RefMover: Diffusion-based Single Image Reflection Removal with Language and Region Guidance

Zifeng Wang†, Yuchen Hong†, Tianyi Xu, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, *Member, IEEE*, Boxin Shi‡, *Senior Member, IEEE*

***Abstract*—Single image reflection removal (SIRR) is a highly ill-posed task. While recent research explores RAW data for physical accuracy, the dominance of processed (8-bit) images in real-world scenarios necessitates robust solutions capable of handling non-linear degradations. Prevailing methods, however, can struggle on such data due to the lack of high-level semantic guidance. Although language-guided diffusion models show promise in bridging this gap, they face challenges in semantic-spatial alignment and recovery faithfulness due to contaminated input conditions. To address these challenges, this paper introduces *RefMover*, a multi-instruction diffusion framework for SIRR. It bridges the semantic-spatial gap by combining language and region guidance. To improve faithful and stable recovery, *RefMover* features a pipeline with a streamlined conditional architecture that mitigates condition conflicts and preserves details. We also introduce a new large-scale dataset (COMRR) for robust training and evaluation. The experimental results demonstrate the effectiveness of the proposed method and its potential for practical interactive and cloud-assisted applications.**

***Index Terms*—Reflection removal, diffusion model, image restoration.**

I. INTRODUCTION

IMAGES captured through reflective surfaces are often degraded by artifacts that impair downstream computer vision tasks [1]–[3]. Single image reflection removal (SIRR) aims to recover the clean transmission scene (denoted as $\mathbf{T}$) from a mixture image (denoted as $\mathbf{M}$) containing an unwanted reflection (denoted as $\mathbf{R}$) [4], [5]. It is an ill-posed problem. Unlike tasks such as deraining [6], [7] or desnowing [8], [9] where the target for removal has relatively distinct patterns, reflections can be as complex as the background itself.

Z. Wang, Y. Hong, T. Xu, and B. Shi are with State Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University. Beijing 100871, China. Email: wzf1835851@stu.pku.edu.cn; yuchenhong.cn@gmail.com; photon@stu.pku.edu.cn; shiboxin@pku.edu.cn.

H. Zhong is with the Wangxuan Institute of Computer Technology, Peking University. Beijing 100871, China. Email: hfzhong@pku.edu.cn

S. Weng is with Beijing Academy of Artificial Intelligence, Beijing 100083, China. Email: scweng@baai.ac.cn.

J. Liang is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Chiyoda-ku 1018430, Japan. Email: cssherryliang@gmail.com.

This work was supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and Beijing Natural Science Foundation (Grant No. QY24037).

† Equal contribution. ‡ Corresponding author.

To mitigate the information loss inherent in standard image signal processing (ISP), recent research [10] proposes shifting the domain to RAW images, capitalizing on the linear physical additivity of light. However, access to RAW data is often restricted in casual photography and web applications, making the restoration of standard 8-bit images—despite their non-linear distortions—a persisting and widely relevant challenge. Consequently, prevailing deep learning-based methods [4], [11]–[13], which operate in a "blind" manner on processed images, often fail in complex scenes. Lacking high-level semantic guidance, they may incompletely remove reflections or damage background details, highlighting the need for an effective form of prior knowledge.

As visually demonstrated in Fig. 1(a) and (b), representative SIRR methods such as DSIT [14] and ZF24 [15] struggle to distinguish reflection and transmission layers, often generating blurry results with noticeable artifacts.

Natural language emerges as a prime candidate to provide this prior knowledge, offering rich semantic information via a convenient user interface. The success of large vision-language models [17] and the generative power of diffusion models [18], [19] have shown potential for language-guided visual tasks like image editing and colorization [20]–[24].

However, applying these technologies to reflection removal faces several challenges. The first is the **semantic-spatial gap**: text prompts like, "remove the fluorescent light," lack the spatial specificity for targeted removal. The second is the challenge of a **contaminated control condition**; unlike many generative tasks [25], [26] that use a clean guide image, SIRR's input is polluted by the very reflections to be removed, which hinders faithful recovery. A synthesis-to-reality gap in existing training data further compounds these issues, as prevailing generation strategies, such as linear addition [4] and regional subtraction [27], fail to faithfully replicate the complex characteristics and distribution of real-world reflections, limiting model generalization.

Our preliminary work, L-DiffER [16], first validated the feasibility of using a language-guided diffusion model for SIRR. However, its **dual-branch condition architecture** (using both the mixture image and its gradient map as parallel inputs) may introduce interactions between the image and gradient conditions. While visually outperforming early SIRR methods (as visually evidenced in Fig. 1(c)), this preliminary design occasionally encountered challenges in preserving the most fine-grained textures under extremely complex lighting

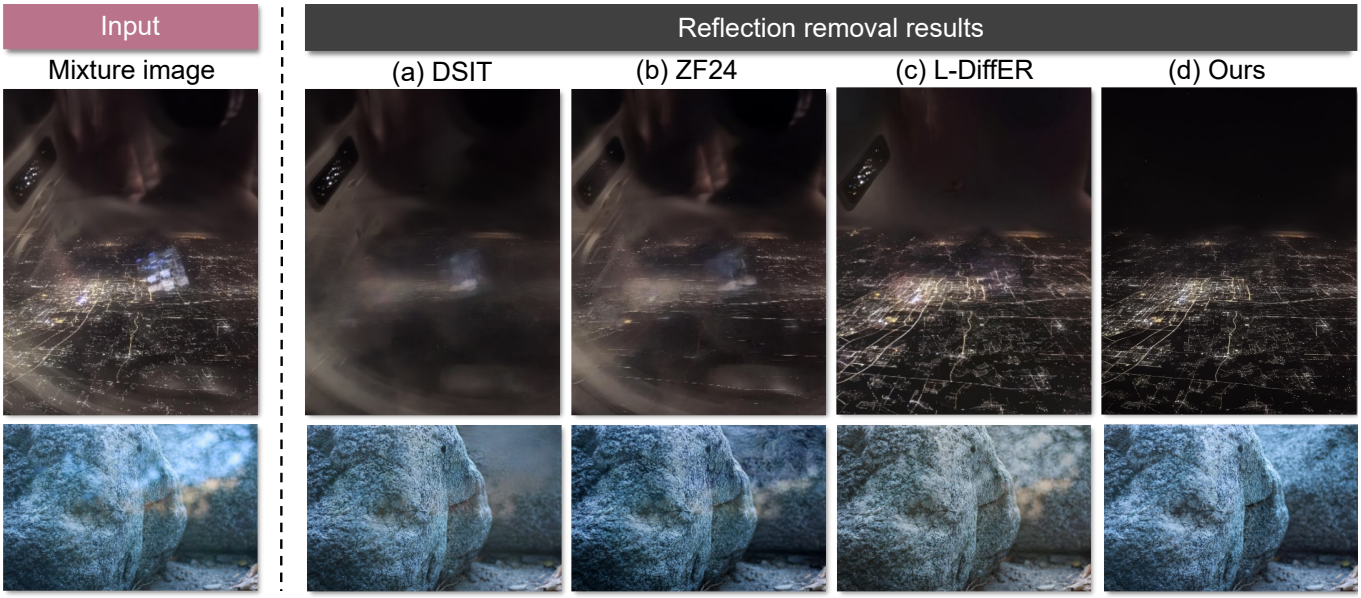


Fig. 1: Examples of reflection removal on challenging real-world images. **(a)** DSIT [14] and **(b)** ZF24 [15] leave visible residuals or blurry structures in these examples. **(c)** Our preliminary work L-DiffER [16] suffers from contaminated conditions and color shift, while **(d)** the proposed *RefMover* more effectively suppresses diverse and complex reflection artifacts to restore high-fidelity transmission layers.

scenarios or dense reflection contamination.

To advance beyond this preliminary exploration, this paper introduces *RefMover* with three main innovations. First, it replaces the dual-branch design with a **streamlined conditional architecture** to improve recovery stability and fidelity, scaling effectively to challenging in-the-wild cases. The direct visual comparison in Fig. 1(d) shows the proposed method successfully restoring high-fidelity details compared to the preliminary framework. Second, it proposes a **multi-instruction mechanism**, which fuses synergistic guidance from both semantic language and a spatial mask to effectively bridge the semantic-spatial gap. Third, to address the synthesis-to-reality gap, we introduce **ComRR**, a new dataset for reflection removal. It includes a collection of authentic, real-captured reflection layers for realistic training data synthesis, alongside two subsets for thorough evaluation, consisting of complex real-world images and a set of test cases with corresponding ground truth transmission images. We summarize our contributions as follows:

- We propose a diffusion-based reflection removal framework with a streamlined conditional architecture and refinement strategies to ensure high faithfulness despite the challenges of contaminated input conditions.
- We present the first multi-instruction diffusion model for reflection removal, which bridges the semantic-spatial gap by enabling precise user control via joint language and spatial guidance.
- We capture a large-scale pure reflection image dataset and a complex real reflection dataset to facilitate method training and evaluation, respectively.

In Sec. II, we start with an introduction to existing reflection removal methods and applications of language-based diffusion models. Sec. III introduces the network architecture and strategies applied in the training and inference stages. We present details of our data generation process and proposed datasets in Sec. IV. Quantitative and qualitative experiments and ablation studies are conducted in Sec. V. Finally, we conclude the paper in Sec. VI.

II. Related Work

A. Reflection removal

Reflection removal is a challenge in computational photography that aims to decompose a corrupted image into its constituent transmission and reflection layers. To alleviate the ill-posedness of this problem, multi-image methods leverage auxiliary physical cues from additional observations, such as polarization information [28]–[34], motion discrepancy between the layers [35]–[37], or transmission scene information from active illuminations [38]–[42]. While these approaches can achieve effective separation by providing additional physical constraints, their reliance on specialized capture setups limits their practical applicability. Consequently, SIRR remains a more general and challenging task. Early non-learning methods employ handcrafted priors into their optimization frameworks [43], [44], such as the gradient sparsity [45], relative smoothness [46], and ghosting cues [47]. Developing flexible neural architectures capable of adaptively processing complex visual inputs has significantly advanced computer vision, benefiting diverse tasks such as video summarization [48]. With the advent of deep learning, learning-based methods have become dominant, exploring various network architectures [4], [14], [49]–[52], loss functions [27], or the iterative recovery strategy [11], [12], [53], [54]. Concurrently, efforts in data synthesis and capture are also an active topic in this area [13], [15], [49], [55]–[59]. Notably, Kee *et al.* [10] recently challenged the physical validity of synthesizing reflections on

tone-mapped images due to non-linear gamma compression, proposing instead a simulation pipeline strictly based on linear RAW data. While this approach ensures physical fidelity, the requirement for RAW inputs limits its applicability in general scenarios where only processed (8-bit) images are available. Despite their progress, prevailing methods for standard images operate blindly, lacking high-level semantic priors to resolve ambiguities. They also cannot bridge the semantic-spatial gap for user-directed, region-specific removal. L-DiffER [16] first explored using language to address this semantic gap, but its dual-branch "multi-condition" architecture introduced instability and compromised recovery faithfulness. Moreover, it did not fully solve the semantic-spatial gap, lacking precise user-steered spatial control. These limitations highlight the challenge of applying generative models to SIRR, underscoring the need for a solution that ensures faithful recovery and is finely-tunable.

B. Diffusion model

Our work is built upon recent advances in Denoising Diffusion Models [18], which have achieved state-of-the-art (SOTA) performance in high-fidelity image generation. The synergy of Diffusion Models with vision-language models such as CLIP [17] has enabled intuitive, language-based control over a wide range of vision tasks. This has been instrumental in applications such as image colorization [24], user-guided image editing [25], [60], [61], and devising sophisticated sampling strategies for tailored generation outcomes [62]. However, applying this paradigm to reflection removal poses a challenge: unlike creative editing, it demands strict fidelity while being guided by a visual condition that is itself corrupted by the artifacts to be removed. L-DiffER [16] confirmed the feasibility of this direction but also highlighted limitations regarding framework stability and guidance precision. *RefMover* is designed to overcome these specific issues by introducing a robust recovery pipeline that ensures faithful restoration, coupled with a more precise, multi-instruction guidance mechanism that enables explicit user control.

III. Proposed Method

The design of *RefMover* is centered on overcoming two obstacles that arise when applying language-guided diffusion models to reflection removal. The first is the insufficient faithfulness of recovery that results from using the reflection-contaminated mixture image as a guiding condition, often leading to color shifts and structural distortions. The second is the inherent semantic-spatial gap, where high-level textual descriptions lack the pixel-level precision required for targeted manipulation. To address these challenges, we introduce a framework with two main components. To ensure faithful and stable recovery, we develop a pipeline featuring a streamlined conditional architecture and discrete refinement strategies. To bridge the semantic-spatial gap, we propose a multi-instruction mechanism for this task, which combines language and region-based spatial guidance to provide precise user control.

This section details our approach as follows: Sec. III-A outlines the problem formulation and our proposed method; Sec. III-B describes the innovations in our recovery pipeline; Sec. III-C details the multi-instruction conditioning mechanism; and Sec. III-D presents the loss function and implementation details.

A. Overview

Problem formulation. SIRR aims to recover the transmission layer $\mathbf{T}$ from a captured mixture image $\mathbf{M}$, which is corrupted by an unwanted reflection layer $\mathbf{R}$. Following widely adopted imaging models in the field [4], [11], [52], the mixture image is commonly approximated as:

$$\mathbf{M} \approx \mathbf{T} + \mathbf{R}, \tag{1}$$

where $\mathbf{M}, \mathbf{T}, \mathbf{R} \in \mathbb{R}^{H \times W \times 3}$, with H and W representing the height and width of the images. This additive relation is not assumed to hold exactly for processed 8-bit sRGB inputs.

In this work, we enhance the traditional reflection removal problem by incorporating user guidance. Specifically, our task is to recover the transmission layer $\mathbf{T}$ from the input mixture image $\mathbf{M}$, conditioned on a language description $\mathbf{c}$ extracted from the CLIP text encoder [17] and a user-provided region awareness mask $\mathbf{S} \in \mathbb{R}^{H \times W \times 1}$. The language condition $\mathbf{c}$ provides semantic information about the unwanted reflection artifacts present in $\mathbf{M}$, while the spatial mask $\mathbf{S}$ provides explicit spatial cues, indicating regions with significant reflection or areas the user wishes to focus on. In the corresponding ablated variants, text guidance is removed by using an empty text condition, while spatial guidance is removed by omitting the mask channel from the U-Net input and disabling mask-dependent blending in Differential Diffusion.

Method overview. *RefMover* redesigns the architecture of L-DiffER [16] to address its practical limitations. Architecturally, its dual-branch ControlNet [25] structure led to cross-modal feature competition and training instabilities. Functionally, its continuous condition refinement strategy resulted in detail over-smoothing and condition drift, while its reliance on a positive-negative prompt mechanism could introduce unintended content modifications. Furthermore, like many diffusion models [63], [64], it was prone to color shifts and brightness biases, compromising the fidelity of the final result.

