L-DiffER: Single Image Reflection Removal with Language-based Diffusion Model

Yuchen Hong[1,2 #] Haofeng Zhong[1,2,3 #] Shuchen Weng[1,2]
Jinxiu Liang[1,2] Boxin Shi[1,2,3 *]

[1] State Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology,
School of Computer Science, Peking University
[3] AI Innovation Center, School of Computer Science, Peking University
{hfzhong, shuchenweng, cssherryliang, shiboxin}@pku.edu.cn,
yuchenhong.cn@gmail.com

Abstract. In this paper, we introduce L-DiffER, a language-based diffusion model designed for the ill-posed single image reflection removal task. Although having shown impressive performance for image generation, existing language-based diffusion models struggle with precise control and faithfulness in image restoration. To overcome these limitations, we propose an iterative condition refinement strategy to resolve the problem of inaccurate control conditions. A multi-condition constraint mechanism is employed to ensure the recovery faithfulness of image color and structure while retaining the generation capability to handle low-transmitted reflections. We demonstrate the superiority of the proposed method through extensive experiments, showcasing both quantitative and qualitative improvements over existing methods.

Keywords: Reflection removal · Language-based diffusion model

1 Introduction

When photographs are taken through transparent mediums like glass windows, the contaminated mixture image (denoted as $\mathbf{M}$) can be considered as the combination of a transmission layer (denoted as $\mathbf{T}$) and a reflection layer (denoted as $\mathbf{R}$) [61, 62], which impairs the performance of downstream computer vision tasks [1, 47, 60]. Consequently, reflection removal, which aims at removing undesired reflections and recovering clear transmission layers from contaminated mixture images, has become an attractive topic in the field of computational photography [8, 9, 16, 31, 59, 65, 77]. State-of-the-art single-image reflection removal methods [8, 22, 31, 61] predominantly learn deep priors from a mixed dataset of synthetic and real data to mitigate reflections. However, due to insufficient

[#] Equal contributions. [*] Corresponding author.

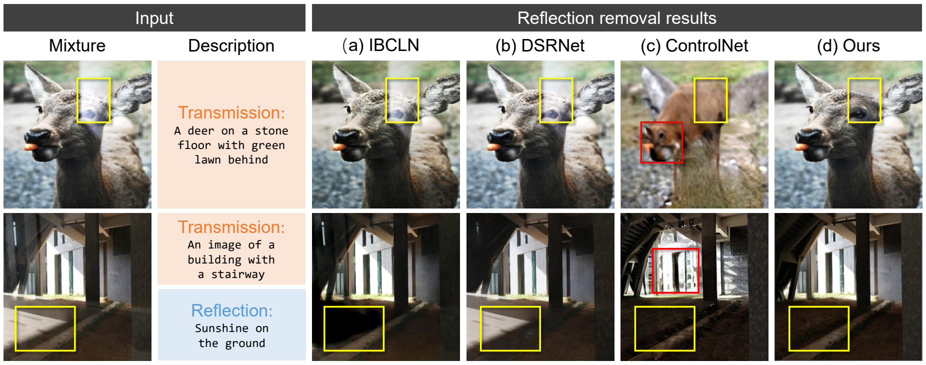

Fig. 1: For mixture images with strong reflections, (a-b) single-image reflection removal methods [22, 31] fail in distinguishing reflection and transmission layers. (c) Using language descriptions as guidance and mixture images as conditions, ControlNet [74] generates results with color shifts and structure distortions, while (d) the proposed L-DiffER successfully achieves high-fidelity transmission recovery in low-transmitted reflection regions.

knowledge about transmission and reflection scenes for addressing such a highly ill-posed problem, they often encounter limitations in scenes with complex or low-transmitted reflections, as shown in yellow boxes of Fig. 1(a-b). By leveraging auxiliary scene information from additional inputs obtained with specialized devices [27, 28, 37] or capture settings [34, 35, 43], multi-image reflection removal methods achieve more robust reflection removal than single-image methods. However, the specialized requirements for data acquisition, in contrast, limit their application scopes, especially for mobile devices and images from the Internet. Hence, it is imperative to find a user-friendly auxiliary input that helps to relieve the ill-posedness of reflection removal while maintaining the applicability and accessibility of single-image methods.

Fortunately, natural language can exactly serve as the aforementioned auxiliary input, since it provides instructive semantic information about images [36, 63, 72], which bridges the gap between human understanding and machine perception. Thanks to the development of vision-language models [48], various vision tasks (*e.g.*, image editing [55, 56, 67, 68] and image colorization [3–5, 69]) benefit from the semantic guidance of language descriptions. Zhong *et al.* [81] make the first attempt to introduce language descriptions for reflection removal, but due to the lack of generation capability and generalizability, it encounters difficulties in addressing scenarios containing low-transmitted reflections.

Recently, by combining semantic information from natural language descriptions and generative priors from pre-trained diffusion models [14, 49], language-based diffusion models (*e.g.*, ControlNet [74]) generate visually pleasant results for image restoration tasks like image colorization [4] and super resolution [54]. For reflection removal, it also shows the potential to provide semantic priors to relieve the ill-posedness, where the generative priors in pre-trained diffusion models can adeptly manage complex scenarios and make it possible to recover transmission layers under low-transmitted reflections. However, as shown in the red boxes of Fig. 1(c), directly employing the existing language-based diffusion model like ControlNet [74] for reflection removal will encounter two primary challenges: **(1) Inaccurate control conditions**. When using language-based diffusion models, other tasks such as image colorization [4] often employ a clear (not superimposed) image to provide accurate structure control. However, in the reflection removal task, the image content of the transmission layer in a mixture image is contaminated by the reflection layer, and thus it is unable to provide an initially accurate condition for both structure and color control. Consequently, achieving transmission recovery under such inaccurate conditions presents a significant challenge. **(2) Insufficient recovery faithfulness**. Since diffusion models are principally designed for image generation, simply applying them for reflection removal may obtain results that are less faithful to conditions (*i.e.*, results with unwanted image content containing color shifts and structural distortions compared with mixture images). Therefore, it is essential to develop a mechanism for constraining the generation capability and improving the recovery faithfulness of diffusion models to recover transmission layers while preventing unwanted alterations in image color and structure.

