FUMO: Prior-Modulated Diffusion for Single Image Reflection Removal

Telang Xu[1*], Chaoyang Zhang[2,3*], Guangtao Zhai[1], and Xiaohong Liu[1,2†]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Xi'an Jiaotong University, Xi'an, China
{luciousdesmon, xiaohongliu, zhaiguangtao}@sjtu.edu.cn
{zcyxjtu65}@gmail.com
*Equal Contribution †Corresponding Author

Abstract. Single image reflection removal (SIRR) is challenging in real scenes, where reflection strength varies spatially and reflection patterns are tightly entangled with transmission structures. This paper presents a diffusion model with prior modulation framework (FUMO) that introduces explicit priors for spatially adaptive conditioning and structurally faithful restoration. Two priors are extracted directly from the mixed image, an intensity prior that estimates spatial reflection severity and a high-frequency prior that captures detail-sensitive responses via multi-scale residual aggregation. We propose a coarse-to-fine training paradigm. In the first stage, these cues are combined to gate the conditional residual injections, focusing the conditioning on regions that are both reflection-dominant and structure-sensitive. In the second stage, a fine-grained refinement network corrects local misalignment and sharpens fine details in the image space. Experiments conducted on both standard benchmarks and challenging images in the wild demonstrate competitive quantitative results and consistently improved perceptual quality. The code is released at https://github.com/Lucious-Desmon/FUMO.

Keywords: Reflection removal · Conditional generation · Diffusion model

1 Introduction

Images captured through glass or other transparent media often contain unwanted reflections that obscure the underlying scene [46, 62]. Such artifacts degrade visual quality and hinder downstream vision applications [36, 47]. Reflection removal aims to suppress reflections and recover a clear transmission image from a mixture input. The task in real-world settings remains challenging due to severe layer entanglement and large spatial variations in reflection appearance [10, 17, 74].

Methods [1, 11, 25, 44, 45, 61, 63] leveraging multiple images depend on specialized acquisition and are less applicable to casual photography or internet images.

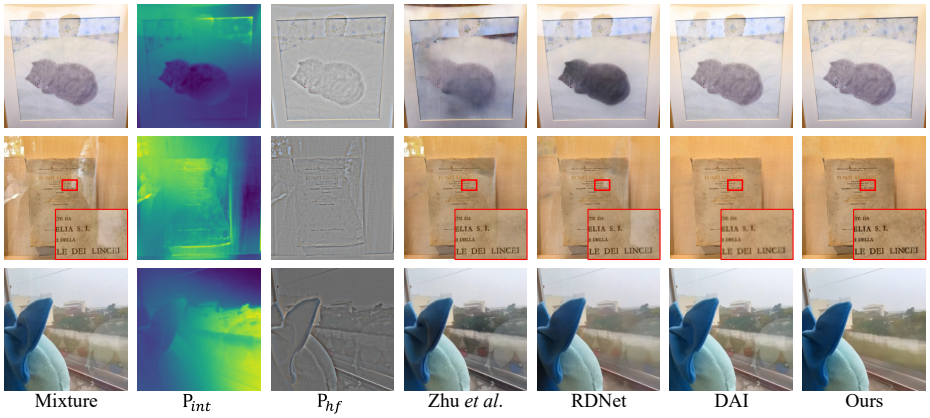

Fig. 1: Failure-mode visualization on in-the-wild reflection mixtures. Three representative real-world mixed images are shown together with two priors. Qualitative comparisons with representative SOTA methods illustrate common challenges in the wild, including incomplete reflection suppression, color inconsistency, and loss of fine details. Red rectangles highlight regions for closer inspection.

In the single-image task, traditional optimization methods [2,24,26,43,60,70,71] impose hand-crafted priors to decompose layers, yet they are often brittle when assumptions are violated and computationally expensive. Learning-based approaches [10,14,16,17,27,28,55,69,71,74] improve performance by learning deep priors from data, but robust reflection removal in the wild remains challenging. More recently, diffusion models [14,15,31,51,58,73] have shown strong potential for image restoration, which calls for appropriate guidance conditions to avoid undesired content drift and structural distortions. Overall, reflection removal faces an intrinsic trade-off, the tension between aggressive reflection suppression and faithful preservation of scene structure under large reflection variability. In practice, this trade-off manifests as three common issues on real mixtures, including incomplete reflection suppression, color inconsistency, and loss of fine details, as illustrated in Fig. 1.

To address this, we introduce an explicit spatial modulation mechanism that suppresses reflection strength spatially and preserves scene structure. We derive two complementary priors from the mixed image, a vision-language model (VLM) [3, 29, 52, 54] based reflection intensity prior and a high-frequency prior by multi-scale residual decomposition. Examples of the priors P_{int} and P_{hf} are shown in Fig. 1. The intensity prior captures both region-level reflection analysis and global localization of reflection phenomena, thereby indicating where reflection removal should be stronger and to what extent. The high-frequency prior, in turn, provides a detailed sensitive signal that helps the model remain attentive to local structures during restoration. We integrate these priors into a spatial gate for selective modulation of the conditional residual injections.

Building on these guided priors, we propose **FUMO**, a dif**FU**sion model with prior **MO**dulation framework for SIRR, which adopts a coarse-to-fine design. In the first stage, a conditional diffusion model [42, 66] is guided by the proposed gate to aggressively remove reflections, driving the low frequency appearance toward the underlying transmission content. This stage delivers effective reflection removal, but strong suppression can still introduce geometric deviations or local inconsistencies. We therefore follow with a fine-grained refinement module, combined with the two priors, to correct distortions and restore coherent structures and details. Experiments demonstrate that our approach achieves strong performance under diverse scenes.

The main contributions are summarized as follows:

- We introduce a dual-prior extraction pipeline for SIRR, supporting both reflection strength estimation and detail aware guidance.
- We propose a prior-modulated diffusion framework that uses the two priors to guide conditional residual injection in a coarse-to-fine restoration process.
- Extensive experiments on benchmarks and in-the-wild images demonstrate competitive quantitative performance and improved visual quality.

2 Related Work

2.1 Image Reflection Removal

Due to the complex superposition of the transmission and reflection layers, recovering the transmission from a single blended image is a mathematically ill-posed problem. An intuitive solution is Multi-Image Reflection Removal (MIRR) [1,11,25,44,45,61,63], which utilizes auxiliary information from multiple images, such as dynamic scenes [44], focused/defocused images [45], or polarization imaging [25,63]. These methods mitigate reflection interference through comparative analysis and can effectively alleviate the ill-posed nature of the problem. However, their reliance on specific equipment or capture conditions limits their practical applicability, especially when processing existing images on the Internet or for mobile devices [14].