To address these framework-level issues, *RefMover* introduces a series of improvements to enhance recovery faithfulness. We replace the dual-branch architecture with a streamlined, concatenation-based U-Net structure, which mitigates feature interference and improves efficiency. This new U-Net denoiser employs a v-prediction objective [65], with its core operation represented as:

$$\mathbf{v} = v_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{y}, \mathbf{S}'), \tag{2}$$

where $\mathbf{z}_t$ is the noisy input, $\mathbf{y}$ is the condition from the mixture image, $\mathbf{c}$ is the language condition, and $\mathbf{S}'$ is the downsampled spatial mask. To further enhance output quality, this framework is also coupled with two strategies: a discrete condition refinement strategy to prevent drift and preserve details, and fidelity preservation techniques, including zero-SNR rescaling and a residual-based refinement module for texture transfer and error correction.

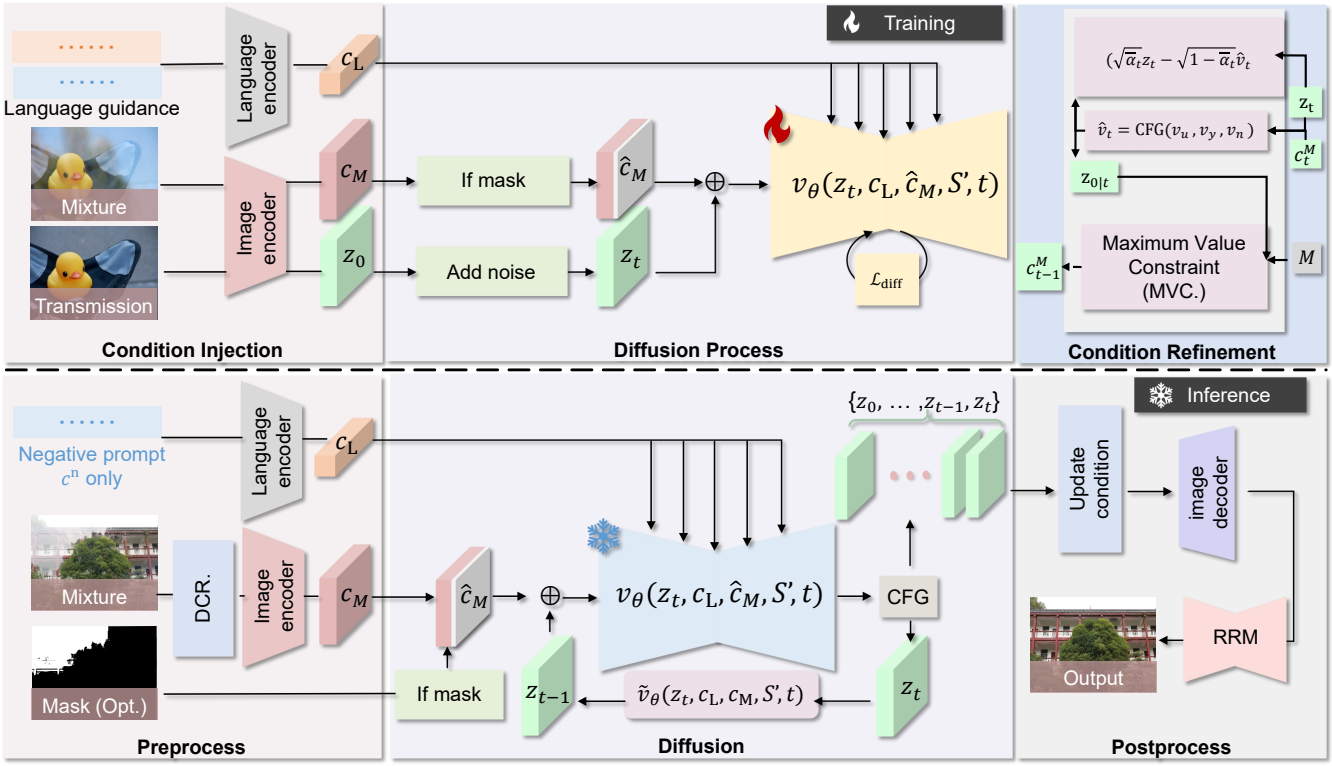


Fig. 2: Framework of the proposed *RefMover*, illustrating both training **(top)** and inference **(bottom)** pipelines with their respective condition injection and refinement mechanisms, which leverage language descriptions with the discrete condition refinement strategy and the multi-instruction conditioning mechanism to achieve faithful and flexible reflection removal. During training, the positive and negative text conditions are jointly optimized through a CFG-aware denoising objective.

In addition to improving the recovery pipeline, *RefMover* introduces a multi-instruction conditioning mechanism for precise user control. This mechanism integrates a negative-only inference prompting strategy with spatial-awareness capabilities. The negative-only prompts prevent unwanted content generation artifacts, while the user-steered spatial mask enables targeted, region-specific reflection removal, thereby improving interaction flexibility and accuracy. Together, these components provide strong reflection-removal performance, improved fidelity, and precise user control with lower computational cost than our preliminary work. The overall framework of the proposed *RefMover*, illustrating both the training and inference pipelines, is depicted in Fig. 2.

B. Faithful and robust recovery design

As established, using the contaminated mixture image as a static control condition poses a threat to recovery faithfulness. To directly address this challenge and prevent artifacts such as color shifts, structural distortions, and loss of details, we designed the robust recovery pipeline of *RefMover*. This section details its three technical components, which work in concert to establish a foundation for high-fidelity and efficient reflection removal: a streamlined diffusion-based conditional U-Net architecture, a discrete condition refinement strategy, and simple but effective color fidelity preservation techniques.

Streamlined condition injection. We propose a streamlined condition injection mechanism that restructures the architectural design of L-DiffER [16]. This approach directly addresses the inherent limitations of L-DiffER's dual-branch ControlNet structure through modality reduction and simplified feature integration.

In L-DiffER, conditional control was achieved via two parallel ControlNet branches processing image and gradient information separately. Each ControlNet branch outputs a set of feature maps that are then added as residuals to the corresponding layers of the main U-Net. This feature integration across multiple resolution levels is expressed as:

$$f_l^{\text{out}} = \text{U-Net}_l(f_l^{\text{in}}) + \text{CN}_{\theta_I}(f_l^{\text{in}}, \mathbf{y}_I) + \text{CN}_{\theta_G}(f_l^{\text{in}}, \mathbf{y}_G), \quad (3)$$

where f_l^{in} and f_l^{out} are the input and output features of the l-th U-Net layer, respectively. CN_{θ_I} and CN_{θ_G} denote the feature maps generated by the image and gradient ControlNet branches with parameters θ_I and θ_G, conditioned on the image encoder output $\mathbf{y}_I$ and gradient encoder output $\mathbf{y}_G$. The subscript l indicates the layer level of the feature.

This dual-branch condition injection design, while allowing multi-modal conditioning, inherently introduced significant computational overhead and complex feature interactions. Our analysis revealed that the gradient condition branch, despite its theoretical benefits for structural guidance, often interfered with the image condition during feature extraction, resulting in inconsistent outputs and training instability. Additionally, the intricate cross-branch feature integration substantially increased parameter count, leading to higher computational demands and memory usage.

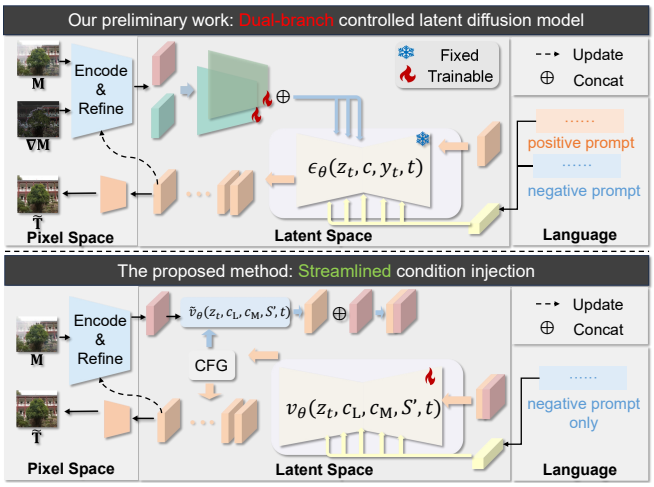


Fig. 3: Architectural comparison between L-DiffER **(top)** and the proposed *RefMover* **(bottom)**. *RefMover* replaces the dual-branch ControlNet with a streamlined, concatenation-based condition injection mechanism, which mitigates feature competition and improves stability.

To mitigate these issues, we reformulate the condition injection process. As visually compared in Fig. 3, *RefMover* replaces the complex dual-branch ControlNet architecture of L-DiffER with a more efficient, streamlined mechanism. Instead of using separate ControlNet branches, *RefMover* directly concatenates the encoded condition features with the latent diffusion representation:

$$\mathbf{z}_t^{cond} = [\mathbf{z}_t; \mathcal{E}_\phi(\mathbf{S}_t); \mathbf{S}'] \in \mathbb{R}^{H' \times W' \times 9}, \tag{4}$$

where $\mathbf{z}_t$ is the noisy latent at timestep t, $\mathcal{E}_\phi$ is the condition encoder, $\mathbf{S}'$ is the downsampled region-awareness mask, and $[\cdot;\cdot]$ denotes channel-wise concatenation. Here, $\mathbf{S}_t$ denotes the time-dependent image condition used by DCR and should not be confused with the region-awareness mask $\mathbf{S}$; it corresponds to $\mathbf{c}_t^M$ in Fig. 2. The image condition is initialized with the mixture image, $\mathbf{S}_T = \mathbf{M}$, and updated once by the discrete refinement strategy detailed subsequently. Consequently, the modified U-Net receives a 9-channel input consisting of the noisy latent, the encoded image condition, and the region mask:

$$f_1^{in} = \text{Conv}_{\text{expanded}}(\mathbf{z}_t^{cond}), \tag{5}$$

Deliberately omitting the gradient condition present in L-DiffER, this modality reduction eliminates cross-modal feature competition while reducing inference computational overhead. For effective integration of both visual and textual guidance, we implement a dual-condition Classifier-Free Guidance (CFG) during the inference process, similar to approaches like InstructPix2Pix [66]:

$$\hat{\mathbf{v}}_t = \mathbf{v}_u + \gamma_y(\mathbf{v}_y - \mathbf{v}_u) + \gamma_c(\mathbf{v}_y - \mathbf{v}_n), \tag{6}$$

Here, $\mathbf{v}_u = v_\theta(\mathbf{z}_t, t, \emptyset, \emptyset, \mathbf{S}')$, $\mathbf{v}_y = v_\theta(\mathbf{z}_t, t, \emptyset, \mathbf{y}, \mathbf{S}')$, and $\mathbf{v}_n = v_\theta(\mathbf{z}_t, t, \mathbf{c}_n, \mathbf{y}, \mathbf{S}')$ denote the unconditional, image-conditioned, and negative-text-conditioned velocity predictions, respectively. The image term $\mathbf{v}_y - \mathbf{v}_u$ strengthens adherence to the mixture-image condition, whereas the text term $\mathbf{v}_y - \mathbf{v}_n$ steers the prediction away from the reflection semantics specified by $\mathbf{c}_n$. The scales γ_y and γ_c independently control these two guidance directions. Comprehensive ablation studies and performance analyses of our streamlined architecture are presented in Sec. V.

Discrete condition refinement. As articulated in Sec. III-A, relying solely on the raw mixture image $\mathbf{M}$ as a static condition throughout inference poses challenges, particularly in regions heavily contaminated by reflections, where the transmission content is obscured. Based on this observation, L-DiffER [16] proposed an iterative condition refinement (ICR) strategy. This strategy aimed to improve the spatial condition over time by blending the initial condition (derived from $\mathbf{M}$ and its gradient $\nabla\mathbf{M}$) with the estimated pseudo transmission layer $\mathbf{T}_{0|t}$ derived from the model's prediction at step t.

In L-DiffER, the spatial condition comprised both a color condition $\mathbf{I}_c$ and a structure condition $\mathbf{I}_s$. At each timestep t during sampling, L-DiffER estimated the clean latent representation $\mathbf{z}_{0|t}$ from its noise prediction, then decoded it to obtain a pseudo transmission layer $\mathbf{T}_{0|t}$ in pixel space:

$$\mathbf{T}_{0|t} = \mathcal{D}\left(\frac{\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\cdot\hat{\epsilon}_t}{\sqrt{\overline{\alpha}_t}}\right), \tag{7}$$

where $\overline{\alpha}_t$ is a term from the diffusion noise schedule [67], [68]. Subsequently, L-DiffER continuously updated the color and structure conditions by blending $\mathbf{T}_{0|t}$ (and its gradient) with the initial mixture image $\mathbf{M}$ (and its gradient). To maintain physical plausibility, constraints such as a Maximum Value Constraint (MVC) were applied to these continuously refined input conditions.

While the continuous iterative condition refinement strategy in L-DiffER offered some adaptability, its dense application at every timestep presented drawbacks that compromised stability and output fidelity in practice. The issues stemmed from the potential of condition drift, where continuous refinement could lead the condition to progressively deviate from the ground truth transmission, especially in challenging regions, causing artifacts or over-smoothing. Furthermore, error propagation occurred as inaccuracies in the initial predictions of $\mathbf{T}_{0|t}$ were continuously fed back into the condition, potentially amplifying errors throughout the sampling process. Lastly, performing full condition updates and constraint applications at every sampling step led to increased computational overhead. To mitigate these issues, *RefMover* introduces a discrete condition refinement strategy. Instead of continuous updates, we strategically refine the conditions at predetermined, sparse timesteps. We refine the condition once at the fixed middle solver state of the three-step trajectory.

For the v-prediction model used in *RefMover*, the corresponding pseudo transmission is recovered as

$$\mathbf{T}_{0|t} = \mathcal{D}\left(\sqrt{\overline{\alpha}_t}\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\hat{\mathbf{v}}_t\right), \tag{8}$$

where $\hat{\mathbf{v}}_t$ denotes the guided velocity prediction.

The process begins by initializing the spatial condition with the mixture image, i.e., $\mathbf{S}_T = \mathbf{M}$, where T denotes the initial high-noise timestep of the reverse process. Instead of continuous updates, the condition is refined only once at the middle solver state of our 3-step DPM-Solver++ inference trajectory. We denote this middle update timestep as t_m. At

this timestep, the condition is updated by blending the current condition $\mathbf{S}_{t_m}$ with the estimated pseudo transmission layer $\mathbf{T}_{0|t_m}$. To limit its deviation from the original mixture image, the updated condition is processed by the MVC operator. This update operation is expressed as:

$$\mathbf{S}_{\text{new}} = \mathbf{MVC}(\mathbf{M}, \beta_{t_m} \cdot \mathbf{T}_{0|t_m} + (1 - \beta_{t_m}) \cdot \mathbf{S}_{t_m}), \quad (9)$$

where β_{t_m} is the blending coefficient at the middle update timestep. Outside this update step, the condition remains fixed, limiting repeated feedback. This discrete refinement mechanism, visually represented in the "Condition Refinement" block of Fig. 2, improves the robustness and fidelity of the final output while remaining consistent with the 3-step inference setting used in all main experiments. Experimental validation concerning the effectiveness and robustness of this sparse strategy is presented in Sec. V.