In this paper, we propose **L-DiffER**, the *first* **L**anguage-based **Diff**usion model for single image r**E**flection **R**emoval, to achieve high-fidelity transmission recovery with language guidance under the interference of low-transmitted reflections, as the results shown in Fig. 1(d). To leverage the language descriptions of two layers, descriptions of the transmission layer serve as positive prompts, guiding the model to preserve essential image content, whereas descriptions of the reflection layer act as negative prompts, aiding in the suppression of unwanted reflections. To address the aforementioned challenges, we propose an iterative condition refinement strategy within the reverse diffusion process to ensure increasingly accurate representations of color and structure conditions as the process evolves. Furthermore, we introduce a multi-condition constraint mechanism designed to safeguard the faithfulness to the specified conditions, which effectively addresses potential color shifts and structural distortions by applying constraints that align the recovery output with the desired color and structure conditions. The proposed method achieves state-of-the-art performance by making the following contributions:

- We present the first language-based diffusion model for reflection removal, leveraging the descriptive power of language to distinguish between transmission and reflection layers.
- We propose an iterative condition refinement strategy to provide more accurate control conditions as color and structure guidance for reflection removal.
- We design a multi-condition constraint mechanism to ensure the recovery faithfulness of image color and structure, which effectively handles low-transmitted reflections.

2 Related Work

Reflection removal. Single image reflection removal aims at suppressing reflections and recovering transmission layers in a single mixture image. Using the presumption that reflections tend to blur and exhibit low intensities, non-learning approaches employ handcrafted priors into their optimization frameworks [57,71], such as the gradient sparsity [30], relative smoothness [33], and ghosting cues [51]. With the development of deep learning, researchers attempt to address the task by exploring different network architectures [9,21,59,61,65], loss functions [76], or the iterative recovery strategy [8,31,70,77]. Concurrently, efforts in data synthesis to meet the data-driven demands is also an attractive topic in this area [9,22,24,39,66,79,80]. Besides, the exploration of special image forms like panoramic images [10,19,20,45] helps to diminish content ambiguity in mixture images. To relieve the ill-posedness of reflection removal, multi-image methods utilize additional input images to introduce auxiliary clues such as polarization information [7,26,28,37,38,44,50], transmission scene information from active illuminations [2,17,18,27,29], and motion discrepancy between the two layers [32,34,35,52], while special data capture requirements limit their practical applications. Inspired by Zhong *et al*. [81], we introduce language-based diffusion models to input auxiliary language descriptions for relieving the ill-posedness of reflection removal while maintaining the applicability of the method, and to leverage generative priors for handling low-transmitted reflection regions.

Language-based diffusion models. The advent of diffusion models [14] (DMs) has marked a significant leap forward, particularly in the realms of image generation [49]. The emergence of pre-trained vision-language models, notably CLIP [48], significantly enhances the flexibility and intuitiveness of utilizing language descriptions for vision tasks. Specifically, these pre-trained models have been instrumental in enabling multi-condition image editing [42,74,78], image colorization [4], and devising sophisticated sampling strategies for tailored generation outcomes [12]. Despite the strides in image editing and conditional generation, the specific challenge of single image reflection removal, *i.e.*, recovering clear transmission layers from the guidance of a superimposed mixture images, still remains unexplored for language-based diffusion models [49,74]. In our work, we seek to address the reflection removal problem by harnessing the generative prowess and flexibility of language-guided diffusion models, which combines the semantic information offered by language descriptions with the sophisticated generative capabilities of DMs while ensuring the recovery fidelity and faithfulness with condition refinement and constraints.

3 Methodology

In this paper, we address the problem of language-based reflection removal using diffusion models. We first provide a brief overview of the framework of our language-based reflection removal diffusion model L-DiffER in Sec. 3.1. Sec. 3.2 proposes the iterative condition refinement strategy designed for addressing inaccurate control conditions, and Sec. 3.3 proposes the multi-condition constraint

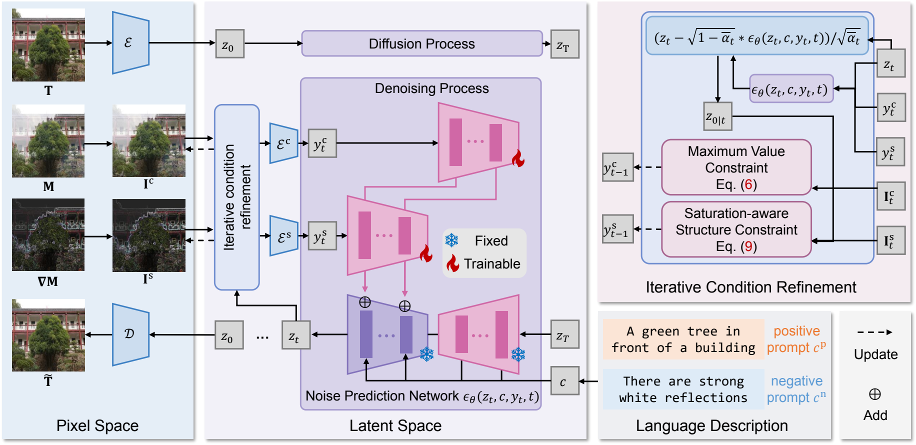
T
M
$\mathbf{I}^c$
∇M
$\mathbf{I}^s$
$\widetilde{\mathbf{T}}$
$\mathcal{E}$
$\mathcal{D}$
Pixel Space
z_0
Diffusion Process
z_T
Denoising Process
Iterative condition refinement
$\mathcal{E}^c$
y_t^c
$\mathcal{E}^s$
y_t^s
Fixed
Trainable
z_0
...
z_t
z_T
c
Noise Prediction Network $\epsilon_\theta(z_t, c, y_t, t)$
Latent Space
$(z_t - \sqrt{1-\overline{\alpha}_t} * \epsilon_\theta(z_t, c, y_t, t))/\sqrt{\overline{\alpha}_t}$
$\epsilon_\theta(z_t, c, y_t, t)$
$z_{0|t}$
z_t
y_t^c
y_t^s
$\mathbf{I}_t^c$
$\mathbf{I}_t^s$
y_{t-1}^c
y_{t-1}^s
Maximum Value Constraint Eq. (6)
Saturation-aware Structure Constraint Eq. (9)
Iterative Condition Refinement
A green tree in front of a building
positive prompt c^p
There are strong white reflections
negative prompt c^n
Language Description
Update
⊕
Add

Fig. 2: Framework of the proposed language-based reflection removal diffusion model, which leverages language descriptions with the iterative condition refinement strategy and the multi-condition constraint mechanism to achieve faithful transmission recovery.

mechanism for improving the recovery faithfulness. Finally, loss functions and training details are presented in Sec. 3.4 and Sec. 3.5, respectively.