In contrast, Single-Image Reflection Removal (SIRR) [62] aims to tackle this problem by exploring intrinsic priors from a single input. Some works relied on hand-crafted priors [2,24,26,43,60,70,71], such as edge annotations [26], ghosting cues [43], and sparsity assumptions [2]. With the development of deep learning, neural network-based methods [10, 14, 16, 17, 27, 28, 55, 69, 71, 74] have demonstrated superior performance and lower costs by autonomously learning implicit priors from the blended image. Dong *et al.* [10] design a reflection detection module to regress a probabilistic reflection confidence map. YTMT [16] enforces the predictions to communicate with each other. DSRNet [17] proposes DSFNet to initially extract transmission and reflection features. Zhu *et al.* [74] introduces MaxRF to explicitly indicate virtual objects. RDNet [69] employs a reversible encoder to secure valuable information. SIRR methods increasingly incorporate reflection localization and structure preserving designs to improve robustness

in real scenes. Following this direction, we leverage explicit reflection strength priors together with detail oriented guidance to improve spatial adaptivity and reconstruction fidelity.

2.2 Diffusion Models

In recent years, diffusion models [4, 13, 37, 38, 40–42] have achieved significant progress. DDPM [13] establishes a powerful new paradigm and has sparked extensive follow up research. DDIM [38] further improves sampling efficiency by enabling a more flexible generation procedure. Latent diffusion models (LDMs) [41, 42] conduct the diffusion process in a compact latent space and facilitate the integration of additional conditioning signals, which significantly broadens their applicability. In addition to U-Net denoisers, studies [4, 40] also explore diffusion transformer denoisers, further enriching architectural choices. These developments collectively strengthen the generative prior of diffusion models and provide flexible design choices for downstream tasks.

Generative models [9, 33–35, 56, 72] have shown strong potential in image and video enhancement by leveraging learned priors for real-world degradations. With the rise of diffusion models as a powerful generative paradigm, diffusion-based methods have also been adapted for image enhancement and restoration, demonstrating powerful capabilities in tasks such as low-light enhancement [19, 20, 73], dehazing [51], and super-resolution [31, 58]. In reflection removal, L-DiffER [14] first introduces diffusion models to this task, where previous predictions are used as conditions to steer an iterative denoising process. DAI [15] strengthens conditional injection with a ControlNet and further improves reconstruction with a refined decoder. PolarFree [63] leverages diffusion based modeling to produce reflection free imaging priors, which support subsequent reflection suppression. These advances suggest that diffusion models are promising for dereflection, and effective results often rely on appropriate conditioning and guidance to preserve faithful transmission content. Motivated by this insight, our framework adopts a diffusion backbone and injects priors through spatially adaptive modulation before fine-grained refinement.

3 Methods

The core of this work is to derive explicit guidance priors from the mixed image and use them to modulate diffusion-based reflection removal. The overall framework is illustrated in Fig. 3.

In this section, we first introduce two complementary priors in Sec. 3.1. These priors form a spatial gate that modulates conditional residual injections in the coarse diffusion stage, followed by a lightweight refinement module for geometric consistency and detail coherence. The architecture and training objectives are described in Sec. 3.2 and Sec. 3.3.

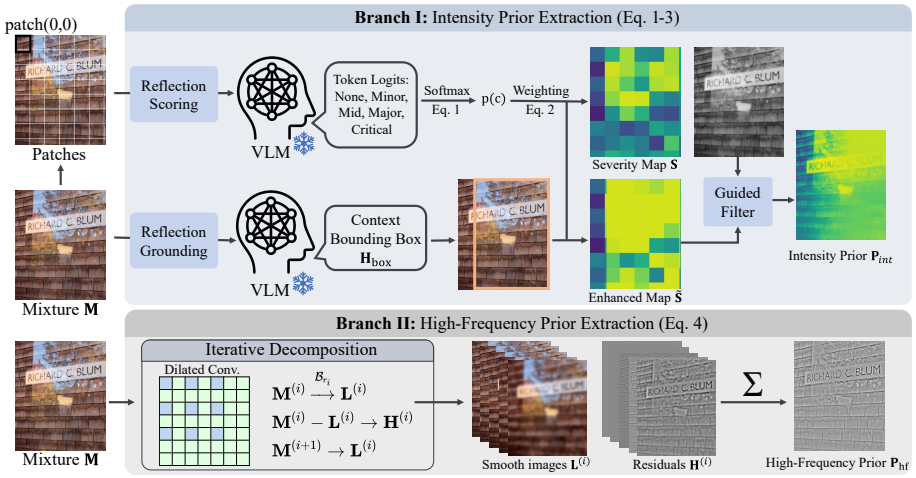

Fig. 2: The pipeline of dual prior extraction. In branch I, the mixed image $\mathbf{M}$ is divided into patches, and the intensity prior $\mathbf{P}_{\text{int}}$ is finally obtained through scoring and localization. In branch II, the mixed image $\mathbf{M}$ is iteratively decomposed and aggregated to yield the high-frequency prior $\mathbf{P}_{\text{hf}}$.

3.1 Dual Prior Extraction

The mixed image alone usually lacks sufficient cues to robustly separate reflections from transmission structures. To alleviate this limitation, we extract two complementary priors, a VLM-derived intensity prior for reflection severity and a high-frequency prior for detail-sensitive responses. The overall prior extraction pipeline is shown in Fig. 2.

Reflection Intensity Prior. Vision Language Models (VLMs) [3, 5, 29, 52, 54] provide a practical interface for extracting semantic cues from a mixed image. Studies such as [7,32,50,57,73] suggest that VLM-derived signals, including intermediate reasoning outputs, can benefit image analysis and processing. Building on this capability, we use a frozen VLM to estimate a reflection intensity prior that indicates where reflection interference is stronger and to what extent. Given a mixed image $\mathbf{M}$, the reflection intensity prior $\mathbf{P}_{\text{int}} \in [0,1]^{H\times W}$ is a pixel-level soft severity map that provides spatially grounded guidance for subsequent conditioning.