Fidelity preservation. *RefMover* incorporates two strategies to ensure the fidelity of the recovered transmission layer, addressing challenges inherent to both the reflection removal task, where colors are distorted by the additive reflection layer [4], [11], [52], and the diffusion models themselves, which can struggle with color consistency [63].

1) Addressing dynamic range with zero terminal SNR: A common issue in diffusion models stems from noise schedules that lack a zero Signal-to-Noise Ratio (SNR) at the terminal timestep ($t = T$) [64]. This discrepancy between training and inference can bias the model towards the mean brightness of the training data. To overcome this, we implement a solution inspired by recent findings [64]. Specifically, we enforce a zero SNR at the terminal timestep by rescaling the noise schedule (i.e., the $\bar{\alpha}_t$ values), train the network with a v-prediction objective [65], and ensure the reverse sampling process starts from pure noise. During inference, the classifier-free guidance scale is also rescaled to prevent over-exposure [64]. This setup provides the model with the capacity to generate samples that span a wider brightness distribution.

2) Texture transfer and error correction via residual refinement: While the adjusted noise schedule improves the dynamic range, the probabilistic nature of the diffusion process can still lead to two issues: loss of fine-grained high-frequency textures (over-smoothing) and subtle structural or color deviations. To address these limitations, we replace the statistical post-processing with a learnable **Residual Refinement Module (RRM)**.

Network Architecture. Unlike standard restoration networks that operate solely on the degraded image, our RRM is designed as a reference-guided detail injector. As shown in our implementation, we adopt a lightweight U-Net architecture [69] rather than a simple stack of residual blocks [70]. This choice is motivated by the need to capture multi-scale context: the encoder extracts structural cues from the diffusion output $\mathbf{T}_{\text{diff}}$ and texture details from the original mixture $\mathbf{M}$, while the decoder progressively recovers spatial resolution. Crucially, skip connections facilitate the direct transfer of high-frequency information from the encoder to the decoder, effectively alleviating the information bottleneck.

Mathematically, the RRM takes the channel-wise concatenation of the diffusion output and the original mixture image as input. The final recovered transmission $\mathbf{T}_{\text{final}}$ is formulated as:

$$\mathbf{T}_{\text{final}} = \mathbf{T}_{\text{diff}} + \mathcal{U}_\psi([\mathbf{T}_{\text{diff}}; \mathbf{M}]), \quad (10)$$

where $[\cdot;\cdot]$ denotes concatenation, and $\mathcal{U}_\psi$ represents the U-Net refiner parameterized by ψ. The network outputs a residual map that corrects artifacts and restores missing textures.

Zero-Initialization Strategy. To stabilize training, we employ a zero-initialization strategy for the final convolution layer of the RRM. Specifically, the weights and biases of the last output layer are initialized to zeros. This ensures that at the beginning of training, the residual term is zero, and $\mathbf{T}_{\text{final}} \approx \mathbf{T}_{\text{diff}}$. Consequently, the RRM essentially starts as an identity mapping for the diffusion output and progressively learns to add missing details only where necessary. This prevents the refinement module from degrading the high-quality semantic content already generated by the diffusion model during the early stages of training.

C. Multi-instruction conditioning mechanism

To bridge the semantic-spatial gap and empower users with precise, intuitive control, *RefMover* incorporates a multi-instruction conditioning mechanism. This mechanism moves beyond unimodal guidance by synergistically combining two distinct channels of user intent: semantic cues derived from natural language (denoted as c) and explicit spatial directives provided by a region-awareness mask (denoted as $\mathbf{S}$). By combining these instructions, *RefMover* achieves a level of user-steerability and fidelity that directly addresses the limitations of purely language-based or fully automatic approaches.

Negative-only text prompt strategy. L-DiffER [16] utilizes both positive prompts (c_p), describing the content in the transmission layer $\mathbf{T}$, and negative prompts (c_n), specifying undesired attributes and characterizing the reflections to be suppressed. While effective, formulating detailed positive prompts can be difficult and risks inducing the model to generate content not originally present in the scene. As illustrated in Fig. 2, *RefMover* simplifies this mechanism by adopting a negative-only text prompt strategy for inference. This refinement stems from the observation that describing the undesired reflection artifacts (e.g., "glare," "window pattern," "streaky lights") is often more direct and less prone to ambiguity than describing the entire, potentially complex, transmission layer. Moreover, using positive prompts in restoration tasks often leads to semantic hallucinations, where the model fabricates background details that do not exist in the original scene. Our negative-only strategy steers the denoising process purely away from reflection artifacts, preserving the physical fidelity of the background. During training, the positive condition c_p and the negative condition c_n are jointly optimized through the CFG-aware denoising objective defined in Eq. (16). During inference, c_p is replaced with an empty text condition, while c_n is retained as the negative CFG branch.

The negative-only strategy is integrated into *RefMover*'s dual-condition CFG mechanism, as detailed in Eq. (6). During inference, the language description c serves exclusively as

the negative prompt c_n. The term relevant to text guidance becomes:

$$\Delta \mathbf{v}_{\text{text}} = \gamma_c(\mathbf{v}_y - \mathbf{v}_n). \tag{11}$$

Unlike L-DiffER, where the negative prompt effect was relative to a positive-prompt-conditioned prediction, here the negative prompt c_n steers the denoising process away from the specified reflection semantics relative to the image-only prediction $\mathbf{v}_y$. This focused, subtractive guidance targets undesired reflection content without introducing a positive semantic generation direction.

User-steered spatial control with region masks. While L-DiffER [16] demonstrated the utility of deriving spatial conditions from the global mixture image, it inherently lacked a direct mechanism for users to pinpoint specific, often localized areas where reflections are most disruptive. Reflections frequently exhibit significant spatial non-uniformity; some regions might be heavily contaminated while others remain relatively clear. A global processing strategy may not optimally address such heterogeneity. To overcome this limitation and empower users with more granular control, *RefMover* introduces user-steered spatial control enabled by a region-awareness mask $\mathbf{S} \in \mathbb{R}^{H\times W\times 1}$. This mask allows users to explicitly designate regions demanding targeted reflection mitigation, thereby aligning the removal process more closely with user intent—a core objective for enhanced interactivity and practical utility.

1) Training strategy: To effectively integrate this user-provided spatial directive, *RefMover* employs a synergistic strategy during training. Our streamlined U-Net architecture is conditioned on a 9-channel input tensor, which is formed by concatenating the noisy latent $\mathbf{z}_t \in \mathbb{R}^{H'\times W'\times 4}$, the downsampled user-provided region mask $\mathbf{S}' \in \mathbb{R}^{H'\times W'\times 1}$, and the latent representation of the reflection-contaminated mixture image, denoted as the encoded primary image condition $\mathbf{y}_{cond} = \mathcal{E}_\phi(\mathbf{S}_t) \in \mathbb{R}^{H'\times W'\times 4}$. For generating diverse and realistic region masks $\mathbf{S}$ to simulate various user inputs, we adopt the mask generation strategy proposed in LaMa [71]. Furthermore, to robustify training and encourage the model to learn strong priors for reflection removal within user-specified areas, we introduce a stochastic masking strategy. Specifically, in 20% of training iterations, the full condition $\mathbf{y}_{cond}$ is replaced by a masked version $\mathbf{y}^m_{cond} = \mathbf{y}_{cond} \odot (\mathbf{1} - \mathbf{S}')$. This effectively tasks the model with an inpainting-like objective within the masked regions, relying on the context from unmasked areas. This stochastic masking encourages the denoiser to infer the transmission content within user-specified regions from the unmasked context and the learned restoration prior.

2) Inference strategy: In inference, this learned spatial prior is leveraged to achieve targeted yet harmonious results, using a strategy inspired by the Differential Diffusion framework [72]. The Differential Diffusion framework enables spatially varying editing by maintaining separate denoising paths for regions requiring different levels of modification. We adapt this concept to reflection removal by progressively blending model-generated content in reflection regions while preserving original content elsewhere. *RefMover* utilizes a soft, grayscale version of the user-provided mask $\mathbf{S}$, which is processed by a Gaussian blur and then normalized to serve as a "change map" $\mathbf{S}_c$. In this map, $\mathbf{S}_c = 0$ signifies regions to be preserved, while $\mathbf{S}_c = 1$ indicates regions for maximal reflection removal.

The idea is to iteratively blend the output of the denoising network with a re-noised version of the input mixture's latent features, guided by the change map $\mathbf{S}_c$. Let $\mathbf{M}$ be the input mixture image and its initial latent representation be $\mathbf{z}^{\text{M, init}} = E_\phi(\mathbf{M})$. The reverse diffusion process proceeds from $t = T-1$ down to 0. At each step t: First, a noised version of the initial mixture latent, $\mathbf{z}_t^{\text{M, noised}}$, is computed by adding noise corresponding to timestep t to $\mathbf{z}^{\text{M, init}}$. This acts as an anchor for regions intended for preservation. Let $\mathbf{z}_{t-1}^{\text{denoised}}$ be the latent produced by one reverse step from $\mathbf{z}_t^{\text{mix}}$. A dynamic binary mask, $\mathbf{B}_t$, is then constructed based on the downsampled change map $\mathbf{S}'_c$ and a progressively evolving threshold, $\tau_t = (T-t)/T$. This threshold can be interpreted as the denoising "progress," as it evolves from 0 to 1 during the process. The binary mask $\mathbf{B}_t(i,j)$ is set to 1 if the required change $\mathbf{S}'_c(i,j)$ exceeds the current progress τ_t:

$$\mathbf{B}_t(i,j) = \begin{cases} 1 & \text{if } \mathbf{S}'_c(i,j) > \tau_t, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

This mechanism ensures that a pixel is updated by the model only as long as its modification is incomplete; once the denoising progress surpasses the required change, the pixel is "locked" to the original image's trajectory to preserve its content. Consequently, regions designated for substantial change (higher $\mathbf{S}'_c$ values) will have $\mathbf{B}_t = 1$ for a larger duration of the denoising process. Finally, the latent representation $\mathbf{z}_{t-1}^{\text{mix}}$, which serves as the input to the denoising U-Net at the next timestep, is composed by blending the two latents according to $\mathbf{B}_t$:

$$\mathbf{z}_{t-1}^{\text{mix}} = \mathbf{z}_{t-1}^{\text{denoised}} \odot \mathbf{B}_t + \mathbf{z}_{t-1}^{\text{M, noised}} \odot (1 - \mathbf{B}_t). \tag{13}$$

This iterative, spatially-variant blending ensures that regions for strong reflection removal are guided by the model's generative capacity, while areas marked for preservation remain anchored to the input's original characteristics. This approach creates smoother transitions and mitigates "pasted-on" artifacts, ultimately enhancing the precision and visual coherence of the restored transmission layer.

D. Loss function and implementation details

Loss function. The training of *RefMover* involves optimizing two distinct components: the core diffusion U-Net and the subsequent Residual Refinement Module (RRM).

Diffusion U-Net Objective. The diffusion backbone is trained based on a v-prediction objective [65], a parameterization known for improving training stability and sample quality. The U-Net denoiser, denoted as v_θ, is trained to predict a target "velocity" $\mathbf{v}_{\text{target}}$. Given a clean image latent $\mathbf{z}_0$, a timestep $t \sim U\{1,\dots,T\}$, and a noise sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I})$, the noisy latent $\mathbf{z}_t$ is formulated as:

$$\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \tag{14}$$

where $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1-\bar{\alpha}_t}$ are terms derived from the noise schedule. The corresponding target velocity $\mathbf{v}_{\text{target}}$ is defined as:

$$\mathbf{v}_{\text{target}} = \sqrt{\bar{\alpha}_t}\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_0. \tag{15}$$

Let $\mathbf{v}_p = v_\theta(\mathbf{z}_t, t, \mathbf{c}_p, \mathbf{y}^m_{\text{cond}}, \mathbf{S}')$ and $\mathbf{v}_n = v_\theta(\mathbf{z}_t, t, \mathbf{c}_n, \mathbf{y}^m_{\text{cond}}, \mathbf{S}')$ denote the velocity predictions conditioned on the positive and negative text conditions, respectively. During training, these predictions are combined using classifier-free guidance, and the mean squared error objective is applied to the resulting guided prediction:

$$\begin{aligned} \hat{\mathbf{v}}_{\text{cfg}} &= \mathbf{v}_n + \gamma_c^{\text{train}}(\mathbf{v}_p - \mathbf{v}_n), \\ \mathcal{L}_{\text{diff}} &= \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon}}\left[\|\mathbf{v}_{\text{target}} - \hat{\mathbf{v}}_{\text{cfg}}\|_2^2\right], \end{aligned} \tag{16}$$

where γ_c^{train} denotes the text CFG scale used during training, while γ_c denotes its inference-time counterpart in Eq. (6). The term $\mathbf{y}^m_{\text{cond}}$ represents the stochastically masked condition introduced in Sec. III-C, defined as:

$$\mathbf{y}^m_{\text{cond}} = \begin{cases} \mathbf{y}_{\text{cond}} & \text{with probability } 0.8, \\ \mathbf{y}_{\text{cond}} \odot (1 - \mathbf{S}') & \text{with probability } 0.2. \end{cases} \tag{17}$$

Residual Refinement Module Objective. To drive the RRM to faithfully recover high-frequency textures and structural boundaries, we train it using a composite objective function that combines pixel-wise reconstruction loss with perceptual, structural (SSIM), and edge-aware losses. This multi-term loss ensures that the refined output maintains pixel fidelity while effectively retrieving lost details from the reference mixture image. The detailed loss definition and weights are provided in the Appendix.

Implementation details. We implement *RefMover* within the Diffusers library [73]. The core architecture is a U-Net adapted from the pre-trained Stable Diffusion 2.1 model [19], which has approximately 860M parameters. We modify the first convolutional layer of the U-Net to accept the additional channels from our concatenated conditions. For encoding the conditions, we use the pre-trained OpenCLIP ViT-H/14 text encoder distributed with the Stable Diffusion 2.1 checkpoint and its corresponding Variational Autoencoder (VAE) for converting images into the latent space.

Training. The training is conducted in two stages. First, the diffusion model is fine-tuned for 50 epochs on our proposed large-scale reflection dataset (detailed in Sec. IV). All training images are resized and center-cropped to a resolution of 768×768. We utilize the Adam optimizer with a constant learning rate of 5×10^{-5} without a learning rate scheduler. To fit training into memory, we use a per-device batch size of 16 with 2 gradient accumulation steps across two GPUs, resulting in a global effective batch size of 64. A conditioning dropout probability of 0.05 is applied to the text and image conditions to enhance robustness. Subsequently, the Residual Refinement Module is trained for 50 epochs with a batch size of 64 on 768×768 resolution crops, utilizing the AdamW optimizer with a cosine annealing scheduler. The entire training procedure is performed with mixed-precision (fp16) training on a system equipped with two NVIDIA RTX PRO 6000 GPUs.

Inference. For image generation, we employ DPM-Solver++ [68], a high-order ODE solver that allows for high-quality synthesis in a few steps. We use 3 sampling steps for all main experiments, except for the sampling-step sensitivity analysis in the Appendix. This setting provides the best speed–quality trade-off according to the ablation in the Appendix.