3.1 Overview

Problem formulation. Given the mixture image $\mathbf{M} \in \mathbb{R}^{H \times W \times 3}$ and the language description condition c, we optimize our model to remove reflection contamination and recover transmission layers $\widetilde{\mathbf{T}} \in \mathbb{R}^{H \times W \times 3}$ under the supervision of the ground truth transmission layers $\mathbf{T} \in \mathbb{R}^{H \times W \times 3}$. We apply the commonly-used assumption [21,31,61] as: $\mathbf{M} = \mathbf{T} + \mathbf{R}$, where the mixture image $\mathbf{M}$ can be considered as the combination of a transmission layer $\mathbf{T}$ and a reflection layer $\mathbf{R}$. We use the intensity and gradient of the mixture image $\mathbf{M}$ to initialize the color condition $\mathbf{I}^{\mathrm{s}} = \mathbf{M}$ and the structure condition $\mathbf{I}^{\mathrm{s}} = \nabla \mathbf{M}$. Then the color latent $y^{\mathrm{c}} = \mathcal{E}^{\mathrm{c}}(\mathbf{I}^{\mathrm{c}})$ and structure latent $y^{\mathrm{s}} = \mathcal{E}^{\mathrm{s}}(\mathbf{I}^{\mathrm{s}})$ are extracted by a color and a structure encoder $\mathcal{E}^{\mathrm{c}}$ and $\mathcal{E}^{\mathrm{s}}$ (denoted by the blue blocks in the middle part of Fig. 2), respectively. Given language descriptions about the content of transmission layers and reflection layers as positive prompt and negative prompt, respectively, the positive and negative latents c^{p} and c^{n} are extracted from the CLIP text encoder [48]. In the proposed method, the transmission layer $\widetilde{\mathbf{T}}$ is recovered from all the spatial and language conditioning information encoded in latents $y = [y^{\mathrm{s}}, y^{\mathrm{c}}]$ and $c = [c^{\mathrm{p}}, c^{\mathrm{n}}]$.

Diffusion models. Diffusion models [14,53] are probabilistic generative models that have been shown to have powerful data modeling ability for images, which can provide an effective image prior for the estimation of transmission layers. To address the challenges of resource consumption in pixel space, Stable Diffusion (SD) [49] introduces latent diffusion models to improve both the training and sampling efficiency of denoising diffusion models without degrading their quality. Following SD [49], we employ a compression encoder $\mathcal{E}$ to encode a transmission layer into the latent space as $z = \mathcal{E}(\mathbf{T})$, and a compression decoder $\mathcal{D}$ to reconstruct the transmission layer from the latent code z as $\widetilde{\mathbf{T}} = \mathcal{D}(z)$. The diffusion process progressively destructs data by injecting noise, then learns to reverse this process for image generation. During training, the forward diffusion process starts from the target image $z_0 = \mathcal{E}(\mathbf{T})$ and generates samples $z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, 1)$ is the Gaussian noise at timestep t, and $\overline{\alpha}_t$ represents the noise scheduler introduced in [49]. During inference, the reverse diffusion process starts from a random noise sample $z_T \sim \mathcal{N}(0, 1)$, where a noise prediction network denoted as ϵ_θ is optimized to predict the noise $\hat{\epsilon}_t$ at each timestep t given the language condition c and the spatial conditions y_t[1] as follows:

$$\mathcal{L}_{\mathrm{ldm}} = \mathbb{E}_{\mathcal{E}(x),t,\epsilon\sim\mathcal{N}(0,1)} \left[\| \epsilon_t - \epsilon_\theta(z_t, c, y_t, t) \|_2^2 \right], \tag{1}$$

until it converges to the desired photorealistic transmission layer $\widetilde{\mathbf{T}}$. The overall framework is shown in Fig. 2.

Condition extraction. During inference, the two conditioning latents start from the initial color and structure latents, *i.e.*, $y_T^{\mathrm{c}} = y^{\mathrm{c}}$ and $y_T^{\mathrm{s}} = y^{\mathrm{s}}$, respectively, and they are refined by employing the iterative condition refinement strategy (Sec. 3.2) to provide more accurate conditions. For condition injection, the proposed framework draws inspiration from ControlNet [74], which learns a parameter-efficient, parallel branch on top of SD. Specifically, two trainable copied modules (denoted by the pink blocks in Fig. 2) are utilized to take the color latent y_t^{c} and the gradient latent y_t^{s} with the latent presentation z_t as input to extract color and structure features, respectively. These features are then added directly to the corresponding scales of the locked module (denoted by the purple block in Fig. 2), which guides ϵ_θ in learning the noise distribution.

Language guidance. To explicitly instruct the model to suppress reflections from the language guidance, classifier-free guidance [15] is used to employ negative prompts to specify undesired reflection contamination. Specifically, two noise predictions with positive prompts c^{p} and negative prompts c^{n} are made at each timestep t. The fusion of these two predictions yields the output $\hat{\epsilon}_{t-1}$:

$$\hat{\epsilon}_{t-1} = \epsilon_\theta(z_t, c^{\mathrm{p}}, y_t, t) + \lambda_{\mathrm{cfg}} \times (\epsilon_\theta(z_t, c^{\mathrm{p}}, y_t, t) - \epsilon_\theta(z_t, c^{\mathrm{n}}, y_t, t)), \tag{2}$$

where λ_{cfg} is a hyperparameter. In our task, as reflections are not always recognizable [81], we set negative prompts as the empty string $c^{\mathrm{n}} = \varnothing$ in the case. For simplicity, we use $\epsilon_\theta(z_t, c, y_t, t)$ to represent the whole noise prediction process.

3.2 Iterative condition refinement strategy

Although with effective data modeling ability, diffusion models probably generate results unfaithful to the input image conditions [40,74]. This issue becomes even

[1] Note that the spatial conditions are varied with timesteps in the proposed method.