We start from patch-level severity scoring to capture local reflection evidence at a manageable granularity. Inspired by Q-ALIGN [57], we partition $\mathbf{M}$ into non-overlapping patches (with patch size a chosen adaptively according to image resolution), and query a VLM to assess reflection severity using a fixed ordinal set $\mathcal{C} = \{\texttt{None}, \texttt{Minor}, \texttt{Mid}, \texttt{Major}, \texttt{Critical}\}$. Instead of relying on free-form text outputs, we exploit the model's next-token logits and restrict the probability mass to $\mathcal{C}$, yielding a more stable confidence distribution over the predefined

levels:

$$p(c) = \frac{\exp(\ell(c)/\tau)}{\sum_{c' \in \mathcal{C}} \exp(\ell(c')/\tau)}, \quad c \in \mathcal{C}, \tag{1}$$

where $\ell(c)$ denotes the logit of the token corresponding to category c and τ is a temperature. Then a continuous severity score is obtained through an ordinal expectation with weights $w(c) \in \{1, 2, 3, 4, 5\}$:

$$s = \sum_{c \in \mathcal{C}} w(c)\, p(c). \tag{2}$$

We bring the patch scores back to the image plane to form a pixel-wise field $\mathbf{S}$ by assigning the same score to all pixels within each patch.

Patch-level scoring provides local estimates, but may overlook reflection regions that are visually prominent on the global scale. We further perform image-level analysis on the mixed image $\mathbf{M}$. Similarly to region localization in object detection, we prompt the VLM to identify areas dominated by reflections and return their bounding boxes. Then the corresponding regions in $\mathbf{S}$ are boosted with a multiplicative factor to obtain an enhanced map $\tilde{\mathbf{S}}$ (capped by a maximum value), which better captures large, coherent reflection patterns without altering the underlying patch scoring mechanism. On 100 real images with annotated reflection-dominant regions, this grounding step achieves 0.65 mIoU, serving as a coarse localization cue.

Subsequently, we densify $\tilde{\mathbf{S}}$ into a spatially coherent guidance map via an edge-aware guided filter [12]:

$$\mathbf{P}'_{\text{int}} = \text{GF}(\tilde{\mathbf{S}}; \mathbf{G}, r, \epsilon), \tag{3}$$

where the guide image $\mathbf{G}$ is obtained from $\mathbf{M}$ by converting to grayscale and applying a mild Gaussian pre-blur. This step removes block artifacts while encouraging intensity transitions to align with major image structures. Ultimately, we clamp the filtered response to a fixed range and linearly rescale it to $[0, 1]$ to obtain the final intensity prior $\mathbf{P}_{\text{int}}$, aligned with the input resolution.

High-Frequency Prior. The intensity prior provides spatial awareness of reflection strength. To complement it with local structural responses, we further extract a high-frequency prior from the same mixture input. Since it is extracted from the mixture $\mathbf{M}$, the prior contains high-frequency components from both the transmission and reflection layers. We therefore use it as a structure-sensitive guidance signal, which is combined with the intensity prior for spatial modulation.

Given the mixed image $\mathbf{M}$, we construct a multi-scale residual decomposition inspired by wavelet color correction [49] to compute the high-frequency prior $\mathbf{P}_{\text{hf}}$. Let $\mathcal{B}_r(\cdot)$ denote a channel-wise smoothing operator at scale r, implemented with dilated convolution [65]. Starting from $\mathbf{M}^{(0)} = \mathbf{M}$, we iteratively compute a smoothed image and its residual at increasing scales: $\mathbf{L}^{(i)} = \mathcal{B}_{r_i}(\mathbf{M}^{(i)})$, $\mathbf{H}^{(i)} =$

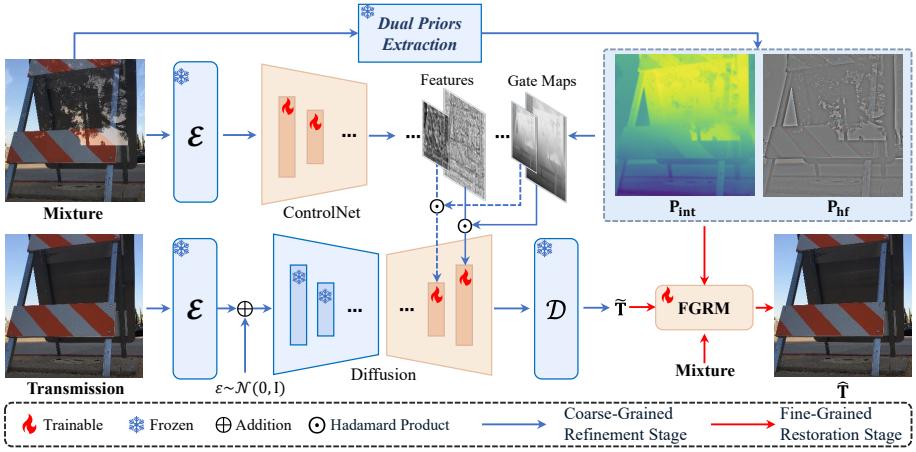

Fig. 3: The framework of the proposed **FUMO** method. Given a mixed image $\mathbf{M}$, we obtain two priors $\mathbf{P}_{\text{int}}$ and $\mathbf{P}_{\text{hf}}$ through dual-prior extraction. A diffusion-based backbone performs conditional denoising, where the extracted features and gates are injected through element-wise fusion operations to guide multi-scale feature aggregation. The decoder produces a coarse restoration, which is further refined by the trainable fine-grained refinement module (FGRM).

$\mathbf{M}^{(i)} - \mathbf{L}^{(i)}$, and update $\mathbf{M}^{(i+1)} = \mathbf{L}^{(i)}$, where $r_i = 2^i$ for $i = 0, \ldots, L-1$. The final high-frequency prior is obtained by accumulating residuals across scales:

$$\mathbf{P}_{\text{hf}} = \sum_{i=0}^{L-1} \mathbf{H}^{(i)}. \tag{4}$$

By aggregating residual components from fine to coarse scales, $\mathbf{P}_{\text{hf}}$ captures local structural responses while remaining simple and efficient to compute. We further clamp and linearly rescale the extracted response to $[0, 1]$ at the input resolution for stable conditioning. Together with the intensity prior, it forms a complementary pair of explicit signals for the subsequent gated modulation design.

3.2 Prior-modulated diffusion framework for SIRR

As illustrated in Fig. 3, FUMO follows a coarse-to-fine prior-modulated diffusion design, consisting of a coarse restoration stage and a fine-grained refinement module. The coarse stage adopts a ControlNet-conditioned [66] diffusion denoiser, where explicit priors are incorporated through a gated modulation mechanism for spatially adaptive conditioning. The refinement module further improves geometric consistency and restores fine structural details, producing the final transmission image.