IV. Dataset

A. Training dataset

Motivation and Synthesis Strategy. Training robust diffusion models benefits from broad semantic diversity. While Kee *et al.* [10] advocate for linear RAW synthesis to ensure physical additivity, this restricts training to explicitly captured RAW data. However, public RAW datasets (e.g., MIT-Adobe FiveK [74], SIDD [75]) are limited in scale ($\sim 10^4$) and semantic diversity compared to the large-scale data required for generative priors. Consequently, we utilize the SA-1B dataset [76], which contains over 11 million diverse images in 8-bit format. To address the lack of linearity and the variety of proprietary ISPs in real-world images, we adopt a pragmatic *learning-in-the-wild* strategy. Instead of ill-posed RAW inversion, we design a synthesis pipeline that operates on 8-bit images while meticulously simulating the optical and non-linear characteristics of real-world photography.

Data Collection and Physics-Inspired Synthesis. We collected 10,243 real-world reflection images (ComRR-R) using a custom rig shown in Fig. 5 with a black light-absorbing cloth to capture pure reflection layers across diverse lighting conditions and glass types. To construct the training set, these layers are composited with SA-1B backgrounds using a rigorous augmentation protocol designed to bridge the domain gap. Specifically, we simulate "geometric variations" via random scaling and flipping. "Depth-of-field effects (defocus blur)" are modeled by applying Gaussian blur with variable kernel sizes ($\sigma \in [0.1, 2.0]$) to the reflection layer, mimicking the optical phenomenon where reflections are out of focus. "Ghosting effects" are simulated by blending the primary reflection with a spatially offset copy. Finally, we apply "non-linear brightness adjustments" to approximate gamma compression in ISP pipelines. An automated screening mechanism filters out excessive overexposure. In total, this pipeline generated 30,729 semi-synthetic scenes. Detailed protocols and parameters are provided in the Appendix.

B. Testing dataset

To rigorously evaluate *RefMover*, we introduce two test datasets designed to address the limitations of existing benchmarks regarding resolution and complexity.

ComRR-Q (Quantitative). This dataset comprises 201 pairs of mixture and ground-truth transmission images captured using detachable glass. We employ homography-based alignment with SIFT features to ensure pixel-level accuracy for metrics such as PSNR and SSIM. The dataset covers challenging "in-the-wild" scenarios with high-frequency textures and dynamic lighting.



Fig. 4: Examples from our ComRR datasets. **First column:** Real-world reflection images from ComRR-R. **Second column:** Sample images from ComRR-Q for quantitative evaluation. **Third column:** Sample images from ComRR-V for visual quality assessment.

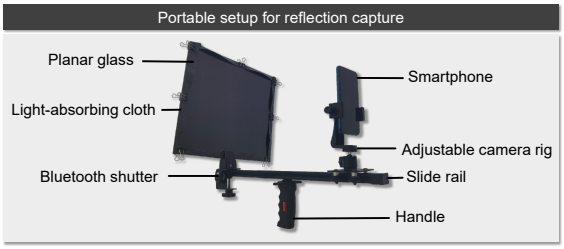


Fig. 5: The custom-designed portable capture setup used to efficiently acquire real-world reflection images. It comprises an adjustable camera rig, a sheet of planar glass, and a black light-absorbing cloth that occludes background light, thereby capturing pure reflection images.

ComRR-V (Visual). Intended for qualitative assessment, ComRR-V consists of 104 single-shot real-world images featuring highly complex reflections (e.g., multi-layer reflections, colored glass) where ground truth capture is impossible. It is used exclusively for qualitative evaluation. The quantitative dataset-complexity analysis in the Appendix is conducted on the aligned ComRR-Q split.

V. Experiments

A. Comparison with state-of-the-art methods

We evaluate our model on existing real reflection datasets SIR^2 [79] (with three subsets Postcard, Object, and Wild), Real [27], Nature [11], and the ComRR-Q dataset, our reflection removal dataset with ground truth collected as described in Sec. IV. Qualitative comparison experiments are performed on our ComRR-V dataset, a test dataset of complex reflection scenes. The compared methods include state-of-the-art single-image reflection removal approaches: IBCLN [11], DX21 [12], YTMT [52], DSRNet [13], SZ23 [78], L-DiffER [16], ZF24 [15], DSIT [14], and RDNet [77]. To evaluate the effectiveness of our training dataset generated from ComRR-R, we finetune SZ23 [78], DSRNet [13], ZF24 [15], and DSIT [14], as the training code of RDNet [77] is not provided. Because Kee *et al.* [10] is designed for linear RAW inputs whereas our benchmarks contain only 8-bit sRGB images, we exclude it from the main quantitative table. The Appendix instead reports a cross-domain diagnostic on 8-bit inputs, which does not represent native RAW-domain performance.

Quantitative comparison. For the quantitative comparison experiments, evaluation metrics include commonly used full-reference metrics in the field of reflection removal [12], [32]: Peak signal-to-noise ratio (PSNR) [80], structural similarity index measure (SSIM) [81], and learned perceptual image patch similarity (LPIPS) [82]. Additionally, no-reference metrics are employed to assess the visual quality of the reflection removal results. These include the multi-scale image quality transformer (MUSIQ) [83], CLIP image quality assessment (CLIP-IQA) [84], and the multi-dimensional attention network for image quality assessment (MANIQA) [85].

RefMover achieves the best aggregate average over the 741-image evaluation set on all six metrics. Although it does not rank first on every individual dataset and metric, the results demonstrate a strong overall balance between reference-based fidelity and perceptual quality. In particular, its high PSNR and SSIM and low LPIPS suggest that the recovered transmission layers retain pixel-aligned structures while maintaining competitive no-reference perceptual quality.

Qualitative comparison. Qualitative comparisons with the aforementioned single-image reflection removal methods [11]–[15], [52], [78] and L-DiffER [16] on recovering transmission layers are shown in Fig. 6.

As the qualitative results in Fig. 6 demonstrate, conventional single-image reflection removal methods encounter obstacles in complex real-world scenes. Lacking a mechanism to incorporate high-level semantic priors, they can struggle to resolve the ambiguities inherent to this ill-posed problem, resulting in residual artifacts or background degradation. Furthermore, without a means to bridge the semantic-spatial gap, they do not support targeted, user-directed removal. In contrast, *RefMover*, guided by its multi-instruction mechanism,

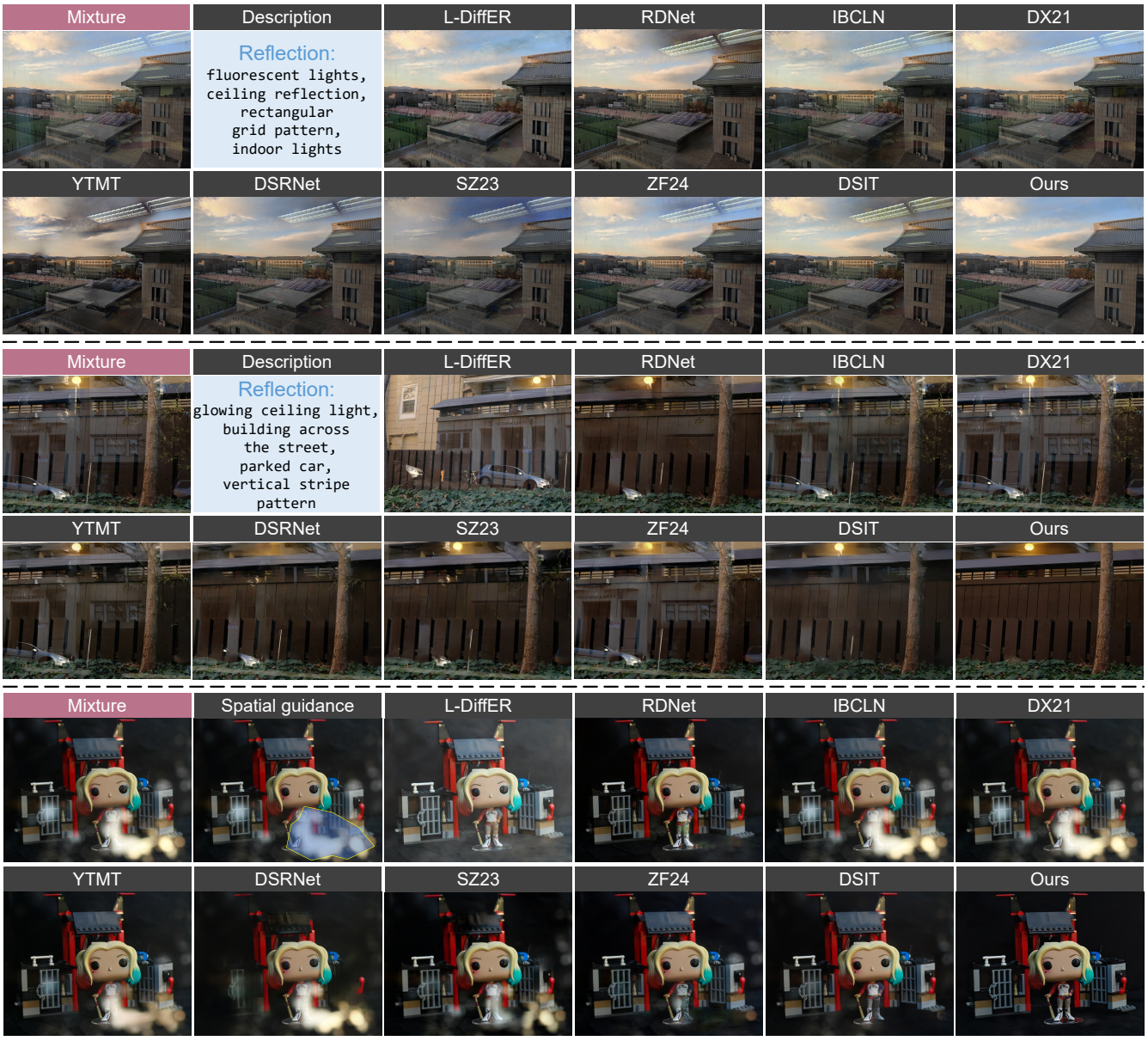


Fig. 6: Qualitative comparison with state-of-the-art methods on real-world images. The methods shown are (from left to right): L-DiffER [16], RDNet [77], IBCLN [11], DX21 [12], YTMT [52], DSRNet [13], SZ23 [78], ZF24 [15], DSIT [14], and our *RefMover*. Please zoom in for details.

more effectively removes the specified reflections while preserving the underlying transmission layer, achieving results with fewer artifacts and clearer details. L-DiffER [16] generates noisy results with its dual-branch network and refinement strategy. The proposed method effectively removes reflections while maintaining high fidelity in the transmission layer with faithful recovery and flexibility to handle user input.

B. Ablation study

We conducted an ablation study to validate the efficacy of our key design choices, with results shown in Table II and visually confirmed in Fig. 7. The ablation results confirm the overall contribution of each component, although removing the RRM produces a slightly higher MUSIQ score.

For "W/o SG.," the mask channel is removed from the concatenated U-Net input, and mask-dependent blending in Differential Diffusion is disabled. The "W/o T.&SG." configuration additionally uses an empty text condition. Despite removing both guidance sources, it remains competitive in the 741-image aggregate average, exceeding the best competing result on five of the six metrics, with SSIM as the exception. This demonstrates the strong restoration capability of the image-conditioned backbone.

Guidance mechanisms. Text and spatial guidance provide complementary forms of user control. Text guidance describes the reflection characteristics to suppress, whereas spatial guidance indicates where stronger modification is desired. Removing either source reduces the aggregate performance and weakens the corresponding form of controllability, although the partially guided variants remain effective reflection-removal models. Moreover, the "W/o T.&SG." configuration exceeds the best competing result on five of the six aggregate metrics, with SSIM as the exception, showing that the

TABLE I: Comparison of quantitative results on real datasets for evaluating the recovery of transmission layers, compared with several state-of-the-art single-image reflection removal methods [13]–[16], [77], [78]. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results, and underline numbers indicate the second-best results. * indicates we finetune the method with our training dataset. For *RefMover*, we prompt Qwen2.5-VL-7B-Instruct [86] with the fixed instruction, "Use a few words to describe the content and visual characteristics of the reflections in this image." The returned keywords are used directly without manual editing. A coarse mask is manually drawn by inspecting only the mixture image; neither guidance source uses the ground truth. The unguided variant is reported in Table II.

Dataset (size)	Metric	SZ23	SZ23*	DSRNet	DSRNet*	ZF24	ZF24*	DSIT	DSIT*	RDNet	L-DiffER	Ours
ComRR-Q (201)	SSIM↑	0.7962	0.8081	0.8143	0.8192	0.8259	0.8280	0.8324	0.8335	0.8322	0.8030	**0.8620**
	PSNR↑	20.59	21.39	20.65	20.61	21.55	22.01	22.06	22.60	21.91	20.90	**24.02**
	LPIPS↓	0.2370	0.2181	0.2030	0.2000	0.1789	0.1782	0.1667	0.1546	0.1646	0.1960	**0.1196**
	CLIPIQA↑	0.4333	0.4174	0.5146	0.4758	0.4822	0.4925	0.5153	0.4937	**0.5276**	0.4870	0.5245
	MANIQA↑	0.3236	0.3337	0.4587	0.4680	0.4532	0.4586	0.4518	0.4576	0.4707	0.4497	**0.4818**
	MUSIQ↑	59.96	60.37	65.63	66.81	66.00	66.21	66.49	66.48	66.64	66.71	**67.26**
SIR^2 (500)	SSIM↑	0.8768	0.8824	0.9211	**0.9302**	0.9126	0.9100	0.9240	0.9260	0.9278	0.9110	0.9296
	PSNR↑	22.68	23.29	25.50	25.61	25.58	25.65	25.90	26.10	26.49	25.18	**29.07**
	LPIPS↓	0.1670	0.1582	0.0912	0.0906	0.1151	0.1040	0.0880	0.0868	0.0809	0.1210	**0.0457**
	CLIPIQA↑	0.3853	0.3651	0.4060	0.4103	0.3898	0.4042	0.3900	0.4062	0.3846	0.3840	**0.4644**
	MANIQA↑	0.2755	0.2689	0.3262	0.3170	0.3025	0.3167	0.3270	0.3350	0.3133	0.2890	**0.3377**
	MUSIQ↑	52.99	53.43	59.48	60.03	57.74	59.35	59.56	60.66	58.60	57.35	**61.02**
Real (20)	SSIM↑	0.8013	0.8291	0.8118	0.8232	0.8222	0.8387	0.8420	**0.8458**	0.8412	0.8210	0.8302
	PSNR↑	23.06	23.70	22.16	23.04	23.43	24.54	24.48	**24.65**	24.20	23.77	23.82
	LPIPS↓	0.2063	0.1855	0.1654	0.1594	0.1604	0.1596	0.1353	0.1346	0.1532	0.1530	**0.1298**
	CLIPIQA↑	0.4208	0.4378	0.5224	0.5212	0.4789	0.4866	**0.5318**	0.5014	0.5240	0.4905	0.5204
	MANIQA↑	0.3159	0.3270	0.4234	0.4356	0.4149	0.4219	0.4374	**0.4399**	0.4362	0.4075	0.4383
	MUSIQ↑	54.53	54.64	62.19	62.50	61.07	62.05	61.40	61.98	60.99	61.25	**62.78**
Nature (20)	SSIM↑	0.7765	0.7958	0.8036	0.8102	0.8440	0.8400	0.8105	0.8133	**0.8447**	0.8310	0.8306
	PSNR↑	21.05	21.08	21.57	21.85	25.98	26.02	22.41	22.89	**26.06**	23.95	24.95
	LPIPS↓	0.2109	0.2067	0.1492	0.1452	0.1055	**0.1042**	0.1394	0.1303	0.1081	0.1450	0.1272
	CLIPIQA↑	0.4106	0.3676	0.4494	0.4490	0.4508	0.4568	0.4447	0.4248	0.4529	0.4459	**0.4610**
	MANIQA↑	0.3023	0.2980	0.3605	0.3584	0.3599	0.3612	0.3756	0.3743	0.3604	0.3602	**0.3912**
	MUSIQ↑	56.49	55.30	63.01	63.79	63.09	63.25	**64.25**	64.18	60.47	63.01	62.85
Average (741)	SSIM↑	0.8502	0.8585	0.8860	0.8940	0.8848	0.8839	0.8939	0.8957	0.8973	0.8771	**0.9059**
	PSNR↑	22.08	22.72	23.99	24.08	24.44	24.65	24.73	25.03	25.17	23.95	**27.45**
	LPIPS↓	0.1882	0.1765	0.1251	0.1236	0.1334	0.1256	0.1120	0.1077	0.1063	0.1429	**0.0702**
	CLIPIQA↑	0.4000	0.3814	0.4398	0.4321	0.4189	0.4318	0.4294	0.4330	0.4290	0.4165	**0.4821**
	MANIQA↑	0.2904	0.2888	0.3656	0.3623	0.3480	0.3592	0.3651	0.3721	0.3606	0.3377	**0.3809**
	MUSIQ↑	55.02	55.39	61.32	62.04	60.21	61.39	61.61	62.37	60.89	60.15	**62.81**