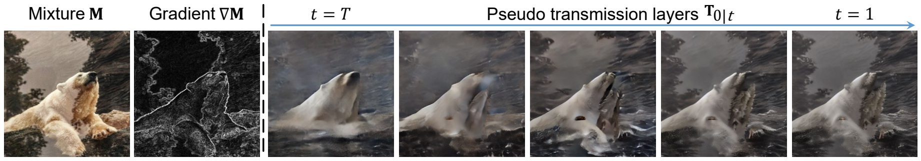
Mixture $\mathbf{M}$
Gradient $\nabla\mathbf{M}$
$t = T$
Pseudo transmission layers $\mathbf{T}_{0|t}$
$t = 1$

Fig. 3: Visualization of pseudo transmission layers by using the original mixture image with its gradient as conditions.

harder for reflection removal, since the desired transmission layers should have a structure without gradients of reflection contaminations, which is different from the input mixture images. That is, the mixture images itself cannot provide an accurate image condition for diffusion-based reflection removal. Therefore, we turn to the recovered transmission layers in the inference process, which have the potential to be combined with the mixture images and help obtain a refined condition. During inference, given a latent noisy sample z_t, we can use the noise prediction network ϵ_θ to predict noise as $\hat{\epsilon}_t = \epsilon_\theta(z_t, t, c, y_t)$. Then we can calculate the estimated latent representation of the transmission layer $z_{0|t}$ at the current timestep t as follows:

$$z_{0|t} = (z_t - \sqrt{1 - \overline{\alpha}_t} * \hat{\epsilon}_t)/\sqrt{\overline{\alpha}_t}. \tag{3}$$

After that, we can obtain the estimated pseudo transmission layer $\mathbf{T}_{0|t}$ by using the compression decoder $\mathcal{D}$ to decode $z_{0|t}$: $\mathbf{T}_{0|t} = \mathcal{D}(z_{0|t})$. Then $\mathbf{T}_{0|t}$ and its gradient $\nabla\mathbf{T}_{0|t}$ can be used as the pseudo-accurate color and structure condition to obtain conditions at the next timestep.

As pseudo transmission layers shown in Fig. 3, we observe two key phenomena: Firstly, the estimated pseudo transmission layers in the first few sampling steps exhibit inaccuracies, with neither structure nor color serving effectively as guiding conditions. However, as the sampling steps progress, the fidelity of the reconstructed images notably improves. Secondly, although employing the mixture image and the gradient of the mixture image as conditions persistently yield inaccurate control, these conditions provide a richer set of usable information than the recovered image at the beginning of sampling. Leveraging the advantages of both observations, we introduce an iterative condition refinement strategy to address the challenge of inaccurate control conditions. Our control conditions transitioned from the mixture image and the gradient of the mixture image to the recovered image and its gradient as the sampling step increases.

Specifically, we introduce predefined time-variant coefficients β_t and γ_t ($t \in \{1, .., T\}$) to combine pseudo-accurate conditions $\mathbf{T}_{0|t}$ with the one initiated from $\mathbf{M}$, which decrease as timestep t increases[2] as follows:

$$\mathbf{I}_t^{\mathrm{c}} = \gamma_t * \mathbf{T}_{0|t} + (1 - \gamma_t) * \mathbf{M}, \quad \mathbf{I}_t^{\mathrm{s}} = \beta_t * \nabla\mathbf{T}_{0|t} + (1 - \beta_t) * \nabla\mathbf{M}. \tag{4}$$

[2] Details of β_t and γ_t will be explained in the supplementary material.

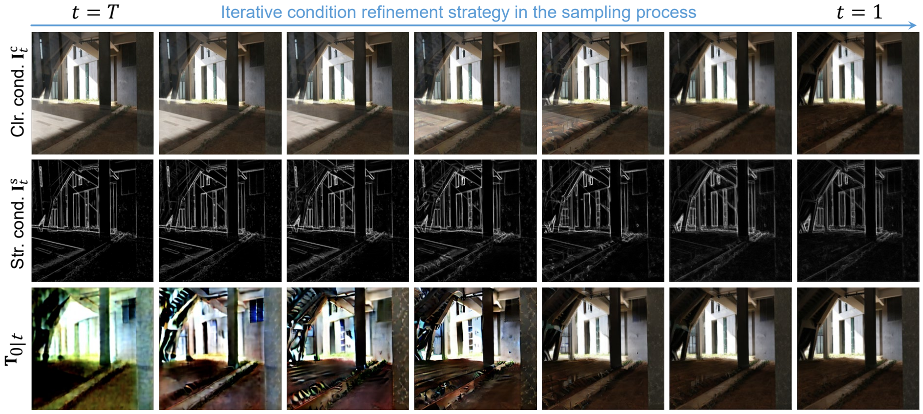
$t = T$
Iterative condition refinement strategy in the sampling process
$t = 1$
Clr. cond. $\mathbf{I}_t^c$
Str. cond. $\mathbf{I}_t^s$
$\mathbf{T}_{0|t}$

Fig. 4: Visualization of the color condition, structure condition, and the pseudo transmission layer by applying the proposed iterative condition refinement strategy.

Fig. 4 shows intermediate conditions and the pseudo transmission layers during sampling, indicating the effect of the iterative condition refinement strategy.

Meanwhile, we also use the iterative condition refinement strategy with a few modifications during training. We replace the pseudo-accurate transmission layers $\mathbf{T}_{0|t}$ estimated from sampling with ground truth $\mathbf{T}$ to refine conditions before estimating noises, which is different from the sampling process. Specifically, given the initial mixture image $\mathbf{M}$ and its gradient $\nabla\mathbf{M}$, we refine the input conditions at timestep t as follows:

$$\widehat{\mathbf{I}}_t^{\mathrm{c}} = \gamma_t * \mathbf{T} + (1-\gamma_t) * \mathbf{M}, \quad \widehat{\mathbf{I}}_t^{\mathrm{s}} = \beta_t * \nabla\mathbf{T} + (1-\beta_t) * \nabla\mathbf{M}, \tag{5}$$

where β_t and γ_t $(t \in \{1, ..., T\})$ are the same predefined time-variant coefficients used in the sampling process.

3.3 Multi-condition constraint mechanism

In order to mitigate the insufficient recovery faithfulness of diffusion models, we employ a multi-condition mechanism that separately utilizes mixture images and their gradients to obtain initial latent conditions for controlling the color and structure of generated results. As the iterative condition refinement strategy obtains new refined conditions during each timestep, no constraints applying on conditions may cause the sampling process out of control, which also impacts the recovery faithfulness. Hence, it is imperative to conduct constraints on refined conditions. Though inaccurate as color conditions, mixture images provide coarse color information about transmission layers, thus we propose a maximum value constraint to prevent color shifts. Similarly, the gradient of the mixture image provides features showing structure information such as edges and textures. Therefore, we propose a saturation-aware structure constraint to prevent structure distortions when recovering transmission layers. Details are as follows.