Coarse-Grained Restoration Diffusion Model. The coarse stage targets strong reflection suppression to obtain an initial transmission estimate that is faithful in global appearance and low-frequency structure. We adopt a pretrained T2I diffusion model as the backbone, such as stable diffusion [42], which provides strong generative priors. The VAE encoder takes the mixed image $\mathbf{M}$ and transmission image $\mathbf{T}$ as input to generate latent tensors $z_{\mathbf{M}}$ and $z_{\mathbf{T}}$, respectively. ControlNet takes $z_{\mathbf{M}}$ as input to guide the SD model output. The diffusion denoiser μ_θ receives noisy input $z_t = \alpha_t z_{\mathbf{T}} + \sigma_t \epsilon$ (where t denotes diffusion step, and α_t and σ_t are determined by noise scheduler with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$), and the control signal $c_m = f_\phi(\mathbf{M})$ from ControlNet f_ϕ to predict the latent $z_0 = z_{\mathbf{T}}$. Following [59,64], we adopt a one-step denoising formulation for the coarse restoration stage.

Gated Modulation Module. The intensity prior $\mathbf{P}_{\text{int}}$ and the high-frequency prior $\mathbf{P}_{\text{hf}}$ provide complementary guidance for spatially adaptive restoration. $\mathbf{P}_{\text{int}}$ indicates where reflections are severe and thus require stronger conditional intervention, while $\mathbf{P}_{\text{hf}}$ highlights structure-sensitive regions where residual modulation should be more attentive to local details. We combine them into a spatial gate that strengthens conditioning in regions with both severe reflection and rich local structure.

Specifically, we define the gate map as

$$\mathbf{g} \;=\; \mathbf{1} \;+\; \beta\, \mathbf{P}_{\text{int}} \odot \mathbf{P}_{\text{hf}}, \tag{5}$$

where $\odot$ denotes element-wise multiplication and β controls the overall modulation strength. The additive identity term ensures that the mechanism reduces to standard conditioning when either guidance signal is weak, while the multiplicative interaction emphasizes their joint presence.

The gate is applied in the coarse stage to modulate the residual injection from the conditioning branch into the restoration network. Let $\mathbf{c}_m = \{\mathbf{c}_s\}$ denote the multi-scale residual tensors produced by ControlNet, where s indexes the injection scale. For each scale, we resize $\mathbf{g}$ to match the spatial resolution of $\mathbf{c}_s$ and perform element-wise modulation to generate the gated modulation condition signal $\widetilde{\mathbf{c}}_s$:

$$\widetilde{\mathbf{c}}_s \;=\; \text{clip}(\mathcal{I}_s(\mathbf{g}),\, 1,\, 1+\beta_{\max}) \odot \mathbf{c}_s, \tag{6}$$

where $\mathcal{I}_s(\cdot)$ denotes per-scale interpolation for alignment and clip$(\cdot)$ clamps the gate to a bounded range for stable injection. In practice, β is warmed up and then increased to $\beta_{\max}$ during training, so that the gated modulation is introduced progressively. We set $\beta_{\max} = 0.25$ and use a warmup ratio of 0.1 empirically.

Fine-Grained Refinement Module. Although the coarse stage provides strong reflection suppression, its aggressive restoration may introduce geometric drift or local structural inconsistency, especially in regions where transmission structures are tightly entangled with reflections. To improve geometric coherence

and recover fine details, we develop a fine-grained refinement module (FGRM) as a deterministic refiner operating in the image space.

The refinement module takes the coarse prediction together with the original mixture and the guidance signals as inputs. Let $\tilde{\mathbf{T}}$ denote the decoded transmission image obtained in the coarse stage. The refiner receives the channel-wise concatenation

$$\hat{\mathbf{T}} \;=\; R_{\phi}(\mathrm{concat}\Big(\mathbf{M},\, \tilde{\mathbf{T}},\, \mathbf{P}_{\mathrm{hf}},\, \mathbf{P}_{\mathrm{int}}\Big)), \tag{7}$$

where R_{ϕ} denotes the Fine-Grained Refinement Module. In practice, we adopt a U-Net but replace the standard activation with a lightweight channel gating operation that splits features into two halves and multiplies them element-wise, following the SimpleGate design in NAFNet [8].

3.3 Training Strategy

The coarse-to-fine design is optimized separately with stage-specific objectives. The coarse stage focuses on one-step latent restoration under guided and gated modulation, while the refinement stage operates deterministically in the image space to improve geometric consistency and recover fine structures.

Coarse-Grained Restoration Stage. In the backward denoising process, conventional diffusion models predict noise at randomly sampled step t, and the corresponding multi-step training loss is formulated as:

$$\mathcal{L}_{\mathrm{ldm}} \;=\; \big\|\epsilon - \mu_{\theta}(z_t, t, \widetilde{\mathbf{c}}_m)\big\|_2^2. \tag{8}$$

where $\widetilde{\mathbf{c}}_m$ is the control signal from ControlNet and gated modulation.

Following [15, 59, 64], we adopt one-step denoising for stable and efficient restoration. Instead of predicting noise, the network directly regresses a less perturbed latent. Specifically, for a target timestep t (with a smaller t indicating a cleaner latent), the model takes a maximally perturbed latent z_N as input and predicts z_t in a single forward pass:

$$\mathcal{L}_{\mathrm{coarse}} \;=\; \big\|z_t - \mu_{\theta}(z_N, t, \widetilde{\mathbf{c}}_m)\big\|_2^2. \tag{9}$$

During training, t is sampled uniformly and N is set to 1000. At inference time, $t = 0$ is set to obtain $\hat{z}_{\mathbf{T}} = \mu_{\theta}(z_N, 0, \widetilde{\mathbf{c}}_m)$ with $z_N \sim \mathcal{N}(0, I)$, and $\hat{z}_{\mathbf{T}}$ is decoded to produce the restored transmission image.

In this training stage, only the ControlNet and the upsampling blocks of denoising U-Net are optimized. The VAE and the remaining parameters of U-Net are kept frozen to preserve the pretrained generative prior and stabilize optimization.

Fine-Grained Refinement Stage. The Fine-Grained Refinement Module refines the coarse prediction in the image space. Denoting the refined output by $\hat{\mathbf{T}}$ and the ground-truth transmission by $\mathbf{T}$, we employ a composite objective that balances pixel fidelity, perceptual similarity and geometric consistency. In this training stage, only the Fine-Grained Refinement Module is updated.