TABLE II: Ablation studies on text guidance (W/o T.), spatial guidance (W/o SG.), their joint removal (W/o T.&SG.), discrete condition refinement (W/o Rfn.), and the residual refinement module (W/o RRM). ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

Metric	W/o T.	W/o SG.	W/o T.&SG.	W/o Rfn.	W/o RRM	Ours
PSNR↑	26.44	27.23	25.98	26.38	26.99	**27.45**
SSIM↑	0.895	0.902	0.885	0.892	0.879	**0.906**
LPIPS↓	0.086	0.074	0.082	0.083	0.080	**0.070**
CLIPIQA↑	0.476	0.481	0.468	0.481	0.471	**0.482**
MANIQA↑	0.374	0.378	0.375	0.377	0.371	**0.381**
MUSIQ↑	62.57	62.71	62.48	62.65	**62.87**	62.81

image-conditioned backbone itself provides strong restoration capability. The full guided configuration nevertheless provides the best overall trade-off across the six metrics and enables explicit semantic and spatial control.

Faithfulness components. Removing discrete condition refinement consistently reduces performance, suggesting that the middle-step condition update improves the image condition and helps preserve structural details. Removing the RRM decreases PSNR, SSIM, CLIPIQA, and MANIQA and increases LPIPS, while slightly improving MUSIQ. This indicates that the RRM mainly improves pixel-aligned and structural fidelity, with a small trade-off in one no-reference metric.

VI. Conclusion

This paper introduced *RefMover*, a robust and user-steerable diffusion model designed to address challenges in SIRR. We began by identifying two main obstacles for language-guided approaches: the semantic-spatial gap that hinders precise control, and the insufficient recovery faithfulness caused by contaminated guidance conditions. The main contribution of this work is the multi-instruction framework for this task, which bridges the semantic-spatial gap by combining language and region guidance. To address the faithfulness challenge, we developed a robust recovery pipeline featuring a streamlined, concatenation-based architecture, a discrete condition refinement strategy, and a **residual refinement module** that help achieve visually authentic results. Furthermore, we collected and introduced the ComRR dataset to facilitate robust model training and evaluation.

Extensive experiments demonstrate that *RefMover* achieves the best aggregate average over the 741-image evaluation set and produces competitive qualitative results on challenging

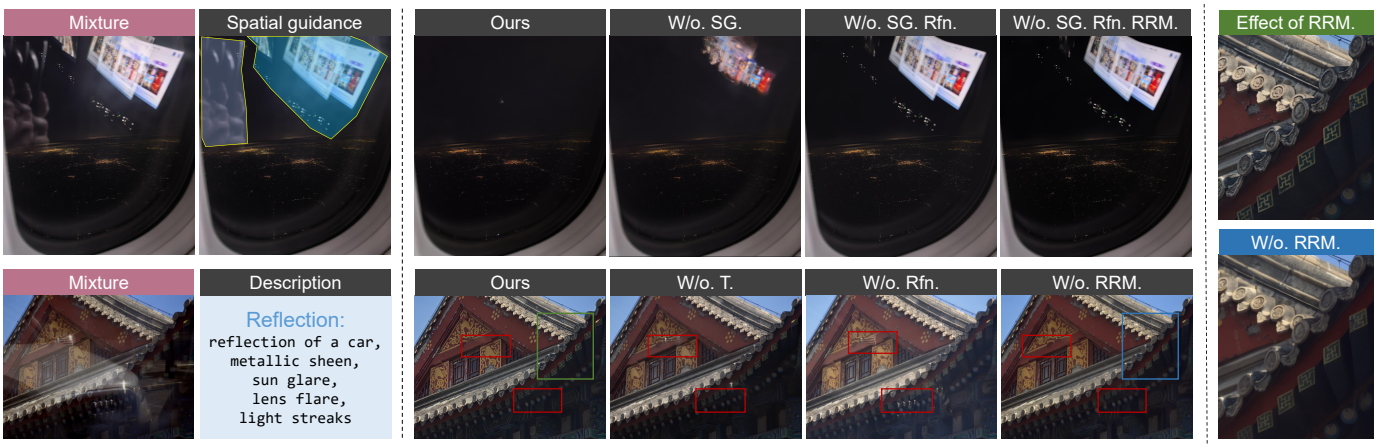


Fig. 7: Ablation of the guidance and faithfulness components. The top and bottom rows evaluate spatial and text guidance, respectively. The highlighted crops visualize the effects of discrete refinement and the RRM on fine-structure preservation in these examples.

real-world scenes. Our ablation studies further demonstrate the contributions of the proposed guidance and refinement components. While *RefMover* performs effectively, it may incorrectly identify the transmission and reflection layers if they exhibit highly similar image content. Future work could explore more granular user guidance, such as annotated edges or precise segmentation masks, to improve the model's disambiguation capabilities in these challenging scenarios.

REFERENCES

[1] Y. Chang, C. Jung, and J. Sun, "Joint reflection removal and depth estimation from a single image," *IEEE Transactions on Cybernetics*, vol. 51, no. 12, pp. 5836–5849, 2020.

[2] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Face image reflection removal," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 385–399, 2021.

[3] J. Qiu, P.-T. Jiang, Y. Zhu, Z.-X. Yin, M.-M. Cheng, and B. Ren, "Looking through the glass: Neural surface reconstruction against high specular reflections," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[4] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CoRRN: Cooperative reflection removal network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 12, pp. 2969–2982, 2019.

[5] R. Wan, B. Shi, H. Li, Y. Hong, L.-Y. Duan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, 2022.

[6] G. Dong, T. Zheng, Y. Cao, L. Qing, and C. Ren, "Channel consistency prior and self-reconstruction strategy based unsupervised image deraining," in *Proc. of Computer Vision and Pattern Recognition*, 2025.

[7] X. Chen, J. Pan, and J. Dong, "Bidirectional multi-scale implicit neural representations for image deraining," in *Proc. of Computer Vision and Pattern Recognition*, 2024.

[8] J. Lai, S. Chen, Y. Lin, T. Ye, Y. Liu, S. Fei, Z. Xing, H. Wu, W. Wang, and L. Zhu, "SnowMaster: Comprehensive real-world image desnowing via MLLM with Multi-Model feedback optimization," in *Proc. of Computer Vision and Pattern Recognition*, 2025.

[9] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I. Chen, J.-J. Ding, S.-Y. Kuo *et al.*, "ALL Snow Removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. of International Conference on Computer Vision*, 2021.

[10] E. Kee, A. Pikielny, K. Blackburn-Matzen, and M. Levoy, "Removing reflections from raw photos," in *Proc. of Computer Vision and Pattern Recognition*, 2025.

[11] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[12] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. Lau, "Location-aware single image reflection removal," in *Proc. of International Conference on Computer Vision*, 2021.

[13] Q. Hu and X. Guo, "Single image reflection separation via component synergy," in *Proc. of International Conference on Computer Vision*, 2023.

[14] Q. Hu, H. Wang, and X. Guo, "Single image reflection separation via dual-stream interactive transformers," in *Proc. of Advances in Neural Information Processing Systems*, 2024.

[15] Y. Zhu, X. Fu, P.-T. Jiang, H. Zhang, Q. Sun, J. Chen, Z.-J. Zha, and B. Li, "Revisiting single image reflection removal in the wild," in *Proc. of Computer Vision and Pattern Recognition*, 2024.

[16] Y. Hong, H. Zhong, S. Weng, J. Liang, and B. Shi, "L-DiffER: Single image reflection removal with language-based diffusion model," in *Proc. of European Conference on Computer Vision*, 2024.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of International Conference on Machine Learning*, 2021.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. of Advances in Neural Information Processing Systems*, 2020.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[20] J. Sun, S. Weng, Z. Chang, S. Li, and B. Shi, "UniCoRN: A unified conditional image repainting network," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[21] S. Weng and B. Shi, "Conditional image repainting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2285–2298, 2024.

[22] J. Tang, H. Zhong, S. Weng, and B. Shi, "LuminAIRe: Illumination-aware conditional image repainting for lighting-realistic generation," in *Proc. of Advances in Neural Information Processing Systems*, 2023.

[23] S. Weng, H. Wu, Z. Chang, J. Tang, S. Li, and B. Shi, "L-CoDe: Language-based colorization using color-object decoupled conditions," in *Proc. of Association for the Advancement of Artificial Intelligence*, 2022.

[24] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, and B. Shi, "L-CAD: Language-based colorization with any-level descriptions using diffusion priors," in *Proc. of Advances in Neural Information Processing Systems*, 2023.

[25] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. of International Conference on Computer Vision*, 2023.

[26] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proc. of Computer Vision and Pattern Recognition*, 2024.

[27] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with

perceptual losses," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[28] S. K. Nayar, X.-S. Fang, and T. Boult, "Separation of reflection components using color and polarization," *International Journal of Computer Vision*, vol. 21, no. 3, pp. 163–186, 1997.

[29] Y. Y. Schechner, N. Kiryati, and R. Basri, "Separation of transparent layers using focus," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 25–39, 2000.

[30] Y. Diamant and Y. Y. Schechner, "Overcoming visual reverberations," in *Proc. of Computer Vision and Pattern Recognition*, 2008.

[31] N. Kong, Y.-W. Tai, and S. Y. Shin, "A physically-based approach to reflection separation," in *Proc. of Computer Vision and Pattern Recognition*, 2012.

[32] Y. Lyu, Z. Cui, S. Li, M. Pollefeys, and B. Shi, "Reflection separation using a pair of unpolarized and polarized images," in *Proc. of Advances in Neural Information Processing Systems*, 2019.

[33] C. Lei, X. Huang, M. Zhang, Q. Yan, W. Sun, and Q. Chen, "Polarized reflection removal with perfect alignment in the wild," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[34] Y. Lyu, Z. Cui, S. Li, M. Pollefeys, and B. Shi, "Physics-guided reflection separation from a pair of unpolarized and polarized images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2151–2165, 2022.

[35] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *Proc. of International Conference on Computer Vision*, 2013.

[36] C. Simon and I. Kyu Park, "Reflection removal for in-vehicle black box videos," in *Proc. of Computer Vision and Pattern Recognition*, 2015.

[37] Y.-L. Liu, W.-S. Lai, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Learning to see through obstructions with layered decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8387–8402, 2021.

[38] Y. Chang, C. Jung, J. Sun, and F. Wang, "Siamese dense network for reflection removal with flash and no-flash image pairs," *International Journal of Computer Vision*, vol. 128, no. 6, pp. 1673–1698, 2020.

[39] C. Lei and Q. Chen, "Robust reflection removal with reflection-free flash-only cues," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[40] C. Lei, X. Jiang, and Q. Chen, "Robust reflection removal with flash-only cues in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 530–15 545, 2023.

[41] Y. Hong, Y. Lyu, S. Li, and B. Shi, "Near-infrared image guided reflection removal," in *Proc. of International Conference on Multimedia and Expo*, 2020.

[42] Y. Hong, Y. Lyu, S. Li, G. Cao, and B. Shi, "Reflection removal with NIR and RGB image feature fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 7101–7112, 2022.

[43] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, W. Gao, and A. C. Kot, "Region-aware reflection removal with unified content and gradient priors," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2927–2941, 2018.

[44] Y. Yang, W. Ma, Y. Zheng, J.-F. Cai, and W. Xu, "Fast single image reflection suppression via convex optimization," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[45] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1647–1654, 2007.

[46] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. of Computer Vision and Pattern Recognition*, 2014.

[47] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. of Computer Vision and Pattern Recognition*, 2015.

[48] W. Zhu, J. Lu, J. Li, and J. Zhou, "DSNet: A flexible detect-to-summarize network for video summarization," *IEEE Transactions on Image Processing*, vol. 30, pp. 948–962, 2021.

[49] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. of International Conference on Computer Vision*, 2017.

[50] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[51] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[52] Q. Hu and X. Guo, "Trash or Treasure? An interactive dual-stream strategy for single image reflection separation," *Proc. of Advances in Neural Information Processing Systems*, 2021.

[53] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. of European Conference on Computer Vision*, 2018.

[54] Y.-N. Zhang, L. Shen, and Q. Li, "Content and gradient model-driven deep network for single image reflection removal," in *Proc. of ACM International Conference on Multimedia*, 2022.

[55] D. Ma, R. Wan, B. Shi, A. C. Kot, and L.-Y. Duan, "Learning to jointly generate and separate reflections," in *Proc. of International Conference on Computer Vision*, 2019.

[56] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proc. of Computer Vision and Pattern Recognition*, 2019.

[57] S. Kim, Y. Huo, and S.-E. Yoon, "Single image reflection removal with physically-based training images," in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[58] Q. Zheng, J. Chen, Z. Lu, B. Shi, X. Jiang, K.-H. Yap, L.-Y. Duan, and A. C. Kot, "What does plate glass reveal about camera calibration?" in *Proc. of Computer Vision and Pattern Recognition*, 2020.