Maximum value constraint (MVC). According to the assumption defined in Sec. 3.1, the pixel values of the mixture image should be greater than or equal to the transmission layer. Therefore, when updating the conditions in the training and sampling processes, we employ maximum value constraint as a function to ensure that any given color condition $\mathbf{I}_t^{\mathrm{c}}$ defined in Eq. (4) does not exceed the mixture image $\mathbf{M}$ in terms of pixel values. This constraint should be adhered to the following:

$$\mathrm{MVC}(\mathbf{I}_t^{\mathrm{c}}) = \min(\mathbf{M}, \mathbf{I}_t^{\mathrm{c}}). \tag{6}$$

Saturation-aware structure constraint (SSC). Due to regional overexposure during photographing, mixture images sometimes contain low-transmitted reflection regions, which lack enough visible content of transmission layers, causing the corresponding gradients to have little usable structure information. Therefore, simply applying the maximum value constraint to edges will force the refined structure condition to abandon generated content in saturated regions, which may hinder optimization. Therefore, we introduce a saturation-aware mask to preserve the generative capability of the structural aspects in regions challenging for recovery. Specifically, we first utilize the mixture image to identify low-transmitted regions, generating a mask $\mathbf{M}^{\mathrm{s}} \in \mathbb{R}^{H\times W}$ as:

$$\boldsymbol{\Omega}^{\mathrm{s}} = \begin{cases} 0, & \text{where } \mathbf{M} > \tau \\ 1, & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbf{M}$ is the mixture image, τ is the saturation threshold we empirically set to 0.95. Besides, to ensure that the structure condition does not contain nonexistent edges in non-saturated regions of the original mixture image, we define a mask to indicate the valid edge in the gradient of the mixture image as:

$$\boldsymbol{\Omega}^{\mathrm{v}} = \begin{cases} 0, & \text{where } \nabla\mathbf{M} < \eta \\ 1, & \text{otherwise,} \end{cases} \tag{8}$$

where η is the valid edge threshold we empirically set to 0.05. After that, for a given gradient condition $\mathbf{I}_t^{\mathrm{s}}$ defined in Eq. (4), we define the saturation-aware structure constraint as follows:

$$\mathrm{SSC}(\mathbf{I}_t^{\mathrm{s}}) = \boldsymbol{\Omega}^{\mathrm{s}} \odot \mathbf{I}_t^{\mathrm{s}} + (1 - \boldsymbol{\Omega}^{\mathrm{s}}) \odot \boldsymbol{\Omega}^{\mathrm{v}} \odot \mathbf{I}_t^{\mathrm{s}}. \tag{9}$$

To extract the color latent y_t^{c} and the structure latent y_t^{s} at each timestep t, we apply the maximum value constraint (MVC) and saturation-aware structure constraint (SSC) on the color condition $\mathbf{I}_t^{\mathrm{c}}$ and structure condition $\mathbf{I}_t^{\mathrm{s}}$ as:

$$y_{t-1}^{\mathrm{c}} = \mathcal{E}^{\mathrm{c}}(\mathrm{MVC}(\mathbf{I}_t^{\mathrm{c}})), \quad y_{t-1}^{\mathrm{s}} = \mathcal{E}^{\mathrm{s}}(\mathrm{SSC}(\mathbf{I}_t^{\mathrm{s}})), \tag{10}$$

where $\mathcal{E}^{\mathrm{s}}$ is the structure condition encoder, $\mathcal{E}^{\mathrm{c}}$ is the color condition encoder, and MVC and SSC are applied to the color and structure conditions for ensuring recovery faithfulness. Finally, we apply the denoising process with the DDIM [53] sampling scheme to obtain the next noisy sample z_{t-1}. Algorithm 1 outlines the complete sampling process.

Algorithm 1 Iterative condition refinement strategy in the sampling process

```
Require: Noise prediction network $\epsilon_\theta$, Compression decoder $\mathcal{D}$, Color encoder $\mathcal{E}^{\mathrm{c}}$,
    Structure encoder $\mathcal{E}^{\mathrm{s}}$
Input: Mixture image $\mathbf{M}$, Language condition $c = [c^{\mathrm{p}}, c^{\mathrm{n}}]$, Initial spatial conditions
    $[y^{\mathrm{s}} = \mathcal{E}^{\mathrm{s}}(\nabla\mathbf{M}), y^{\mathrm{c}} = \mathcal{E}^{\mathrm{c}}(\mathbf{M})]$, Coefficients $\{\overline{\alpha}_t\}_{t=1}^{T}$, $\{\beta_t\}_{t=1}^{T}$, $\{\gamma_t\}_{t=1}^{T}$
Output: Recovered latent representation $z_0$
$z_T \sim \mathcal{N}(0,1)$
$y_T = [y^{\mathrm{s}}, y^{\mathrm{c}}]$
for each $t \in [T, 1]$ do
    $\hat{\epsilon}_t = \epsilon_\theta(z_t, c, y_t, t)$
    if $t > 1$ then
        $z_{0|t} = (z_t - \sqrt{1-\overline{\alpha}_t} * \hat{\epsilon}_t)/\sqrt{\overline{\alpha}_t}$          ▷ Compute $z_{0|t}$ by Eq. (3)
        $\mathbf{T}_{0|t} = \mathcal{D}(z_{0|t})$
        $y^{\mathrm{c}}_{t-1} = \mathcal{E}^{\mathrm{c}}(\mathrm{MVC}(\gamma_t * \mathbf{T}_{0|t} + (1-\gamma_t) * \mathbf{M}))$          ▷ Refine color condition
        $y^{\mathrm{s}}_{t-1} = \mathcal{E}^{\mathrm{s}}(\mathrm{SSC}(\beta_t * \nabla\mathbf{T}_{0|t} + (1-\beta_t) * \nabla\mathbf{M}))$          ▷ Refine structure condition
        $y_{t-1} = [y^{\mathrm{s}}_{t-1}, y^{\mathrm{c}}_{t-1}]$          ▷ Update spatial conditions
    end if
    $z_{t-1} = \sqrt{\overline{\alpha}_{t-1}} z_{0|t} + \sqrt{1-\overline{\alpha}_{t-1}} \hat{\epsilon}_t$          ▷ DDIM
end for
```

3.4 Loss functions

In this section, we introduce the latent diffusion model loss to minimize the discrepancy between the learned latent space distribution and the target distribution, as well as the pixel losses to prevent color shifts and structure distortions.

