real-world values into the disparity equation yields expected pixel shifts ranging from 15 to 80 pixels. Specifically, capturing an object at a typical distance of $0.5m$ using a $25mm$ camera baseline and a 1000-pixel focal length yields a theoretical disparity of exactly 50 pixels.

Furthermore, when considering the alternative user-friendly scenario of utilizing Live Photos [2] or burst captures from a single lens, the natural hand tremor and body sway over a brief 0.5-second interval introduce an equivalent effective baseline of $1 \sim 3cm$, converging to an identical disparity distribution. Therefore, our sampling distribution rigorously simulates the intrinsic parallax of dual-lens hardware and the dynamic motion of hand-held capture, ensuring our quantitative evaluation accurately reflects real-world performance.

B Additional results

B.1 Comparison with powerful image editing models

With the rapid development of general-purpose image editing models, some may argue that the study of reflection removal, as a branch of low-level image restoration, has lost its significance. Therefore, we experiment to compare our results with those of powerful image editing models, including OpenAI GPT 5.2 [47] and



Fig. A2: Visual results for qualitative comparison with powerful image editing models on challenging real-world scenarios in our DualRef-V dataset. For each scene divided by dashed lines, the upper row and the bottom row displays view 1 and 2, respectively.

Gemini Nano Banana 2 [15]. The experiments are carried out on our real-world reflection dataset DualRef-V.

For both models, we give the prompt *"Remove the reflection in the image, keep the background intact, keep the original input resolution."*. However, GPT 5.2 still outputs a different resolution. We also tried giving both views together and the prompt *"Remove the reflection for two views of the same scene leveraging the view parallax information."* yet both models always output a single image and show little improvement over results given a single image. Qualitative results in Figure A2 show that these general-purpose image editing models (*i.e.* GPT 5.2 [47] and Gemini Nano Banana 2 [15]) tend to modify the color and texture or the content of the input image, resulting in unsatisfactory background fidelity. Additionally, they fail when confronted with complex or high-intensity reflections. This suggests that current state-of-the-art image editing models or image restoration models, which are not specifically optimized for reflection removal, may still be inadequate for removing reflections. Therefore, the study of effective reflection removal methods remains necessary and important.

We also evaluate the performance of Adobe Photoshop 2026's reflection removal model, which is associated with Kee *et al.* [33] on DualRef-V. As is shown in Figure A2, it fails to remove intense or structurally complex reflections. This is probably because Kee *et al.* [33] is trained on RAW images, where reflections and transmissions are roughly linearly blended, and reflection removal in linear space is fundamentally a different problem. Given that we are consid-

Table A3: Quantitative comparison on single-image benchmarks (Part 1/3). We evaluate our method in Dual separate mode against best SIRR methods [26, 27, 30, 87, 90]. Best results are in **bold**, second-best are underlined. ↑ or ↓: higher or lower is better.

Method	SIR^2 [65] (454 images)						Real [89] (109 images)					
	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑
DSIT [30]	27.06	0.9333	0.0690	0.3706	0.3908	41.25	27.29	**0.8490**	0.1047	0.4922	0.3887	44.64
L-DiffER [90]	25.18	0.9110	0.1210	-	-	-	23.77	0.8210	0.1530	-	-	-
RDNet [90]	27.00	0.9285	0.0737	0.3501	0.3869	40.78	26.52	0.8405	**0.0996**	0.5046	0.3901	44.20
DAI [27]	26.75	0.9333	0.0848	0.3377	0.3859	39.50	27.57	0.8455	0.1390	0.4919	0.3966	45.07
WSeat [87]	28.87	**0.9405**	0.0628	0.4592	0.4025	43.78	**27.93**	0.8386	0.1965	0.5239	0.4061	47.48
Ours	**29.15**	0.9349	**0.0477**	**0.4822**	**0.4128**	**46.42**	27.23	0.8482	0.1037	**0.5803**	**0.4143**	**48.32**

ering mobile phone photography as our main setting, we limit the discussion to tone-mapped inputs.