[59] Q. Zheng, B. Shi, J. Chen, X. Jiang, L.-Y. Duan, and A. C. Kot, "Single image reflection removal with absorption effect," in *Proc. of Computer Vision and Pattern Recognition*, 2021.

[60] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-ControlNet: All-in-one control to text-to-image diffusion models," in *Proc. of Advances in Neural Information Processing Systems*, 2024.

[61] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[62] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-Prompt image editing with cross attention control," in *Proc. of International Conference on Learning Representations*, 2023.

[63] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon, "Perception prioritized training of diffusion models," in *Proc. of Computer Vision and Pattern Recognition*, 2022.

[64] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proc. of Winter Conference on Applications of Computer Vision*, 2024.

[65] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. of International Conference on Learning Representations*, 2022.

[66] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *Proc. of Computer Vision and Pattern Recognition*, 2023.

[67] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[68] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025.

[69] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention*, 2015.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of Computer Vision and Pattern Recognition*, 2016.

[71] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proc. of Winter Conference on Applications of Computer Vision*, 2022.

[72] E. Levin and O. Fried, "Differential diffusion: Giving each pixel its strength," *arXiv preprint arXiv:2306.00950*, 2023.

[73] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, S. Liu, W. Berman, Y. Xu, and T. Wolf, "Diffusers: State-of-the-art diffusion models," 2022. [Online]. Available: https://github.com/huggingface/diffusers

[74] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input / output image pairs," in *Proc. of Computer Vision and Pattern Recognition*, 2011.

[75] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. of Computer Vision and Pattern Recognition*, 2018.

[76] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao *et al.*, "Segment Anything," in *Proc. of International Conference on Computer Vision*, 2023.

[77] H. Zhao, M. Li, Q. Hu, and X. Guo, "Reversible decoupling network for single image reflection removal," in *Proc. of Computer Vision and Pattern Recognition*, 2025.
[78] Z. Song, Z. Zhang, K. Zhang, W. Luo, Z. Fan, W. Ren, and J. Lu, "Robust single image reflection removal against adversarial attacks," in *Proc. of Computer Vision and Pattern Recognition*, 2023.
[79] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. of International Conference on Computer Vision*, 2017.
[80] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, pp. 800–801, 2008.
[81] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[82] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of Computer Vision and Pattern Recognition*, 2018.
[83] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. of International Conference on Computer Vision*, 2021.
[84] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. of Association for the Advancement of Artificial Intelligence*, 2023.
[85] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2022.
[86] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-VL technical report," *arXiv preprint arXiv:2502.13923*, 2025.



Zifeng Wang received the B.S. degree from Peking University in 2026 and is currently pursuing the M.S. degree in the School of Computer Science, Peking University. His research interests include low-level vision and computational photography.



Yuchen Hong received the Ph.D. degree from Peking University in 2025. His research interests include computational photography and computer vision.



Tianyi Xu is currently a Ph.D. student in the School of Computer Science of Peking University. He received the B.S. degree from Peking University in 2025. His research interests include low-level vision and computational photography.



Haofeng Zhong received the B.E. and M.E. degrees from Peking University in 2022 and 2025, respectively. He is currently a Ph.D. student at Peking University. His research interests include 3D and 4D generation.



Shuchen Weng is a research staff member at Beijing Academy of Artificial Intelligence. He received the Ph.D. degree from Peking University in 2024. His research interests include language-based visual content creation and manipulation.



Jinxiu Liang (Member, IEEE) received the B.E. and Ph.D. degrees from the South China University of Technology in 2016 and 2021, respectively. She was a Postdoctoral Fellow at Peking University from 2021 to 2025 and is currently a Researcher by Special Appointment at the National Institute of Informatics, Tokyo. Her research interests include computational photography and computer vision. Her paper received a Best Paper Runner-Up award at CVPR 2024. She serves as a reviewer for top-tier journals and conferences (e.g., IJCV, IEEE TIP, CVPR, ICCV, ECCV, NeurIPS, ICML), and was recognized as an Outstanding Reviewer by IJCV in 2023 and a Top Reviewer by NeurIPS in 2024.



Boxin Shi (Senior Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, the M.E. degree from Peking University, and the Ph.D. degree from the University of Tokyo in 2007, 2010, and 2013, respectively. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he conducted research at the MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, and National Institute of Advanced Industrial Science and Technology from 2013 to 2017. His papers received Best Paper Runner-Up awards at CVPR 2024 and ICCP 2015 and were selected as Best Paper candidates at ICCV 2015, 3DV 2026, and CVPR 2026. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV.

RefMover: Diffusion-based Single Image Reflection Removal with Language and Region Guidance (Supplementary Material)

Zifeng Wang†, Yuchen Hong†, Tianyi Xu, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, *Member, IEEE*, Boxin Shi‡, *Senior Member, IEEE*

APPENDIX A
EVALUATION IN THE 8-BIT AND RAW DOMAINS

Recently, Kee *et al.* [1] proposed a physically accurate reflection removal pipeline based on linear RAW data, resolving the physical additivity issue. We acknowledge the value of operating in the linear RAW domain. However, a significant *domain gap* exists between linear RAW sensor data and the non-linear, 8-bit sRGB images (e.g., JPEG) that dominate real-world applications, web sharing, and mobile photography. The standard Image Signal Processing (ISP) pipeline introduces complex, often proprietary, non-linear transformations (such as gamma compression, tone mapping, and local contrast adjustments), rendering the simple linear additive assumption mathematically invalid in the 8-bit domain.

To evaluate our semi-synthesis strategy on 8-bit images, we conduct a cross-domain experiment. We evaluate four distinct configurations on our real-world 8-bit test sets, using reference-based metrics including PSNR [2], SSIM [3], and LPIPS [4], as well as no-reference perceptual metrics including CLIPIQA [5], MANIQA [6], and MUSIQ [7]. First, we evaluate the method by Kee *et al.* [1]† directly using its native integration in Adobe Photoshop 2026, applying it to 8-bit images. Second, we explicitly adapt the algorithmic pipeline of Kee *et al.* [1]‡ to operate on 8-bit/JPEG inputs. Third, following Kee *et al.* [1], we synthesize a training set using public RAW datasets, such as MIT-Adobe FiveK [8], convert the synthesized images to 8-bit, and train the same *RefMover* architecture, denoted as *RefMover**. Finally, we evaluate our default *RefMover* trained on our 8-bit synthesized dataset, which features massive semantic diversity from SA-1B [9].

TABLE S1: Quantitative comparison between RAW-based methods and our 8-bit-based strategy. ↑ (↓) indicates larger (smaller) values are better. **Bold** indicates the best performance.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
Kee *et al.* [1]†	23.25	0.866	0.118	**0.498**	0.356	61.83
Kee *et al.* [1]‡	23.47	0.901	0.135	0.391	0.340	58.97
*RefMover**	26.77	0.900	0.081	0.477	0.376	62.61
RefMover (Ours)	**27.45**	**0.906**	**0.070**	0.482	**0.381**	**62.81**

† Native PS2026 application. ‡ Adaptation to 8-bit domain. * Training on RAW datasets and evaluating on 8-bit images.

As demonstrated by the results in Table S1, methods strictly operating with RAW-domain priors [1] († and ‡) show lower performance when transferred to the 8-bit domain, and generally underperform our 8-bit-trained model, especially in PSNR and LPIPS, as they are less effective in handling the non-linear distortions introduced by the ISP. Interestingly, *RefMover** trained on the RAW-synthesized (and subsequently 8-bit converted) dataset achieves competitive results, performing only slightly below our default model. This suggests that while RAW-based physical synthesis is theoretically sound, relying solely on it may be affected by the limited scene diversity of public RAW datasets (typically containing $\sim 10^4$ images). These datasets contain less semantic diversity than the 11-million SA-1B dataset [9] utilized in our 8-bit synthesis. These findings suggest that semantic diversity is also important for robust generalization in the evaluated setting.

APPENDIX B
8-BIT DATA SYNTHESIS AND COMPLEXITY STATISTICS

A. Detailed synthesis pipeline

To reduce the synthesis-to-reality gap under 8-bit JPEG conditions, we provide a physically motivated yet scalable synthesis pipeline. We use 10,243 unique real-world pure reflection images (COMRR-R) and composite them with SA-1B backgrounds [9], yielding 30,729 unique training scenes.

Z. Wang, Y. Hong, T. Xu, and B. Shi are with State Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University. Beijing 100871, China. Email: wzf1835851@stu.pku.edu.cn; yuchenhong.cn@gmail.com; photon@stu.pku.edu.cn; shiboxin@pku.edu.cn.

H. Zhong is with the Wangxuan Institute of Computer Technology, Peking University. Beijing 100871, China. Email: hfzhong@pku.edu.cn

S. Weng is with Beijing Academy of Artificial Intelligence, Beijing 100083, China. Email: scweng@baai.ac.cn.

J. Liang is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Chiyoda-ku 1018430, Japan. Email: cssherryliang@gmail.com.

This work was supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and Beijing Natural Science Foundation (Grant No. QY24037).

† Equal contribution. ‡ Corresponding author.

Unlike global additive blending, our synthesis models two key properties: globally near-uniform glass reflectivity and locally non-uniform residual variations. We therefore construct a reflectivity field as a global base term plus a low-frequency local residual, and perform composition in a pseudo-linear domain via pre/post gamma transforms.

Algorithm S1 Proposed 8-bit JPEG-oriented Data Synthesis

Input: Transmission image $\mathbf{T} \in$ SA-1B, Reflection image $\mathbf{R} \in$ ComRR-R
// 1. Overexposure Filtering on Transmission Layer
if count$(\mathbf{T} > 235)$/total_pixels$(\mathbf{T}) > 0.10$ **then**
  Discard $\mathbf{T}$ and re-sample
end if
Normalize $\mathbf{T}$ and $\mathbf{R}$ to $[0, 1]$
// 2. Global-Uniform + Local-Residual Reflectivity
Sample global reflectivity $a_0 \sim \mathcal{U}(0.35, 0.65)$
Generate a zero-mean low-frequency map $\delta\mathbf{A}$ (Gaussian/Perlin), then normalize to $[-1, 1]$
Sample local amplitude $\beta \sim \mathcal{U}(0.05, 0.20)$
$\mathbf{A} = \text{clip}(a_0 + \beta \cdot \delta\mathbf{A}, 0, 1)$
// 3. Defocus and Ghosting
Sample $\sigma \sim \mathcal{U}(0.1, 2.0)$, $\Delta x, \Delta y \sim \mathcal{U}(-10, 10)$, $\rho \sim \mathcal{U}(0.15, 0.35)$
Randomly scale and horizontally flip $\mathbf{R}$
$\mathbf{R}_{blur} = \text{GaussianBlur}(\mathbf{R}, \sigma)$
$\mathbf{R}_{total} = (1 - \rho)\mathbf{R}_{blur} + \rho \cdot \text{Translate}(\mathbf{R}_{blur}, \Delta x, \Delta y)$
// 4. Independent Appearance Perturbation
Apply independent random color temperature / color jitter to $\mathbf{T}$ and $\mathbf{R}_{total}$
// 5. JPEG-oriented Nonlinear Imaging with Pre/Post Gamma
Set $\gamma_{in} = \gamma_{out} = 2.2$
$\mathbf{T}_{lin} = \mathbf{T}^{\gamma_{in}}, \quad \mathbf{R}_{lin} = \mathbf{R}_{total}^{\gamma_{in}}$
$\mathbf{M}_{lin} = \mathbf{T}_{lin} + (\mathbf{R}_{lin} \odot \mathbf{A})$
$\mathbf{M} = \text{clip}(\mathbf{M}_{lin}, 0, 1)^{1/\gamma_{out}}$
Optional JPEG re-encoding: quality $q \sim \mathcal{U}(70, 100)$
return $\mathbf{M}, \mathbf{T}, \mathbf{R}_{total} \odot \mathbf{A}$

B. Physical motivation: reflectivity field under Fresnel and thickness variations

In practical glass imaging, reflectivity is not purely random over space. For a single glass pane, the dominant reflectivity is usually globally smooth and approximately uniform at the image level, while local deviations arise from thickness non-uniformity, slight surface contamination, micro-geometry, and viewpoint-dependent Fresnel modulation. This motivates our decomposition $\mathbf{A} = a_0 + \beta\delta\mathbf{A}$: a_0 captures the pane-level global reflectivity, and $\beta\delta\mathbf{A}$ models localized residual variation. Such a design avoids unrealistic large-scale random alpha swings while still preserving local heterogeneity that frequently appears in real captures.

This formulation also improves learning behavior. Instead of relying on simple global color subtraction, *RefMover* is encouraged to detect localized reflection statistics and textures under spatially varying intensity, which leads to better robustness in unconstrained real-world environments.

C. Semantic diversity and model generalization

A key reason for adopting our 8-bit semi-synthetic strategy over purely RAW-based datasets (*e.g.*, FiveK [8]) is semantic coverage. Although RAW datasets offer stronger physical linearity, they typically contain only around 5,000 unique scenes. For our approximately 860M-parameter diffusion restorer, broad scene diversity is as important as photometric fidelity. Our dataset provides substantially richer semantic combinations, reducing overfitting to specific transmission-reflection pairs and improving disentanglement of heavily overlapped textures.

D. Complexity statistics of the held-out ComRR-Q *test set*

To quantitatively characterize the difficulty and diversity of our held-out ComRR-Q evaluation set, we introduce a multidimensional Dataset Complexity Radar (Fig. S1), comparing ComRR-Q against established benchmarks including SIR^2 [10], Nature [11], and Real [12]. Since our training corpus is a large-scale semi-synthetic set generated from a much broader source pool, we focus this analysis on the held-out test split, for which a compact statistical summary is more directly relevant to evaluation difficulty. Instead of relying solely on average metrics, which often mask the presence of severe localized artifacts, we evaluate dataset complexity across three complementary aspects. To ensure fair visualization across different physical units, all radar axes are transformed and anchored to fixed physical bounds such that a larger outward span strictly indicates higher degradation complexity.

First, we measure the general intensity of reflections (Average Degradation) via inverted PSNR ($-$PSNR), structural corruption ($1 -$ SSIM), and perceptual discrepancy (LPIPS). Second, to assess the datasets' inclusion of severe, challenging reflections (Extreme Hard Cases), we report the 10th percentile (p10) for PSNR and SSIM, and the 90th percentile (p90) for LPIPS. Finally, we evaluate Semantic Entanglement by measuring the CLIP [13] feature similarity between the corrupted image and the text prompt "picture with glass reflection," where a higher value indicates that the reflection heavily dominates the global image semantics. The detailed absolute raw values for these metrics are provided in Table S2.