We follow the training objective of ControlNet [74] but replace task-specific conditions with the mixture image and the gradient of the mixture. Thus our network ϵ_θ learns to predict the noise added to the noisy latent z_t with $\mathcal{L}_{\mathrm{ldm}}$ defined in Eq. (1). According to Eq. (3), we can obtain $\mathbf{T}_{0|t}$ in pixel space by decoding $z_{0|t}$:

$$\mathbf{T}_{0|t} = \mathcal{D}(z_{0|t}) = \mathcal{D}((z_t - \sqrt{1-\overline{\alpha}_t} * \hat{\epsilon}_t)/\sqrt{\overline{\alpha}_t}). \tag{11}$$

Then RGB loss is introduced to constrain the consistency of the reconstructed transmission $\mathbf{T}_{0|t}$ and ground truth $\mathbf{T}$, minimizing their errors as follows:

$$\mathcal{L}_{\mathrm{RGB}} = \|\mathbf{T} - \mathbf{T}_{0|t}\|_2^2. \tag{12}$$

For better structure consistency, we use gradient loss to constrain the prediction's gradient and its ground truth as below:

$$\mathcal{L}_{\mathrm{grad}} = \|\nabla(\mathbf{T}) - \nabla(\mathbf{T}_{0|t})\|_2^2, \tag{13}$$

where $\nabla(\cdot)$ denotes the computation of image gradients. Furthermore, we also compute the standard deviation and the mean of the predicted transmission and its ground truth, and use a ℓ_2 loss to minimize the color distortion:

$$\mathcal{L}_{\mathrm{num}} = \|\mu(\mathbf{T}) - \mu(\mathbf{T}_{0|t})\|_2^2 + \|\sigma(\mathbf{T}) - \sigma(\mathbf{T}_{0|t})\|_2^2, \tag{14}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the computation of mean and standard deviation, respectively, and $i = 1, ..., N$ denotes the i-th sample in the dataset. Overall, the pixel loss is formulated as follows:

$$\mathcal{L}_{\text{pix}} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{num}}. \tag{15}$$

Gathering the latent diffusion model loss and pixel losses yields the final objective as:

$$\mathcal{L}_{\text{all}} \coloneqq \mathcal{L}_{\text{ldm}} + \lambda \mathcal{L}_{\text{pix}}, \tag{16}$$

where λ is set to 1.0 empirically.

3.5 Implementation details

We implement the proposed method using PyTorch [46]. The model is trained for 20 epochs with a batch size of 16 on two NVIDIA GeForce RTX 3090 GPUs. Weights are initialized as in [11] and updated using the Adam optimizer [25]. The learning rate is set to 1×10^{-5}. For sampling, we use the DDIM sampling method [53] with 50 sampling steps.

Dataset. Our training dataset contains both synthetic and real data. For the synthetic data, following Zhong *et al.* [81], we generate 50000 triplets of data using COCO Captions [6] and Flickr30k [73] for network training. For real data, we use 200 image pairs from Nature dataset [31] and 90 pairs from Zhang *et al.* [76] for training and the rest of the real data are used for evaluation. We further collect 50 mixture images from the Internet for qualitative evaluation.

4 Experiments

4.1 Comparison with state-of-the-arts

For evaluating model performances, experiments are conducted on existing real reflection datasets [31,58,76] and our collected mixture images from the Internet (with manually annotated language descriptions). We compare the proposed L-DiffER with state-of-the-art single-image reflection removal methods, including Zhang *et al.* [76], CoRRN [61], ERRNet [65], IBCLN [31], Dong *et al.* [8], YTMT [21], and DSRNet [22]. To compare with a representative language-based diffusion model ControlNet [74][3], we make it fit our setting by using mixture images as conditions, and transmission and reflection descriptions as positive and negative prompts, respectively. For fair comparisons, we finetune the above methods on our training data if their training codes are provided.

Quantitative comparison. Quantitative experiments are conducted on three real datasets for reflection removal, *i.e.*, Nature [31], Real20 [76], and SIR^2 [58] dataset. To evaluate the recovery of transmission layers[4], we utilize PSNR [23]

[3] Since ControlNet [74] destructs image color and structure in a generative manner as shown in Fig. 1(c), we only run it for qualitative comparisons.

[4] Evaluations on reflection layers are provided in the supplementary material.

Table 1: 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 [8,21,22,31,61,76]. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

Dataset (size)	Metric	Method							
		Zhang *et al.*	CoRRN	ERRNet	IBCLN	Dong *et al.*	YTMT	DSRNet	Ours
SIR^2 (500)	PSNR↑	22.45	22.96	23.13	23.36	23.09	23.05	24.97	**25.18**
	SSIM↑	0.872	0.879	0.878	0.881	0.893	0.886	0.907	**0.911**
	LPIPS↓	0.172	0.161	0.164	0.153	0.131	0.149	0.124	**0.121**
	NIQE↓	4.911	4.826	4.793	4.501	4.729	4.738	4.624	**4.352**
	FID↓	78.61	69.58	73.33	60.92	49.39	58.69	46.60	**44.66**
Real20 (20)	PSNR↑	22.51	21.17	22.06	21.59	21.73	22.31	23.46	**23.77**
	SSIM↑	0.806	0.786	0.803	0.771	0.811	0.805	0.806	**0.821**
	LPIPS↓	0.204	0.202	0.185	0.213	0.170	0.178	0.165	**0.153**
	NIQE↓	3.819	3.891	3.989	3.895	3.959	3.977	4.065	**3.665**
	FID↓	113.25	111.25	85.07	120.95	86.20	94.67	77.92	**72.64**
Nature (20)	PSNR↑	20.37	20.54	21.11	23.69	23.61	21.03	21.70	**23.95**
	SSIM↑	0.772	0.778	0.806	0.828	0.825	0.802	0.820	**0.831**
	LPIPS↓	0.217	0.205	0.183	0.170	0.157	0.186	0.168	**0.145**
	NIQE↓	4.413	4.433	4.442	4.345	4.581	4.387	4.373	**4.223**
	FID↓	109.77	89.12	79.68	75.17	71.55	81.77	74.84	**68.57**
Average (540)	PSNR↑	22.38	22.80	23.02	23.31	23.06	22.95	24.79	**25.08**
	SSIM↑	0.866	0.872	0.873	0.875	0.887	0.880	0.900	**0.905**
	LPIPS↓	0.175	0.164	0.165	0.156	0.133	0.151	0.127	**0.123**
	NIQE↓	4.852	4.777	4.750	4.473	4.695	4.697	4.594	**4.322**
	FID↓	81.05	71.85	74.00	63.67	51.57	60.88	48.81	**46.58**

and SSIM [64] as error metrics following reflection removal methods [8,37]. We further adopts LPIPS [75], NIQE [41], FID [13] to measure the perceptual quality of the recovered results. As quantitative results shown in Table 1, the proposed method achieves the best performance among all competing methods, especially in metrics for perceptual quality, validating its effectiveness.