B.2 Evaluation on single-image datasets

While our method is designed for dual-view inputs, it can operate in single-image mode by setting both input views to the same image with a zero relative pose (which corresponds to the Dual separate model in Sec 5.3). This capability is enabled by our training strategy, which randomly drops the second view with a 30% probability, as described in Sec. 4.3 of the main paper. Although our ablation studies demonstrate significant performance gains from dual-view information (see Table 2 in the main paper, comparing "Dual separate" with "Dual w/o pose"), we conduct single-image evaluations to ensure a fair comparison with existing methods and demonstrate the versatility of our model. We employ six complementary metrics: SSIM and PSNR for structural similarity and signal quality; LPIPS [88] for perceptual distance; CLIPIQA [67], MANIQA [84] and MUSIQ [32] for no-reference image quality assessment. These metrics collectively evaluate pixel accuracy and perceptual quality. We compare with state-of-the-art SIRR methods [26,27,30,87,90] on three real-captured SIRR datasets [13,65,89]. Some statistics for L-DiffER [26] are unavailable because it is not open-sourced.

As shown in Table A3, Table A4 and Table A5, our method achieves competitive performance across all single-image benchmarks despite being primarily designed for dual-view inputs. We note again that **the results in these tables are acquired under the condition that neither pose data nor available second views are provided**, different from the dual-view setting which we explore.

B.3 Ablation Study on In-the-Wild Dataset

As a direct extension of the ablation study presented in Section 5.3 of the main manuscript, we further evaluate the effectiveness of our proposed frame-aware spatial attention mechanism on the in-the-wild Niklaus dataset [46]. Because this dataset lacks synchronized camera pose annotations, we perform inference using

Table A4: Quantitative comparison on single-image benchmarks (Part 2/3). Results on Nature [39] and CRLAV [13] datasets.

Method	Nature [39] (20 images)						CRLAV [13] (600 images)					
	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑
DSIT [30]	26.54	0.8467	0.1205	0.4195	0.3685	42.66	20.01	0.7146	0.2046	0.4277	0.3474	44.13
L-DiffER [90]	23.95	0.8310	0.1450	-	-	-	-	-	-	-	-	-
RDNet [90]	26.13	0.8457	**0.1024**	0.4350	0.3694	42.14	20.18	0.7305	0.1804	0.4464	0.3467	42.96
DAI [27]	26.89	0.8492	0.1840	0.3695	0.3687	41.53	20.90	0.7333	0.2139	0.3552	0.3400	42.43
WSeat [87]	27.58	**0.8555**	0.1352	0.5639	0.3917	46.32	21.99	0.7517	0.1484	0.4841	0.3783	46.45
Ours	**27.66**	0.8341	0.1109	**0.5806**	**0.3989**	**49.74**	**25.28**	**0.7568**	**0.1089**	**0.6209**	**0.3870**	**48.88**

Table A5: Quantitative comparison on single-image benchmarks (Part 3/3). Results on the OpenRR [83] dataset and weighted average across all 1,283 test images.

Method	OpenRR [83] (100 images)						Weighted Average (1,283 images)					
	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	MANIQA↑	MUSIQ↑
DSIT [30]	26.20	0.9307	0.0847	0.5667	0.3652	46.37	23.71	0.8223	0.1375	0.4237	0.3680	43.31
L-DiffER [90]	-	-	-	-	-	-	-	-	-	-	-	-
RDNet [90]	25.78	0.9308	0.0788	0.5654	0.3732	46.08	23.66	0.8273	0.1266	0.4264	0.3670	42.52
DAI [27]	26.72	0.9332	**0.0714**	0.5670	0.3712	47.65	24.08	0.8310	0.1503	0.3774	0.3639	42.01
WSeat [87]	**26.44**	**0.8482**	0.1627	0.5602	0.3769	46.34	25.36	**0.8350**	0.1231	0.4858	0.3893	45.58
Ours	27.12	0.8593	0.1025	**0.5707**	**0.3873**	**47.68**	**27.00**	0.8323	**0.0863**	**0.5638**	**0.3987**	**47.88**

our "Dual w/o pose" configuration. In this setting, the pose embedding is nullified, compelling the network to rely exclusively on visual cross-view correspondences established by the attention mechanism. The quantitative comparisons are detailed in Table A6.