A pixel-level reconstruction loss constrains the global color and luminance:

$$\mathcal{L}_{\text{pix}} \;=\; \|\hat{\mathbf{T}} - \mathbf{T}\|_1. \tag{10}$$

Then we include a perceptual loss based on LPIPS [67] to discourage oversmoothing and to better preserve visually salient structures:

$$\mathcal{L}_{\text{perc}} \;=\; \mathcal{L}_{\text{LPIPS}}(\hat{\mathbf{T}}, \mathbf{T}) = \sum_i w_i \|\phi_i(\hat{\mathbf{T}}) - \phi_i(\mathbf{T})\|_2, \tag{11}$$

where $\phi_i(\cdot)$ denotes the feature map extracted from the i-th layer of AlexNet [23]. Finally, to explicitly promote edge alignment and mitigate local misalignment artifacts, we impose an edge-aware gradient loss and measure the discrepancy of both horizontal and vertical derivatives:

$$\mathcal{L}_{\text{grad}} = \|\nabla_x \hat{\mathbf{T}} - \nabla_x \mathbf{T}\|_1 + \|\nabla_y \hat{\mathbf{T}} - \nabla_y \mathbf{T}\|_1 \tag{12}$$

where $\nabla(\cdot)$ denotes image gradients computed by fixed Sobel filters.

Gathering all the loss terms yields the refinement objective as:

$$\mathcal{L}_{\text{refine}} \;=\; \lambda_{\text{pix}}\,\mathcal{L}_{\text{pix}} \;+\; \lambda_{\text{perc}}\,\mathcal{L}_{\text{perc}} \;+\; \lambda_{\text{grad}}\,\mathcal{L}_{\text{grad}}, \tag{13}$$

where the weights $\lambda_{\text{pix}} = 0.5$, $\lambda_{\text{perc}} = 0.25$, and $\lambda_{\text{grad}} = 0.25$ are set empirically.

4 Experiments

4.1 Implementation Details

We implement the proposed method in PyTorch [39]. All experiments are conducted on 4 NVIDIA GeForce RTX 4090 GPUs with a batch size of 1. The two stages are trained separately, with 100k steps and 10k steps respectively. In the coarse stage, we initialize the U-Net and ControlNet with pretrained weights of Stable Diffusion V2.1 [42], and update the weights using AdamW [22] with a learning rate of 5×10^{-5}. In the refinement stage, we freeze the coarse restoration part and train the fine-grained refinement module using AdamW with a learning rate of 1×10^{-4}. At 1K resolution on one RTX 4090, prior extraction takes 3.40s, and the restoration network takes 0.63s with 1.72B parameters, 1512.96G FLOPs, and 4.96GB peak memory.

Datasets Training uses a combination of real paired data and synthetic data. The real subset contains 89 image pairs from Real [68], 200 pairs from Nature [27], 1230 pairs from RR4k [6], 6600 pairs from RRW [74] and 23303 pairs

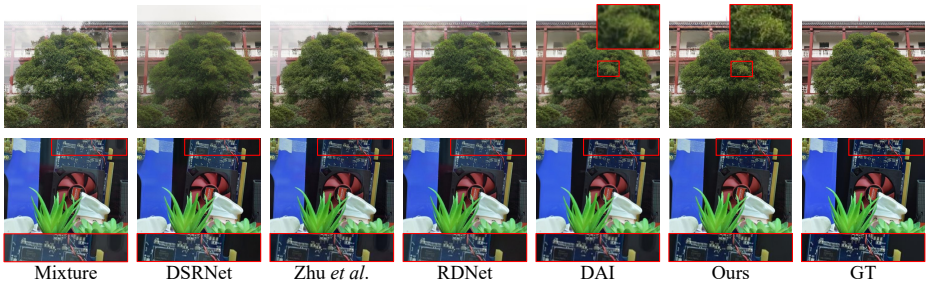

Fig. 4: Qualitative comparisons on representative examples from the three benchmarks. The red rectangles highlight key regions for comparison.

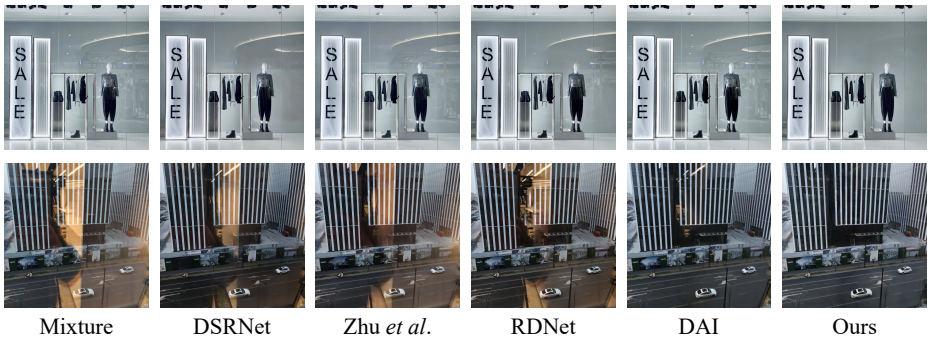

Fig. 5: Qualitative comparisons on challenging real-world mixed images.

from DRR [15]. The remaining image pairs from Real and Nature are used for evaluation. For synthetic data, we generate 16120 training pairs by blending images sampled independently from the COCO [30] dataset, using formulation from DSRNet [17] $M = \gamma_1 T + \gamma_2 R - \gamma_1\gamma_2 T \odot R$. Following [69], we sample γ_1 and γ_2 for 3 channels individually. During training, all images are resized and randomly cropped to a resolution of 768, with random flipping and color jitter applied for augmentation.

4.2 Comparison with State-of-the-Art Methods

We evaluate reflection removal on three widely used real benchmarks, including Nature [27], Real [68], and SIR2 [46], following their standard test splits and evaluation protocols. To assess robustness under unconstrained imaging conditions, we further report qualitative results on a small set of in-the-wild mixed images collected from the Internet, which exhibit diverse reflection patterns and challenging scene content. We compare our method with representative single image reflection removal approaches, including IBCLN [27], Dong *et al.* [10], YTMT [16], DSRNet [17], Zhu *et al.* [74], RDNet [69] and DAI [15]. For methods with publicly available training code, we follow the official implementations

Table 1: Quantitative comparisons on three reflection removal benchmarks (Nature, Real, and SIR^2). For each evaluation metric, the arrow ↑ (↓) indicates that larger (smaller) values are better. The best results are highlighted in **bold**, and the second-best results are underlined.