The radar chart shows progressively larger complexity profiles from legacy benchmarks to our proposed dataset. The SIR^2 dataset forms the innermost polygon, occupying the lowest complexity tier across most axes, which is consistent with its lab-controlled capture setting where reflections are typically flatter and less destructive. Relative to prior benchmarks, the clearest distinction of ComRR-Q lies in its structural tail behavior and semantic entanglement rather than uniformly worse mean degradation on every average metric. As shown in Table S2, Real attains a lower mean PSNR, while ComRR-Q shows the lowest mean SSIM, the highest mean LPIPS by a very small margin, the most severe structural tail case ($\text{SSIM}_{\text{p10}} = 0.453$), and the highest CLIP mean. These results indicate that ComRR-Q is broadly comparable to the hardest existing benchmarks in mean-level degradation, but stands out

TABLE S2: **Raw statistical values of dataset complexity metrics.** To highlight the difficulty, for PSNR and SSIM, lower values indicate more severe degradation (↓); for LPIPS and CLIP similarity, higher values indicate greater discrepancy or stronger reflection semantics (↑). The most and second-most severe values are highlighted in **bold** and underlined, respectively. ComRR-Q notably demonstrates the most severe structural destruction in extreme cases ($\text{SSIM}_{\text{p10}} = 0.453$).

Dataset	Average Degradation (Mean)			Extreme Cases (Tail)			CLIP (mean) ↑	#Images
	PSNR ↓	SSIM ↓	LPIPS ↑	PSNR (p10) ↓	SSIM (p10) ↓	LPIPS (p90) ↑		
SIR^2	22.756	0.892	0.131	19.089	0.832	0.226	0.184	500
Real	**18.756**	0.722	0.238	14.834	0.547	**0.431**	0.204	20
Nature	20.324	0.771	0.170	**14.683**	0.654	0.272	0.187	20
ComRR-Q	19.731	**0.719**	**0.239**	15.199	**0.453**	0.362	**0.213**	201

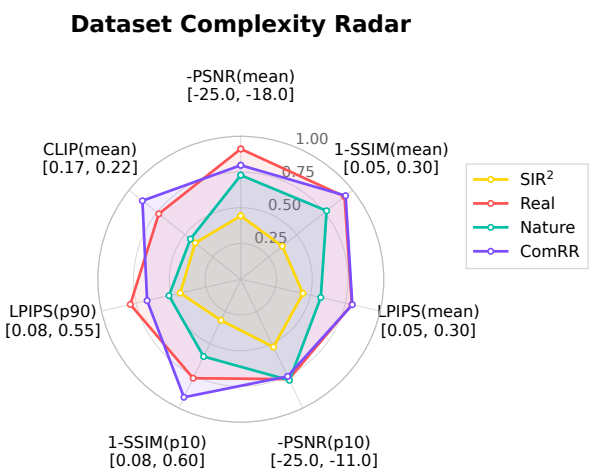


Fig. S1: **Dataset Complexity Radar.** Axes are transformed and bounded such that larger values (outer edges) indicate higher complexity. Compared with existing benchmarks, ComRR-Q shows more pronounced tail-case severity and stronger CLIP-based semantic entanglement, while its mean-level degradation is broadly comparable to the most challenging existing benchmarks.

more clearly in severe structural tail cases and reflection-related semantic entanglement. These statistics indicate that ComRR-Q contains challenging real-world cases, such as severe overexposure, thick-glass artifacts, and dense dirt, that complicate restoration.

Appendix C
Extended Ablation Studies and Robustness Analysis

A. *Empirical justification for negative-only prompting*

As discussed in the main text, inaccurate positive prompts describing the transmission layer can induce *semantic hallucinations*. We illustrate this in Fig. S2 using a real-world image of a cardboard box on a concrete ground, obstructed by hazy reflections.

We compare three prompting configurations using specific text descriptions for the transmission (positive) and reflection (negative) layers. When guided by a **Positive-Only** prompt (*e.g.*, "*A high-quality photo of a brown cardboard box... sitting on grass*"), the generative prior acts as an unconstrained repainting tool. To satisfy the textual condition, the model aggressively hallucinates the background, incorrectly transforming the original grey concrete on the left side of the box into a sandy, brown dirt surface. Adding an artifact-suppressing negative prompt (**Positive & Negative**), which is the standard practice in image generation, fails to resolve this issue. While the reflection is successfully suppressed, the severe semantic hallucination of the ground persists because the positive text condition still actively drives the scene generation.

Conversely, the **negative-only** setting avoids this prompt-induced background modification in this example. By using only an accurate negative prompt describing the artifacts (*e.g.*, "*hazy white glare, smoky glass reflection...*") and leaving the positive prompt empty, we steer the generation away from the artifact semantics without forcing background synthesis. As shown in the top right result, the model effectively strips away the reflections while faithfully preserving the true grey concrete background and the box's original physical structure, remaining closer to the ground truth.

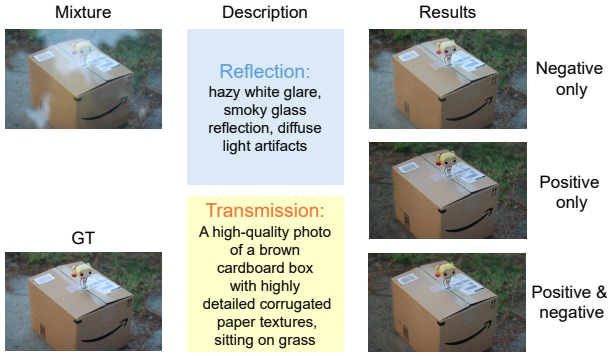


Fig. S2: Visual ablation of prompting strategies. (Middle Right) The inaccurate positive-only prompt leads the model to semantically hallucinate the background, incorrectly repainting the grey concrete into brown dirt/sand. (Bottom Right) Using both Positive and Negative prompts removes the reflection but still suffers from the same background hallucination. (Top Right) Our negative-only strategy removes the hazy glare while more faithfully preserving the true physical background.

B. *Robustness to coarse user guidance*

In practical interactive applications, users often provide rapid, coarse inputs rather than time-consuming, pixel-level segmentation masks. We evaluate the robustness of *RefMover* against varying levels of spatial precision, as illustrated in Fig. S3.

We benchmark the restoration performance under three distinct interaction settings: (i) a **precise mask**, (ii) a loose

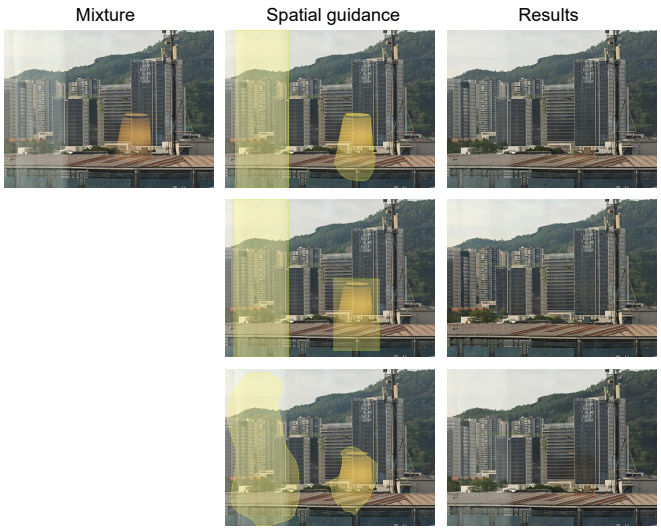


Fig. S3: Robustness to spatial guidance precision. From top to bottom: precise masks, coarse bounding boxes, and casual scribbles. *RefMover* consistently removes reflections while maintaining smooth structural and color transitions at the guidance boundaries, illustrating its applicability for practical interactive use.



Fig. S4: Visual ablation on CFG scales. Increasing the text guidance scale (γ_c) leads to relatively minor visual changes, as the RRM module already provides strong structural constraints. In contrast, increasing the image guidance scale (γ_y) makes the output more strongly follow the target image condition and leads to more thorough reflection removal. Overall, the visual results remain stable across a reasonable range of values.

bounding box that covers substantial non-reflection areas, and (iii) **casual scribbles** that only partially cover the reflection artifacts. As shown in the *Results* column, *RefMover* shows consistent results across all three scenarios. Despite the significant inclusion of clean background pixels within the coarse bounding box and the sparse coverage of the scribbles, the model restricts the restoration to the physical reflection layers with limited visible boundary artifacts or color inconsistencies at the mask edges. This spatial robustness supports a practical and user-friendly experience in real-world deployment.

TABLE S3: Ablation on sampling steps (evaluated on 741 test images). The performance is highly stable across different steps, demonstrating the efficiency of our model even at lower step counts.

Steps	SSIM↑	PSNR↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
1	0.904	27.264	0.074	0.480	0.379	62.717
3	0.906	27.448	0.070	0.482	0.381	62.812
7	0.905	27.423	0.070	0.482	0.380	62.779
14	0.906	27.418	0.070	0.481	0.380	62.701
28	0.903	27.341	0.071	0.485	0.381	62.752

C. Sensitivity to inference hyperparameters

We further evaluate the sensitivity of our framework to Classifier-Free Guidance (CFG) scales and the number of sampling steps. As visualized in Fig. S4, we explore the effects of varying the text guidance scale (γ_c) and the image guidance scale (γ_y). As γ_c increases, the visual differences remain relatively minor; this is expected because the text condition in our setting mainly acts as a negative constraint. In contrast, increasing γ_y strengthens adherence to the target image condition and leads to more thorough reflection removal. Overall, the visual results remain stable across a reasonable range of scales.

Furthermore, we analyze the impact of sampling steps on quantitative performance averaged over 741 test images, as summarized in Table S3. We evaluate the results using standard reference metrics including SSIM [3], PSNR, and LPIPS [4], as well as no-reference perceptual metrics such as CLIPIQA [5], MANIQA [6], and MUSIQ [7]. The model achieves competitive reference metrics in as few as 3 steps, indicating efficient inference. Importantly, the no-reference perceptual metrics do not show a uniform optimum at 28 steps: CLIPIQA is highest at 28 steps, MANIQA is effectively tied between 3 and 28 steps up to rounding, and MUSIQ is highest at 3 steps. Since 3 steps already provide the best or tied-best overall trade-off while being substantially faster, we adopt 3 sampling steps as the default setting for all other experiments.

D. Residual refinement objective

The RRM is optimized using the following composite objective:

$$\mathcal{L}_{\text{RRM}} = \mathcal{L}_{\text{pix}} + 0.1\mathcal{L}_{\text{perc}} + 0.1\mathcal{L}_{\text{edge}} + 0.1\mathcal{L}_{\text{SSIM}}. \quad (1)$$

Here, $\mathcal{L}_{\text{pix}}$ is the pixel-wise MSE loss, while $\mathcal{L}_{\text{perc}}$, $\mathcal{L}_{\text{edge}}$, and $\mathcal{L}_{\text{SSIM}}$ denote the VGG-based perceptual loss, Sobel edge loss, and SSIM loss, respectively. The RRM is trained for 50 epochs on 768×768 crops with a batch size of 64 using AdamW and a cosine learning-rate schedule.

APPENDIX D
COMPUTATIONAL COMPLEXITY AND DEPLOYMENT FEASIBILITY

We compare the computational overhead of *RefMover* with representative SOTA methods in Table S4. While diffusion-based models inherently possess larger parameter counts and higher FLOPs than lightweight CNNs (*e.g.*, ZF24) or Transformer-based architectures (*e.g.*, DSIT), our streamlined

TABLE S4: Comparison of model complexity and inference time (measured on a single NVIDIA RTX PRO 6000 GPU). For *RefMover*, the parameters include both the U-Net (865.9M) and the RRM (0.65M).

Method	Params (M)	FLOPs (G)	Time/Image (s)
ZF24 [15]	17.2	61.0	0.041
DSRNet [16]	124.6	1240.0	0.133
RDNet [17]	315.9	695.0	0.086
DSIT [18]	327.0	5688.0	0.495
RefMover (Ours)	866.6	29005.4	0.632

condition injection mechanism and the adoption of DPM-Solver++ [14] reduce the inference cost. Specifically, *RefMover* processes an image in approximately 0.63 seconds on a single NVIDIA RTX PRO 6000 GPU. Given its heavier computational footprint, we view this result as supporting practical high-fidelity interactive editing and cloud-assisted deployment, rather than direct on-device mobile execution.

APPENDIX E
EXTENDED COMPARISON WITH STATE-OF-THE-ART METHODS

We compare *RefMover* with specialized Single Image Reflection Removal (SIRR) algorithms, commercial software, and recent generative foundation models. The comparison characterizes the behavior of task-specific restoration models and general-purpose image editors under the reflection removal setting, especially in terms of pixel-aligned fidelity and background preservation.

A. Qualitative evaluation

We conduct qualitative comparisons on the REAL, SIR^2, and COMRR-V datasets, as visualized in Fig. S5. We benchmark *RefMover* against specialized SIRR algorithms, including ZF24 [15], DSIT [18], DSRNet [16], and RDNet [17], as well as commercial and generative image-editing tools, including Adobe Photoshop 2026 (PS2026), Nano Banana 2 [19] (NB2), and GPT Image 2 [20].

Comparison with specialized SIRR methods. As shown in Fig. S5, previous specialized methods often struggle with either incomplete reflection suppression or over-smoothing of the transmission layer. For instance, in the REAL examples, methods such as ZF24 and DSIT leave visible hazy reflections. In the SIR^2 examples, DSRNet and RDNet tend to blur structural details while suppressing reflections. These results indicate that purely blind restoration models can still be limited when the reflection and transmission layers are strongly entangled.

Comparison with generative foundation models. For Nano Banana 2 and GPT Image 2, we use the same concise user-style instruction:

"remove glass reflection in this photo"

Only the reflection-contaminated input image and this text instruction are provided. No ground-truth transmission image, manual mask, image-specific prompt, manual selection, or post-processing is used, except resizing to the original resolution when required for metric computation.

As shown in Fig. S5 and Fig. S6, the generative foundation models can remove visible reflections and produce visually clean images, confirming their strong perceptual editing capability. However, their unconstrained editing behavior may alter the underlying scene. In the examples of Fig. S6, Nano Banana 2 and GPT Image 2 modify the sky, illumination, color tone, and local background structures, producing plausible but non-identical transmission images. Such behavior is acceptable for perceptual image editing, but can be penalized under paired restoration protocols that require pixel-aligned faithfulness to the ground truth.

Behavior of *RefMover*. In contrast, *RefMover* is designed for reflection removal with language and region guidance while preserving the original transmission scene. It better maintains the input-aligned background structure, color appearance, and camera geometry, and avoids large content rewriting in non-reflection regions. For COMRR-V, ground-truth images are not available because it consists of unconstrained real-world captures; therefore, these examples are used for qualitative inspection only.

B. Diagnostic quantitative comparison with generative foundation models

We further conduct a diagnostic quantitative comparison with Nano Banana 2 [19] and GPT Image 2 [20]. We evaluate them on two public paired real-world benchmarks, REAL and NATURE, which contain 40 images in total and provide ground-truth transmission images for reference-based evaluation.

This experiment is not intended to represent a fully prompt-optimized upper bound of these commercial models. Instead, it evaluates their off-the-shelf zero-shot behavior under a fixed user-style instruction. For both generative baselines, the exact prompt is:

"remove glass reflection in this photo"

We generate one output per input without manual cherry-picking. Output coverage is reported over all 40 inputs. For one NB2 request and two GPT Image 2 requests, the corresponding API services declined to return an output. To ensure a matched quantitative comparison, all six metrics are computed on the common subset of 37 images for which all three methods produced a valid output.