The quantitative results consistently align with our findings on the DualRef benchmark. The baseline Single-view and Dual separate configurations exhibit sub-optimal performance, underscoring the necessity of genuine cross-view feature interaction for resolving complex real-world reflections. The SD2.1+RAFT baseline provides only marginal gains over the single-view and dual-separate baselines, showing that explicit flow alignment remains fragile under reflection-induced correspondence ambiguity even with the same SD2.1 prior. When incorporating cross-view information, alternative attention designs fall short. The Cat3D-style [14] mechanism introduces unconstrained bidirectional flow, resulting in cross-view feature ambiguity and noise leakage. More severely, the SIR-style [45] approach—which incorporates 3D ResNet blocks—disrupts the delicate pre-trained 2D generative priors, leading to a drastic degradation in perceptual fidelity (evidenced by the significant drop in LPIPS and MUSIQ scores). Width-concat also underperforms our channel-wise directed attention, indicating that such concatenation method seems not suitable for pixel-aligned guidance. In contrast, our directed frame-aware attention mechanism achieves the highest performance across all metrics. This strongly confirms that even without explicit geometric pose guidance, our directed attention effectively prevents noise propagation and robustly disentangles unstructured parallax motion in diverse, in-the-wild photographic scenarios.

Table A6: Ablation study on the Niklaus dataset [46]. Since camera poses are unavailable for this in-the-wild dataset, all cross-view configurations are evaluated without pose embeddings. Our directed frame-aware attention significantly outperforms alternative multi-view attention designs even in the absence of explicit geometric guidance.

Configuration	Niklaus Dataset [46]				
	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Single-view	22.84	0.7635	0.1240	0.5512	53.45
Dual separate	22.79	0.7612	0.1265	0.5483	53.27
SD2.1+RAFT	22.89	0.7641	0.1281	0.5530	54.08
Cat3D-Style Attn	23.45	0.7750	0.1182	0.5622	54.15
SIR-Style Attn	21.88	0.7386	0.1530	0.4854	49.59
Width-concat	22.79	0.7641	0.1328	0.5350	54.01
Ours (Dual w/o pose)	**25.09**	**0.7973**	**0.0865**	**0.5937**	**57.27**

B.4 Zero-Shot Generalization on Dynamic Scenes

In the main paper, our primary assumption is that modern smartphone dual-camera systems naturally capture two viewpoints simultaneously with a relatively small, fixed baseline. However, in everyday mobile photography, users may also capture image pairs sequentially by slightly moving the device between shots. This sequential capture inevitably introduces temporal gaps, resulting in varying viewpoint baselines and dynamic scene elements (e.g., moving people, vehicles, or moving reflection sources).

To demonstrate the versatility and robustness of our framework, we conduct a zero-shot evaluation on dynamic scenes—a highly challenging setting that is completely unseen during our training phase. For this evaluation, we collect testing pairs comprising real-world sequential captures made by ourselves and internet-sourced dynamic image pairs, each accompanied by camera pose information.

As illustrated in Figure A3, our dual-view diffusion framework gracefully handles these dynamic scenarios, successfully recovering the clean transmission layers without introducing ghosting artifacts or severe structural distortions. We attribute this strong zero-shot generalization capability to two fundamental reasons:

– **Exploitation of Residual Parallax:** Even in dynamic environments, the majority of the scene (e.g., architectural structures, distant backgrounds, and static objects) typically remains stationary. Consequently, the depth-induced parallax between the physical transmission background and the virtual reflection layer is preserved across the two views. Our implicit frame-aware spatial attention (Section 4.1), guided by the soft camera pose embedding, is flexible enough to extract and utilize this consistent global geometric cue, even when localized dynamic foregrounds exhibit inconsistent motion.



Fig. A3: Zero-shot reflection removal results on dynamic scenes. The input pairs are captured sequentially, introducing temporal gaps, dynamic objects, and larger camera baselines. Despite these challenging conditions completely unseen during training, our framework effectively removes reflections. It leverages the residual parallax from the static background and gracefully falls back to strong single-image generative priors for dynamic or occluded regions, avoiding the catastrophic failures common in explicit multi-view matching methods.