Dataset (size)	Metric	Method							
		IBCLN	Dong *et al.*	YTMT	DSRNet	Zhu *et al.*	RDNet	DAI	Ours
Nature (20)	PSNR↑	23.77	23.33	20.76	24.58	25.68	25.74	26.81	**26.93**
	SSIM↑	0.786	0.812	0.768	0.817	0.826	0.828	**0.840**	**0.840**
	LPIPS↓	0.145	0.117	0.183	0.120	0.103	0.109	0.203	**0.088**
	CLIPIQA↑	0.408	0.419	0.400	**0.440**	0.432	0.431	0.362	0.439
	MUSIQ↑	58.83	59.56	59.63	62.09	61.37	60.26	54.70	**62.87**
Real (20)	PSNR↑	21.55	22.34	22.87	23.37	21.85	24.81	25.21	**25.95**
	SSIM↑	0.767	0.811	0.808	0.801	0.781	0.838	0.838	**0.852**
	LPIPS↓	0.210	0.150	0.157	0.157	0.183	0.118	0.150	**0.097**
	CLIPIQA↑	0.444	0.441	0.464	0.483	0.453	**0.538**	0.414	0.509
	MUSIQ↑	58.08	57.55	58.72	60.44	58.46	61.42	58.20	**62.12**
SIR^2 (500)	PSNR↑	23.89	24.29	23.73	25.51	25.37	26.46	**27.35**	27.22
	SSIM↑	0.886	0.902	0.891	0.913	0.904	0.921	0.923	**0.925**
	LPIPS↓	0.127	0.099	0.119	0.094	0.112	0.080	0.093	**0.067**
	CLIPIQA↑	0.384	0.394	0.385	0.405	0.408	0.396	0.365	**0.419**
	MUSIQ↑	58.16	57.55	58.09	58.82	58.11	58.53	54.40	**59.85**
Average (540)	PSNR↑	23.79	24.17	23.57	25.39	25.24	26.36	**27.24**	27.15
	SSIM↑	0.877	0.895	0.882	0.905	0.896	0.914	0.916	**0.918**
	LPIPS↓	0.131	0.102	0.123	0.098	0.114	0.083	0.100	**0.069**
	CLIPIQA↑	0.387	0.397	0.388	0.410	0.411	0.403	0.367	**0.424**
	MUSIQ↑	58.19	57.63	58.18	59.02	58.26	58.71	54.57	**59.88**

and finetune them on our training data to reduce the impact of mismatched data distributions. All competing methods are evaluated using the same test images, preprocessing pipeline, and metrics.

Quantitative Comparison. In quantitative evaluation, we employ PSNR [18] and SSIM [53] to measure fidelity with respect to the ground truth transmission, together with LPIPS [67] for perceptual similarity. We additionally report CLIP-IQA [48] and MUSIQ [21] as complementary no-reference quality indicators. Quantitative results are summarized in Tab. 1. Across the three benchmarks, our method remains competitive on distortion-oriented metrics (PSNR/SSIM) and shows clear advantages on perceptual metrics (LPIPS/CLIPIQA/MUSIQ), indicating that the proposed guidance and gated modulation improve visual quality while better preserving transmission structures under complex reflections.

Qualitative Comparison. Fig. 4 shows qualitative comparisons on representative examples from the three benchmarks. In addition, Fig. 5 presents results on extra real-world mixed images, including photos captured by us and images collected from the Internet, to further assess robustness under diverse reflection

Fig. 6: Qualitative ablation results on gated modulation.

conditions. The results demonstrate that previous methods often struggle under complex reflections, leading to residual reflection traces, local over-smoothing, or color inconsistencies. In contrast, our results better retain geometric coherence and fine structures, particularly around edges and textured regions, while removing reflection patterns that are entangled with transmission content.

4.3 Ablation Study

Ablation on Gated Modulation. This part isolates the role of gated modulation in the coarse stage, where conditional residual injections are spatially modulated before entering the diffusion denoiser. All variants share the same training methods, and the only change lies in how the gate is formed from the guidance signals. Specifically, we compare: (i) *w/o gate*, where the injection tensors are passed without modulation ($g \equiv 1$); (ii) *intensity only*, where the gate is driven solely by the intensity prior ($g = 1+\beta P_{\text{int}}$); (iii) *high-frequency only*, where the gate is driven solely by the high-frequency prior ($g = 1+\beta P_{\text{hf}}$); and (iv) *full*, where the gate combines both signals as in Eq. (5). We additionally evaluate a *concat* variant that directly concatenates $\mathbf{P}_{\text{int}}$ and $\mathbf{P}_{\text{hf}}$ with the ControlNet condition instead of modulating residual injections. Tab. 2 reports quantitative results on the same benchmarks and metrics as in Sec. 4.2, and Fig. 6 provides representative visual comparisons.

Across datasets, removing gated modulation degrades perceptual quality most noticeably and the outputs exhibit more residual reflection artifacts and more obvious restoration traces in Fig. 6. Using a single guidance signal partially alleviates these artifacts, but the two priors emphasize different failure modes. The intensity driven gate primarily improves suppression in high severity regions, whereas the high-frequency driven gate improves responses around edges and textured regions. Direct concatenation improves over the no-gate baseline, but remains weaker than gated modulation, indicating that the priors are more effective when used to modulate residual injections. Combining P_{int} and P_{hf} in the gate consistently gives the most favorable perceptual scores among the variants and produces cleaner reflection removal with more faithful transmission

Table 2: Quantitative ablation results on gated modulation and refinement module. For each evaluation metric, the arrow ↑ (↓) indicates that larger (smaller) values are better. The best results are highlighted in **bold**. HF-only refers to *high-frequency only* variant

Metric	Abl. on gated modulation				Abl. on refinement module		
	W/o gate	Concat	Intensity-only	HF-only	Coarse-only	Fine-decoder	Ours
PSNR↑	26.55	26.87	26.94	26.78	26.27	26.75	**27.15**
SSIM↑	0.897	0.906	0.905	0.906	0.865	0.891	**0.918**
LPIPS↓	0.076	0.072	0.073	0.071	0.133	0.084	**0.069**
CLIPIQA↑	0.390	0.406	0.403	0.393	0.385	0.398	**0.424**
MUSIQ↑	57.13	58.55	58.40	58.08	57.56	58.43	**59.88**