Qualitative comparison. Qualitative comparisons with the aforementioned single-image reflection removal methods [8,21,22,31,61,76] and a language-based diffusion model ControlNet [74] on recovering transmission layers are shown in Fig. 5. As can be observed, single-image reflection removal methods encounter obstacles in dealing with complex reflections due to the lack of auxiliary semantic information, especially in low-transmitted reflection regions. ControlNet [74] performs significant modifications on the image color and structure in a generative manner, which deviates significantly from the input mixture image, emphasizing the pivotal role of accurate conditions in the sampling process. The proposed method generates visually pleasant reflection removal results with faithful recovery, demonstrating that it capitalizes on the synergy of the auxiliary information from language descriptions, the iterative condition refinement strategy, and the multi-condition constraint mechanism.

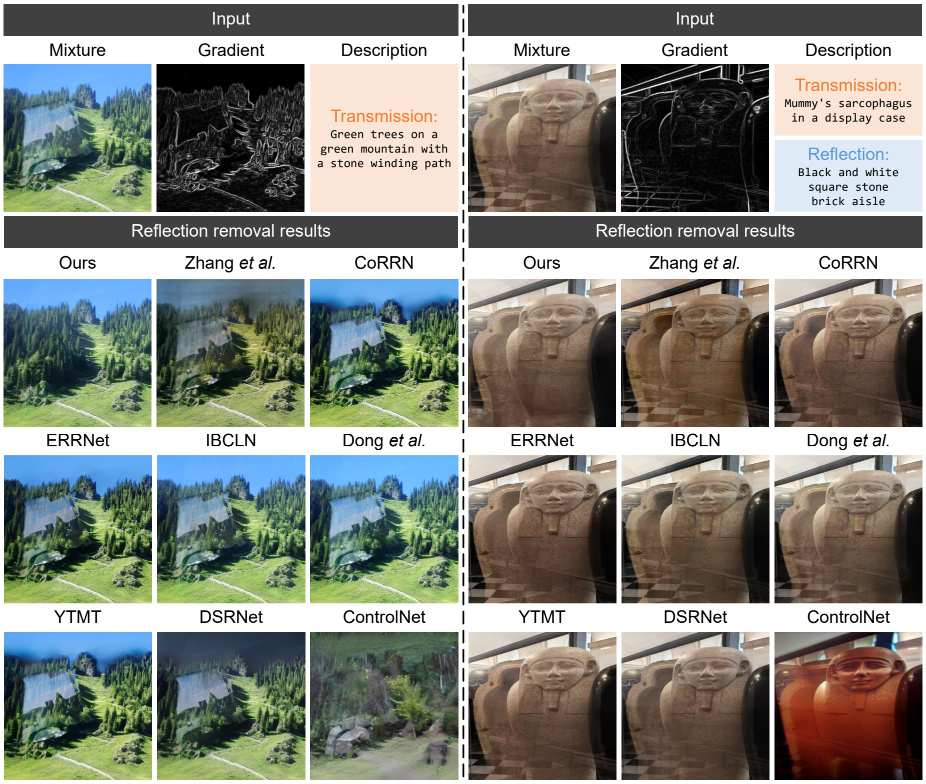

Fig. 5: Qualitative comparison of estimated transmission layers on real mixture images collected from the Internet, compared with several single-image methods [8,21,22,31,61,76] and a diffusion-based method ControlNet [74]. Please zoom in for details.

4.2 Ablation study

Ablation on control conditions[5]**.** We conduct ablation studies to investigate the effectiveness of the iterative condition refinement strategy and the multi-condition constraint mechanism. As shown in Table 2 and Fig. 6, most reflection contaminations are retained when missing iterative condition refinement (W/o rfn.), and disabling the refinement on color (W/o clr. rfn.) or structure (W/o str. rfn.) will cause inaccurate recovery, which exists color shift or retains the reflection content. In addition, as shown in Table 2 and Fig. 7, due to the generative prior of the diffusion model, the absence of the max value constraint (W/o MVC) or the saturation-aware structure constraint (W/o SSC), or both (W/o cnst.) leads to obvious color shifts, structure distortion, and uncontrolled recovery. Therefore, the degradation observed in the above ablations indicates the necessity of the refinement and constraints on conditions.

[5] More ablation studies are provided in the supplementary material.

Table 2: Ablation studies on the iterative condition refinement strategy and the multi-condition constraint mechanism. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

Metric	Abl. on iterative condition refinement			Abl. on multi-condition constraint			
	W/o clr. rfn.	W/o str. rfn.	W/o rfn.	W/o MVC	W/o SSC	W/o cnst.	Ours
PSNR↑	24.42	24.68	24.39	24.73	24.16	23.35	**25.08**
SSIM↑	0.898	0.896	0.891	0.897	0.887	0.879	**0.905**
LPIPS↓	0.132	0.136	0.139	0.129	0.143	0.154	**0.123**
NIQE↓	4.501	4.645	4.868	4.553	4.709	4.561	**4.322**
FID↓	49.13	53.71	60.12	51.06	56.46	63.68	**46.58**

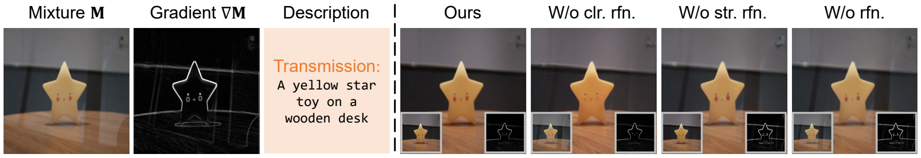

Fig. 6: The effect of the iterative condition refinement strategy. For each result, we place the corresponding final color and structure condition at its lower left and lower right. Please zoom in for details, respectively.