– **Robust Fallback to Single-Image Prior:** In localized regions where dynamic motion completely breaks cross-view correspondence (or when the baseline is exceptionally large), explicit multi-view methods relying on strict optical flow or 3D warping typically fail catastrophically. In contrast, our model relies on a powerful diffusion prior. Thanks to the carefully designed curriculum training strategy (Section 4.3)—which incorporates stochastic view replacement and pose dropout—our network has learned a robust fallback mechanism. In the worst-case scenario where cross-view information is contradictory or unavailable due to object motion, the model seamlessly degrades to a state-of-the-art single-image reflection-removal model, leveraging learned generative priors to restore the transmission layer independently.

B.5 Failure cases

Despite the strong performance of our method, specific challenging scenarios remain problematic. When viewpoint changes provide minimal parallax (e.g., pure rotation without translation, as is shown in the upper rows in Fig A4), the complementary information between views becomes insufficient for robust reflection separation. In such cases, our method degrades to essentially single-image performance. Moreover, in the specific case, the reflection on a washing machine is out-of-distribution data, leading to the complete failure of all compared state-of-the-art reflection removal methods [26, 27, 30, 87, 90].

We also observe color drift in some challenging examples, as shown in Fig. A4. This is a known limitation of diffusion-based restoration methods: the generative prior may slightly shift the global color or contrast even when the target content is mostly preserved, and there is no obvious general solution without sacrificing the model's ability to remove severe reflections. Our two-view design alleviates this issue to some extent because cross-view observations provide additional constraints on scene appearance, but it does not completely eliminate the problem.

B.6 Analysis of Pose Integration Strategies

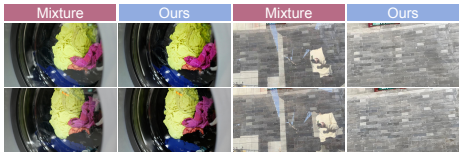


Fig. A4: Failure case of our method. The upper row and the bottom row displays view 1 and 2, respectively.

To address concerns regarding the optimality of our camera pose integration strategy, we provide a comprehensive analysis evaluating both the geometric representation of the pose and its injection mechanism into the diffusion framework. Quantitative comparisons across both the DualRef-V and DualRef-R datasets are summarized in Table A7. We note that there are no available dual-view datasets with pose data provided other than our proposed dataset. The Niklaus [46] dataset lacks pose data.

Table A7: Quantitative comparison of different pose representation and injection strategies. All models are evaluated on the DualRef-V and DualRef-R test sets. Our proposed strategy (Relative Pose via Time Embedding pathway) achieves the best performance by providing invariant geometric guidance without spatial or semantic domain gaps.

Configuration	Representation	DualRef-V			DualRef-R		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Dual w/o pose (Baseline)	N/A	23.53	0.7468	0.1694	27.47	0.8310	0.1176
Direct Concatenation	Relative	23.68	0.7482	0.1681	27.55	0.8321	0.1152
Textual Prompt Encoding	Relative	23.59	0.7471	0.1690	27.49	0.8313	0.1168
Sin/Cos Positional Encoding	Absolute	23.82	0.7510	0.1670	27.76	0.8345	0.1110
Plücker Coordinates	Absolute	23.95	0.7535	0.1652	27.91	0.8358	0.1065
Ours (MLP to Time Emb)	**Relative**	**24.17**	**0.7591**	**0.1638**	**28.19**	**0.8378**	**0.0983**

Pose Representation Analysis. A critical design choice is the geometric representation of the camera pose. In multi-view synthesis and 3D-aware generation, absolute camera parameters, such as high-frequency Sin/Cos positional encodings or Plücker coordinates, are widely adopted. While these representations excel at novel view synthesis, where absolute 3D spatial structure is paramount, they are suboptimal for the dual-view reflection removal task. The disentanglement of transmission and reflection relies entirely on depth-induced parallax, which is strictly a function of the relative baseline between the two viewpoints. Consequently, the physical correlation between the two frames remains identical regardless of the absolute camera orientation or global translation in the world coordinate system.