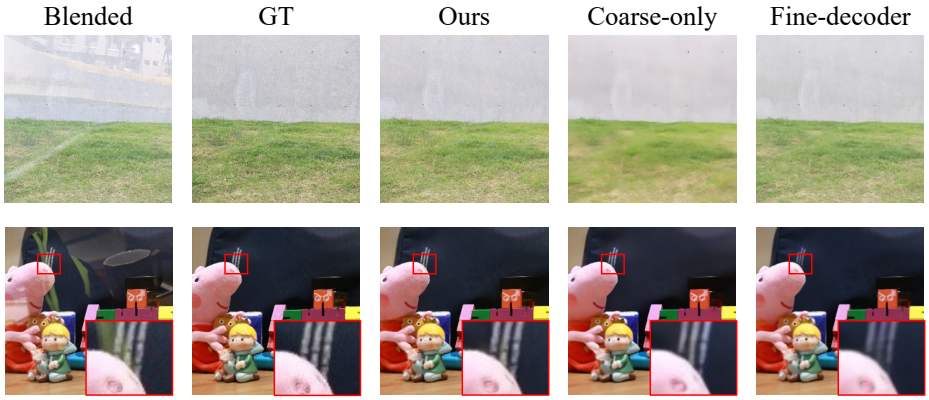

Fig. 7: Qualitative ablation results on refinement module.

structures, supporting the design choice of modulating coarse stage injections with a severity aware and detail sensitive signal.

Ablation on Refinement Module. This experiment examines the contribution of the Fine-Grained Refinement by comparing the full pipeline with two variants: (i) the *coarse-only* variant and (ii) the *fine-decoder* variant. While the *coarse-only* variant directly decodes the Coarse-Grained Restoration output as the final prediction, the *fine-decoder* variant utilizes a fine-tuned decoder following the training protocol in DAI [15]. All settings use the same coarse model.

As summarized in Tab. 2 and illustrated in Fig. 7, the *coarse-only* variant is already effective at suppressing dominant reflections, confirming the capability of the one-step coarse restoration. However, its outputs often exhibit noticeably reduced sharpness and visible restoration traces, particularly around edges and textured regions where reflection patterns intertwine with transmission structures. With a fine-tuned decoder, *fine-decoder* variant shows partial improvements in artifact removal and detail enhancement, but falls short of fully addressing the

underlying problem. Introducing the refinement module largely improves visual clarity and structural coherence. Residual blur and local artifacts are substantially reduced. These observations support the role of the refinement network as a deterministic correction step that consolidates the coarse restoration into a sharper and more geometrically consistent transmission estimate.

5 Conclusion

This paper proposed a diffusion framework for single image reflection removal that improves spatial adaptivity and structural faithfulness under varying reflections. The method derives two guidance priors from the mixed image and uses them to guide restoration in a complementary manner. The coarse stage applies gated modulation to strengthen reflection suppression where it is needed and to better preserve transmission structures in detail sensitive regions. The refinement module improves geometric consistency and visual clarity and reduces blur and restoration traces. Experiments on standard benchmarks and additional real-world mixtures show competitive performance and clear improvements on perceptual quality measures.

Limitations. The method can be ambiguous when reflection and transmission content overlap with similar intensity and similar structures. In such cases, it is difficult to determine which content should be preserved. User-provided spatial guidance such as coarse masks or sparse edge annotations is a practical direction to reduce this ambiguity. In addition, the VLM-derived prior extraction introduces extra inference overhead compared with purely feed-forward restoration networks. It can be mitigated by replacing prior extraction with a lightweight predictor.

Acknowledgements

The work was supported in part by the National Natural Science Foundation of China under Grant 62301310 and 62572317.

References

1. Agrawal, A., Raskar, R., Nayar, S.K., Li, Y.: Removing photography artifacts using gradient projection and flash-exposure sampling. In: ACM SIGGRAPH 2005 Papers. pp. 828–835 (2005)
2. Arvanitopoulos, N., Achanta, R., Süsstrunk, S.: Single image reflection suppression. In: CVPR. pp. 4498–4506 (2017)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)

4. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A ViT backbone for diffusion models. In: CVPR. pp. 22669–22679 (2023)
5. Chen, F.L., Zhang, D.Z., Han, M.L., Chen, X.Y., Shi, J., Xu, S., Xu, B.: Vlp: A survey on vision-language pre-training. Machine Intelligence Research **20**(1), 38–56 (2023)
6. Chen, G.Y., Zheng, C.W., Fan, G., Su, J.N., Gan, M., Chen, C.P.: Real-world image reflection removal: An ultra-high-definition dataset and an efficient baseline. IEEE TCSVT **35**(5), 4397–4408 (2025)
7. Chen, L., Bai, S., Chai, W., Xie, W., Zhao, H., Vinci, L., Lin, J., Chang, B.: Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. arXiv preprint arXiv:2502.20172 (2025)
8. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV. pp. 17–33 (2022)
9. Dong, W., Zhou, H., Zhang, Y., Liu, X., Chen, J.: Ecmamba: Consolidating selective state space model with retinex guidance for efficient multiple exposure correction. In: NeurIPS. pp. 53438–53457 (2024)
10. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: ICCV. pp. 5017–5026 (2021)
11. Han, B.J., Sim, J.Y.: Reflection removal using low-rank matrix completion. In: CVPR. pp. 5438–5446 (2017)
12. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE transactions on pattern analysis and machine intelligence **35**(6), 1397–1409 (2012)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. pp. 6840–6851 (2020)
14. Hong, Y., Zhong, H., Weng, S., Liang, J., Shi, B.: L-DiffER: Single image reflection removal with language-based diffusion model. In: ECCV. pp. 58–76 (2024)
15. Hu, J., Yang, C., Zhou, Z., Fang, J., Yang, X., Tian, Q., Shen, W.: Dereflection any image with diffusion priors and diversified data. arXiv preprint arXiv:2503.17347 (2025)
16. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. In: NeurIPS. pp. 24683–24694 (2021)
17. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: ICCV. pp. 13138–13147 (2023)
18. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)
19. Jiang, H., Guan, B., Liu, Z., Liu, X., Yu, J., Liu, Z., Han, S., Liu, S.: Learning to see in the extremely dark. In: ICCV. pp. 7676–7685 (2025)
20. Jiang, H., Luo, A., Liu, X., Han, S., Liu, S.: Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In: ECCV. pp. 161–179 (2024)
21. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021)
22. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NeurIPS. pp. 1097–1105 (2012)
24. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: CVPR. pp. 14811–14820 (2021)
25. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: CVPR. pp. 1750–1758 (2020)

26. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. IEEE TPAMI **29**(9), 1647–1654 (2007)
27. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: CVPR. pp. 3565–3574 (2020)
28. Li, Y., Liu, M., Yi, Y., Li, Q., Ren, D., Zuo, W.: Two-stage single image reflection removal with reflection-aware guidance. Applied Intelligence **53**(16), 19433–19448 (2023)
29. Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G.: A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. arXiv preprint arXiv:2501.02189 (2025)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
31. Lin, X., Yu, F., Hu, J., You, Z., Shi, W., Ren, J.S., Gu, J., Dong, C.: Harnessing diffusion-yielded score priors for image restoration. arXiv preprint arXiv:2507.20590 (2025)
32. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1X-Edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
33. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: ICCV. pp. 7314–7323 (2019)
34. Liu, X., Shi, K., Wang, Z., Chen, J.: Exploit camera raw data for video super-resolution via hidden markov model inference. IEEE TIP **30**, 2127–2140 (2021)
35. Liu, X., Shi, Z., Wu, Z., Chen, J., Zhai, G.: Griddehazenet+: An enhanced multi-scale network with intra-task knowledge transfer for single image dehazing. IEEE Transactions on Intelligent Transportation Systems **24**(1), 870–884 (2022)
36. Liu, Y., Ma, X., Bailey, J., Lu, F.: Reflection backdoor: A natural backdoor attack on deep neural networks. In: ECCV. pp. 182–199 (2020)
37. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research **22**(4), 730–751 (2025)
38. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML. pp. 8162–8171 (2021)
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: NeurIPS. pp. 8024–8035 (2019)
40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
41. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
43. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: CVPR. pp. 3193–3201 (2015)
44. Simon, C., Kyu Park, I.: Reflection removal for in-vehicle black box videos. In: CVPR. pp. 4231–4239 (2015)

45. Szeliski, R., Avidan, S., Anandan, P.: Layer extraction from multiple images containing reflections and transparency. In: CVPR. pp. 246–253 (2000)
46. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: ICCV. pp. 3922–3930 (2017)
47. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. IJCV **129**(2), 385–399 (2021)
48. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: AAAI. pp. 2555–2563 (2023)
49. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. IJCV **132**(12), 5929–5949 (2024)
50. Wang, R., Li, W., Liu, X., Li, C., Zhang, Z., Min, X., Zhai, G.: Hazeclip: Towards language guided real-world image dehazing. In: ICASSP. pp. 1–5 (2025)
51. Wang, R., Zheng, Y., Zhang, Z., Li, C., Liu, S., Zhai, G., Liu, X.: Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In: CVPR. pp. 23091–23100 (2025)
52. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Yang, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Zhou, B., Su, W., Chen, K., Qiao, Y., Wang, W., Luo, G.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
53. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers. pp. 1398–1402 (2003)
54. Wei, H., Xu, B., Liu, H., Wu, C., Liu, J., Peng, Y., Wang, P., Liu, Z., He, J., Xietian, Y., Tang, C., Wang, Z., Wei, Y., Hu, L., Jiang, B., Li, W., He, Y., Liu, Y., Song, X., Li, E., Zhou, Y.: Skywork UniPic 2.0: Building kontext model with online RL for unified multimodal model. arXiv preprint arXiv:2509.04548 (2025)
55. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: CVPR. pp. 8178–8187 (2019)
56. Wu, G., Tao, X., Li, C., Wang, W., Liu, X., Zheng, Q.: Perception-oriented video frame interpolation via asymmetric blending. In: CVPR. pp. 2753–2762 (2024)
57. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023)
58. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. In: NeurIPS. pp. 92529–92553 (2024)
59. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
60. Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. ACM TOG **34**(4), 1–11 (2015)
61. Yang, J., Li, H., Dai, Y., Tan, R.T.: Robust optical flow estimation of double-layer images under transparency or reflection. In: CVPR. pp. 1410–1419 (2016)

62. Yang, K., Cai, J., Ouyang, L., Vasluianu, F.A., Timofte, R., Ding, J., Sun, H., Fu, L., Li, J., Ho, C.M., Meng, Z., Li, M., Wang, H., Hu, Q., Wang, J., Zhao, H., Hu, J., Guo, X., Yang, M., He, J., Wang, Y., Chen, Z., Fang, H., Zhang, W., Cong, R., Hegde, D.D., Kalal, J., Akalwadi, N., Tabib, R.A., Mudenagudi, U., Lin, Y.F., Lee, C.M., Hsu, C.C., Zhang, M., Nathan, S., Uma, K., Sasithradevi, A., Bama, B.S., Roomi, S.M.M., Benjdira, B., Ali, A.M., Boulila, W., Dong, W., Li, Y., Hussein, A., Zhou, H., Chen, J., Xiao, Z., Li, Z.: Ntire 2025 challenge on single image reflection removal in the wild: Datasets, methods and results. In: CVPRW. pp. 1301–1311 (2025)
63. Yao, M., Wang, M., Tam, K.M., Li, L., Xue, T., Gu, J.: PolarFree: Polarization-based reflection-free imaging. In: CVPR. pp. 10890–10899 (2025)
64. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: StableNormal: Reducing diffusion variance for stable and sharp normal. ACM TOG **43**(6), 1–18 (2024)
65. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
66. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023)
67. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
68. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: CVPR. pp. 4786–4794 (2018)
69. Zhao, H., Li, M., Hu, Q., Guo, X.: Reversible decoupling network for single image reflection removal. In: CVPR. pp. 26430–26439 (2025)
70. Zheng, Q., Shi, B., Chen, J., Jiang, X., Duan, L.Y., Kot, A.C.: Single image reflection removal with absorption effect. In: CVPR. pp. 13395–13404 (2021)
71. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. In: CVPR. pp. 24913–24922 (2024)
72. Zhou, H., Dong, W., Liu, X., Liu, S., Min, X., Zhai, G., Chen, J.: Glare: Low light image enhancement via generative latent feature based codebook retrieval. In: ECCV. pp. 36–54 (2024)
73. Zhou, H., Dong, W., Liu, X., Zhang, Y., Zhai, G., Chen, J.: Low-light image enhancement via generative perceptual priors. In: AAAI. pp. 10752–10760 (2025)
74. Zhu, Y., Fu, X., Jiang, P.T., Zhang, H., Sun, Q., Chen, J., Zha, Z.J., Li, B.: Revisiting single image reflection removal in the wild. In: CVPR. pp. 25468–25478 (2024)