As summarized in Table S5, on the common valid subset, Nano Banana 2 and GPT Image 2 obtain higher no-reference perceptual scores, reflecting their strong image synthesis and aesthetic priors. However, their PSNR and SSIM are lower and their LPIPS is higher than *RefMover*. This supports the qualitative observation in Fig. S6: general-purpose image editors can produce visually plausible reflection-free images, but may modify non-reflection regions, synthesize plausible but non-identical textures, or deviate from the pixel-aligned ground truth. *RefMover* shows stronger reference-based fidelity on this paired diagnostic subset.

C. Relation to NTIRE 2026 SIRR

The NTIRE 2026 SIRR report [21] provides a complementary perspective on reflection removal in the wild. It

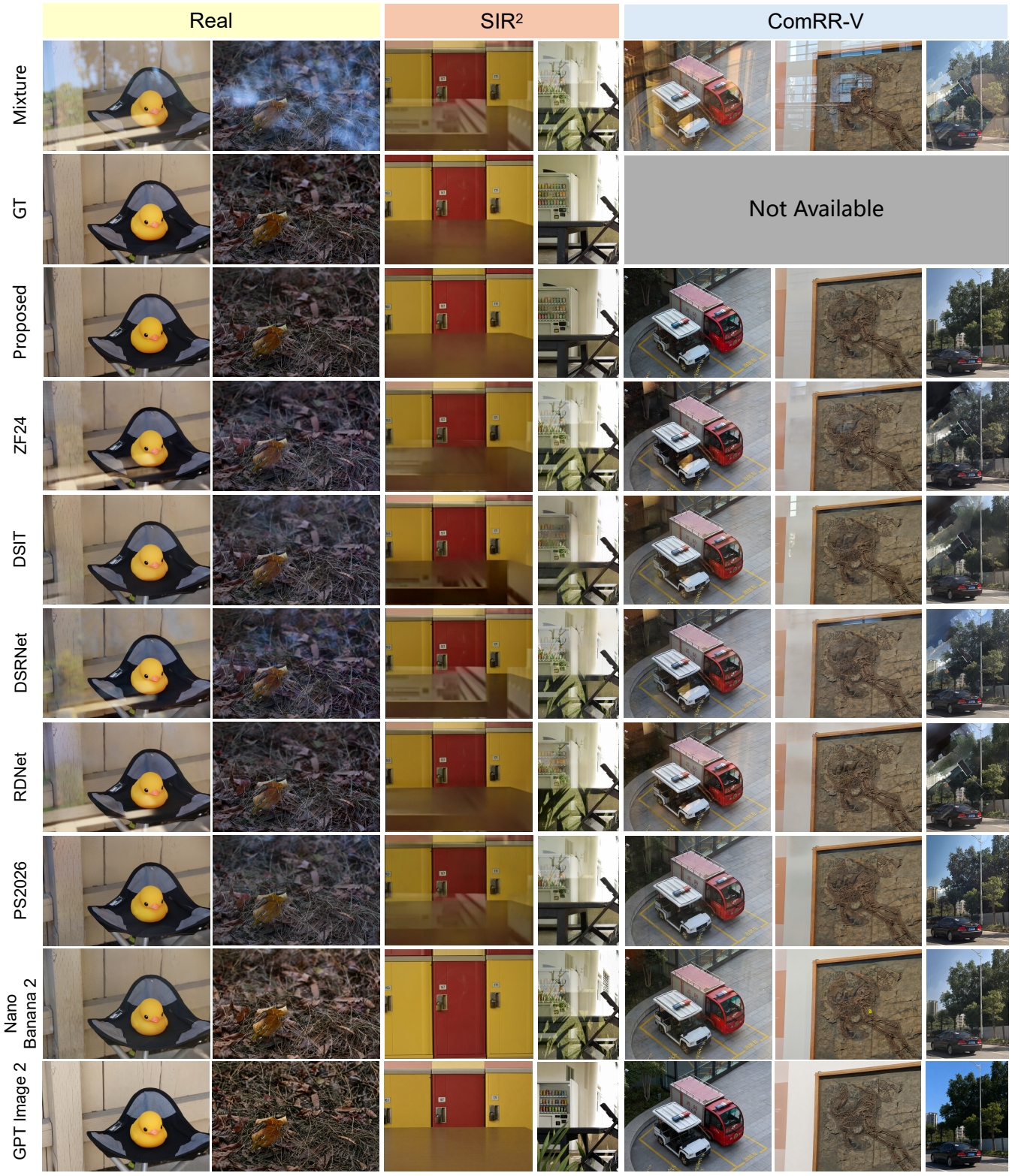


Fig. S5: Qualitative comparison on challenging real-world cases. Specialized SIRR methods may leave residual reflection artifacts or over-smooth the transmission layer, while general-purpose generative editors can produce visually plausible reflection-free images but may alter the underlying scene content. In contrast, *RefMover* better preserves the input-aligned transmission structures while suppressing reflection artifacts.

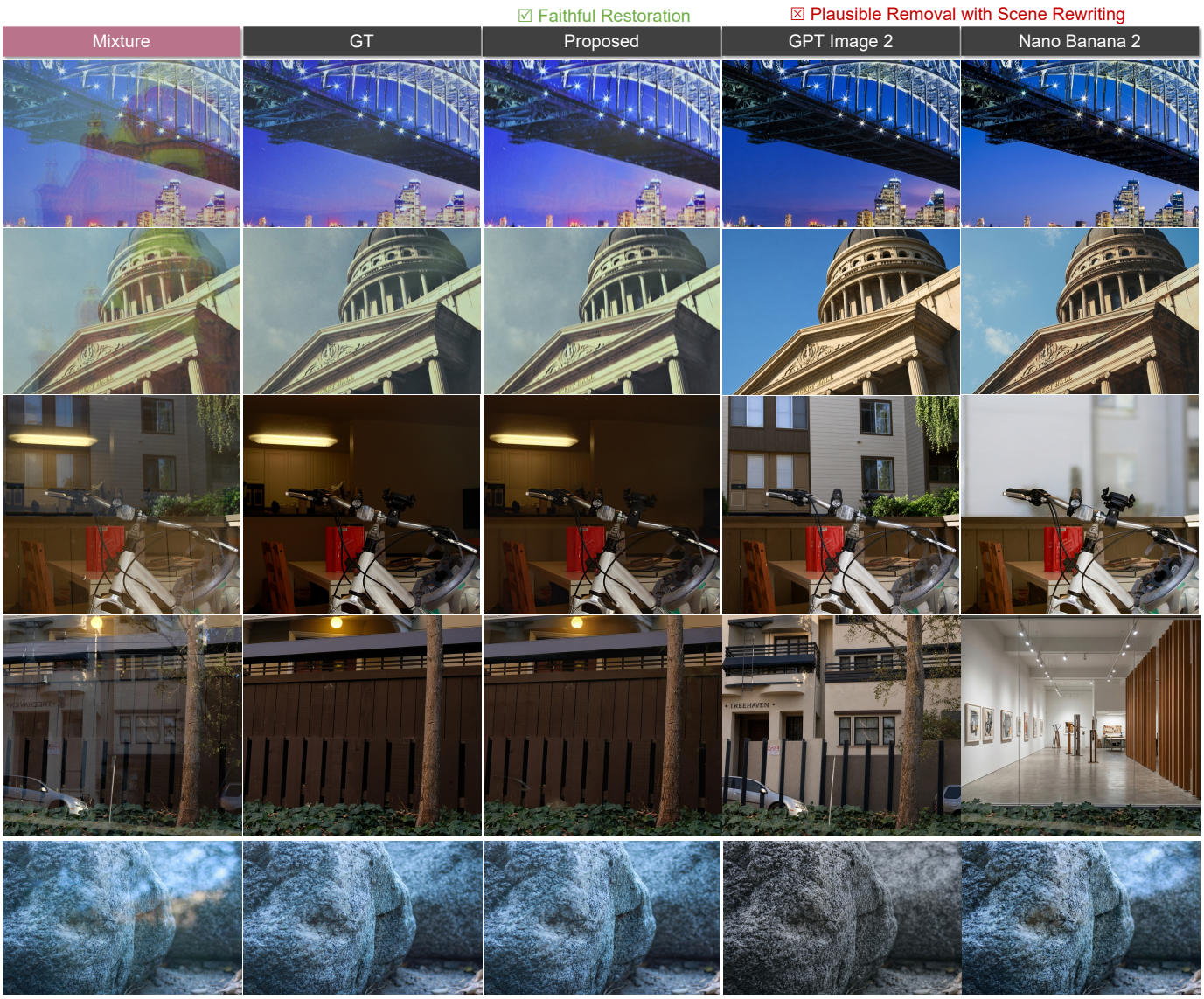


Fig. S6: Focused comparison with generative foundation models. Nano Banana 2 and GPT Image 2 use the fixed instruction, "remove glass reflection in this photo," and can remove visible reflections and produce visually clean images. However, they may also rewrite non-reflection regions, including the sky, illumination, color tone, and background structures. *RefMover* better preserves the pixel-aligned transmission scene and remains closer to the ground truth, which is critical for image restoration rather than unconstrained image editing.

shows that recent generative image models, including the Nano Banana series, can achieve strong perceptual restoration quality in challenging real-world cases. This observation is consistent with our qualitative findings that generative editors can produce visually clean reflection-free images.

The evaluation emphasis, however, is different. NTIRE 2026 focuses on in-the-wild perceptual quality and subjective assessment, while our paired benchmark evaluates pixel-aligned fidelity against ground-truth transmission images. Under such paired restoration protocols, visually plausible edits may receive lower PSNR and SSIM and higher LPIPS if they alter non-reflection regions, change color or illumination, or synthesize plausible but non-identical textures. Therefore, our conclusion is not that generative foundation models cannot understand reflections. Rather, our results indicate that under a paired image restoration protocol, their unconstrained editing behavior can trade pixel-level faithfulness for perceptual plausibility.

Appendix F
Limitations and Failure Cases

While *RefMover* demonstrates robust performance across diverse scenarios, it still encounters challenges in certain extreme boundary cases. Fig. S7 illustrates two typical failure modes. First, when the high-frequency textures of the transmission layer are severely polluted or densely occluded by reflections, the model struggles to recover the authentic high-frequency details, occasionally resulting in over-smoothed or

TABLE S5: Diagnostic quantitative comparison with generative foundation models on two public paired real-world benchmarks, REAL and NATURE. All six metrics are computed on the common valid subset of $N = 37$ images. Output coverage is reported over the complete 40-image set. All generative baselines use the same prompt, "remove glass reflection in this photo."

Method	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
NB2 [19]	18.86	0.687	0.212	0.630	0.491	70.50
GPT Image 2 [20]	17.60	0.599	0.262	**0.662**	**0.566**	**71.53**
RefMover	**24.34**	**0.834**	**0.129**	0.490	0.414	62.64

Note: NB2, GPT Image 2, and *RefMover* returned usable outputs for 39/40, 38/40, and 40/40 inputs, respectively. For the remaining requests, the corresponding API services declined to return an output. Because the missing samples do not overlap, all six metrics are computed on the common valid subset of $N = 37$ images.

corrupted local structures. Second, despite the use of explicit text descriptions to guide the restoration, the model's strong training priors for reflection removal can sometimes overpower the textual prompts. This can lead to the unintended elimination of genuine non-reflection objects that share visual characteristics with glares or glass artifacts.

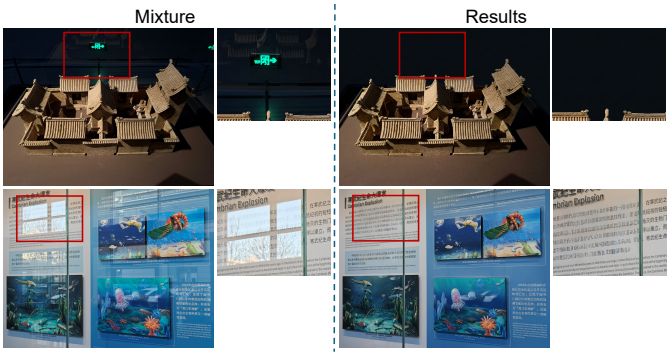


Fig. S7: Failure cases. The model may struggle when (Top) the high-frequency textures of the background are severely polluted by reflections, making detail recovery difficult, or (Bottom) the strong reflection-removal prior unintentionally eliminates genuine non-reflection objects, despite text guidance.

To address these limitations, future work could explore incorporating explicit frequency-domain structural guidance to preserve fine details, or leveraging more advanced vision-language alignment techniques to strictly enforce semantic preservation during the diffusion process.

REFERENCES

[1] E. Kee, A. Pikielny, K. Blackburn-Matzen, and M. Levoy, "Removing reflections from raw photos," in *Proc. of Computer Vision and Pattern Recognition*, 2025.
[2] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, pp. 800–801, 2008.
[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of Computer Vision and Pattern Recognition*, 2018.
[5] J. Wang, K. C. Chan, and C. C. Loy, "Exploring CLIP for assessing the look and feel of images," in *Proc. of Association for the Advancement of Artificial Intelligence*, 2023.
[6] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2022.
[7] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. of International Conference on Computer Vision*, 2021.
[8] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input / output image pairs," in *Proc. of Computer Vision and Pattern Recognition*, 2011.
[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao *et al.*, "Segment Anything," in *Proc. of International Conference on Computer Vision*, 2023.
[10] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. of International Conference on Computer Vision*, 2017.
[11] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proc. of Computer Vision and Pattern Recognition*, 2020.
[12] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. of Computer Vision and Pattern Recognition*, 2018.
[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of International Conference on Machine Learning*, 2021.
[14] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025.
[15] Y. Zhu, X. Fu, P.-T. Jiang, H. Zhang, Q. Sun, J. Chen, Z.-J. Zha, and B. Li, "Revisiting single image reflection removal in the wild," in *Proc. of Computer Vision and Pattern Recognition*, 2024.
[16] Q. Hu and X. Guo, "Single image reflection separation via component synergy," in *Proc. of International Conference on Computer Vision*, 2023.
[17] H. Zhao, M. Li, Q. Hu, and X. Guo, "Reversible decoupling network for single image reflection removal," in *Proc. of Computer Vision and Pattern Recognition*, 2025.
[18] Q. Hu, H. Wang, and X. Guo, "Single image reflection separation via dual-stream interactive transformers," in *Proc. of Advances in Neural Information Processing Systems*, 2024.
[19] Google Inc., "Gemini 3.1 Flash Image (Nano Banana 2)," https://aistudio.google.com/models/gemini-3-1-flash-image, 2026, google AI Studio model page, accessed 2026-03-14.
[20] OpenAI, "GPT Image 2 Model," https://developers.openai.com/api/docs/models/gpt-image-2, 2026, openAI API documentation, accessed 2026-06-09.
[21] J. Cai, K. Yang, Z. Li, F.-A. Vasluianu, R. Timofte *et al.*, "NTIRE 2026 Challenge on Single Image Reflection Removal in the Wild: Datasets, Results, and Methods," in *Proc. of Computer Vision and Pattern Recognition Workshops*, 2026.