As demonstrated in Table A7, although incorporating absolute pose information via Plücker coordinates or Sin/Cos encodings yields improvements over the baseline without pose guidance, these methods underperform compared to our relative representation. This performance gap arises because absolute coordi-

nates inadvertently introduce coordinate bias, forcing the model to overfit to the absolute world-space trajectories observed during training rather than learning the generalized relative transformation. By explicitly providing only the normalized relative translation direction and relative rotation (quaternion), our method inherently enforces SE(3) equivariance. This perfectly isolates the essential geometric cue—the viewpoint shift—while completely shielding the network from irrelevant global coordinate interference.

Pose Injection Mechanism Analysis. Beyond the pose representation, the mechanism for injecting this geometric information into the U-Net architecture significantly impacts performance. We explored several alternative strategies to validate our design. First, we evaluated direct concatenation, where the pose vector is spatially replicated and concatenated with the image latents along the channel dimension. This approach yielded minimal performance gains, primarily because of a substantial domain gap: spatial convolutional layers are optimized to process localized, high-frequency visual features rather than globally uniform numerical constants. Second, we investigated textual prompt encoding by converting numerical pose values into text descriptions and injecting them via the CLIP text encoder and cross-attention layers. This strategy yielded the lowest performance among the conditioned variants. Pre-trained text encoders are designed to map high-level semantic concepts but lack the granularity to accurately process high-precision floating-point geometric coordinates, thereby introducing unnecessary computational overhead.

In contrast, our proposed method uses a lightweight MLP to map the relative pose to a geometric embedding, which is linearly combined with the timestep embedding. Within the diffusion U-Net, this combined time-pose embedding naturally acts as a global modulator via standard Feature-wise Linear Modulation (FiLM) layers across all network scales. This design elegantly biases the global denoising trajectory based on the camera baseline, without polluting spatial feature maps or introducing semantic domain gaps, achieving the best quantitative and visual results with negligible computational cost.

In summary, our strategy of using a relative pose representation and injecting it via the temporal modulation pathway is grounded in the physical properties of dual-view parallax and the architectural characteristics of diffusion models, proving optimal and highly efficient for this task.

C Implementation details

C.1 Training details

Our model training follows a two-stage curriculum learning strategy to ensure stable convergence and optimal performance. In the foundation training stage, we first fine-tune the pre-trained InstructPix2Pix model (built on Stable Diffusion 2.1) on our DualRef-R dataset for 30,000 steps using simpler samples with clean backgrounds and moderate reflection intensities. This stage employs a batch size of 32, a learning rate of 1e-5 with a constant schedule, a resolution

of 768×768, and disables mixed precision for stability. For the advanced training stage, we continue for an additional 20,000 steps, incorporating more challenging samples with complex backgrounds, severe reflections, and diverse lighting conditions. During this stage, we enable pose conditioning with 25% dropout probability to maintain single-image capability and a pose scale factor of 0.15 to control the influence of pose embeddings. The curriculum learning approach ensures the model first learns fundamental reflection-transmission decomposition before tackling more complex scenarios. Training was conducted on an RTX 6000 PRO BLACKWELL GPU with 96GB of VRAM, utilizing gradient checkpointing and xFormers' memory-efficient attention to optimize memory usage while maintaining full 768×768 resolution throughout training.

C.2 Inference

During inference, we employ 50 denoising steps. For input images of arbitrary resolution, we resize them to 768×768, perform inference at this resolution, and then resize the output back to the original dimensions. On an RTX 6000 PRO BLACKWELL GPU, the inference process completes in 2 seconds for 2 views, producing high-quality, reflection-free outputs. This inference speed is practical for real-world applications while maintaining the quality advantages of diffusion-based generation.

Text conditioning For computational efficiency, we remove text conditioning entirely from our pipeline. The text encoder's cross-attention layers receive dummy zero inputs during both training and inference, resulting in approximately a $2\times$ speedup.