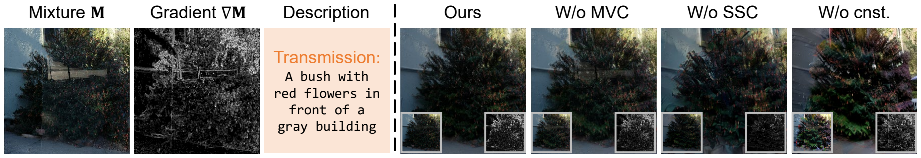

Fig. 7: The effect of the color and structure constraints in the multi-condition constraint mechanism. For each result, we place the corresponding final color and structure condition at its lower left and lower right, respectively. Please zoom in for details.

5 Conclusion

This paper introduces the first language-based diffusion model for single image reflection removal, which exploits auxiliary semantic information from language descriptions and generative priors from diffusion models to recover transmission layers from mixture images with low-transmitted reflection regions. An iterative condition refinement strategy is proposed to address the problem of inaccurate control conditions, and a multi-condition constraint mechanism is introduced to ensure recovery faithfulness. Experiments on real data demonstrate the effectiveness of the proposed method.

Limitations. The proposed method may incorrectly identify transmission and reflection layers if they exhibit similar image content. For such cases, more accurate spatial guidance from users such as annotated edges or region masks could help to relieve the ambiguity, which is left as our future work.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 62136001, 62088102.

References

1. Chang, Y., Jung, C., Sun, J.: Joint reflection removal and depth estimation from a single image. IEEE Transactions on Cybernetics (2020)
2. Chang, Y., Jung, C., Sun, J., Wang, F.: Siamese dense network for reflection removal with flash and no-flash image pairs. International Journal of Computer Vision (2020)
3. Chang, Z., Weng, S., Li, Y., Li, S., Shi, B.: L-CoDer: Language-based colorization with color-object decoupling transformer. In: Proc. of European Conference on Computer Vision (2022)
4. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In: Proc. of Advances in Neural Information Processing Systems (2023)
5. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-CoIns: Language-based colorization with instance awareness. In: Proc. of Computer Vision and Pattern Recognition (2023)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Diamant, Y., Schechner, Y.Y.: Overcoming visual reverberations. In: Proc. of Computer Vision and Pattern Recognition (2008)
8. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proc. of International Conference on Computer Vision (2021)
9. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proc. of International Conference on Computer Vision (2017)
10. Han, B.J., Sim, J.Y.: Zero-shot learning for reflection removal of single 360-degree image. In: Proc. of European Conference on Computer Vision (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proc. of International Conference on Computer Vision (2015)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Proc. of Advances in Neural Information Processing Systems (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proc. of Advances in Neural Information Processing Systems (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Hong, Y., Chang, Y., Liang, J., Ma, L., Huang, T., Shi, B.: Light flickering guided reflection removal. International Journal of Computer Vision (2024)

17. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with NIR and RGB image feature fusion. IEEE Transactions on Multimedia (2022)
18. Hong, Y., Lyu, Y., Li, S., Shi, B.: Near-infrared image guided reflection removal. In: Proc. of International Conference on Multimedia and Expo (2020)
19. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Panoramic image reflection removal. In: Proc. of Computer Vision and Pattern Recognition (2021)
20. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: PAR^{2}Net: End-to-end panoramic image reflection removal. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
21. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. Proc. of Advances in Neural Information Processing Systems (2021)
22. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proc. of International Conference on Computer Vision (2023)
23. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters (2008)
24. Kim, S., Huo, Y., Yoon, S.E.: Single image reflection removal with physically-based training images. In: Proc. of Computer Vision and Pattern Recognition (2020)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Kong, N., Tai, Y.W., Shin, S.Y.: A physically-based approach to reflection separation. In: Proc. of Computer Vision and Pattern Recognition (2012)
27. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proc. of Computer Vision and Pattern Recognition (2021)
28. Lei, C., Huang, X., Zhang, M., Yan, Q., Sun, W., Chen, Q.: Polarized reflection removal with perfect alignment in the wild. In: Proc. of Computer Vision and Pattern Recognition (2020)
29. Lei, C., Jiang, X., Chen, Q.: Robust reflection removal with flash-only cues in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
30. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(9), 1647–1654 (2007)
31. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proc. of Computer Vision and Pattern Recognition (2020)
32. Li, Y., Brown, M.S.: Exploiting reflection change for automatic reflection removal. In: Proc. of International Conference on Computer Vision (2013)
33. Li, Y., Brown, M.S.: Single image layer separation using relative smoothness. In: Proc. of Computer Vision and Pattern Recognition (2014)
34. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions. In: Proc. of Computer Vision and Pattern Recognition (2020)
35. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions with layered decomposition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
36. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3D-SPS: Single-stage 3D visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16454–16463 (June 2022)
37. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. In: Proc. of Advances in Neural Information Processing Systems (2019)

38. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Physics-guided reflection separation from a pair of unpolarized and polarized images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
39. Ma, D., Wan, R., Shi, B., Kot, A.C., Duan, L.Y.: Learning to jointly generate and separate reflections. In: Proc. of International Conference on Computer Vision (2019)
40. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: Proc. of International Conference on Learning Representations (2021)
41. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013)
42. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
43. Nam, S., Brubaker, M.A., Brown, M.S.: Neural image representations for multi-image fusion and layer separation. In: Proc. of European Conference on Computer Vision (2022)
44. Nayar, S.K., Fang, X.S., Boult, T.: Separation of reflection components using color and polarization. International Journal of Computer Vision (1997)
45. Park, J., Kim, H., Park, E., Sim, J.Y.: Fully-automatic reflection removal for 360-degree images. In: Proc. of Winter Conference on Applications of Computer Vision (2024)
46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Proc. of Advances in Neural Information Processing Systems (2019)
47. Qiu, J., Jiang, P.T., Zhu, Y., Yin, Z.X., Cheng, M.M., Ren, B.: Looking through the glass: Neural surface reconstruction against high specular reflections. In: Proc. of Computer Vision and Pattern Recognition (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. of International Conference on Machine Learning. PMLR (2021)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of Computer Vision and Pattern Recognition (2022)
50. Schechner, Y.Y., Kiryati, N., Basri, R.: Separation of transparent layers using focus. International Journal of Computer Vision (2000)
51. Shih, Y., Krishnan, D., Durand, F., Freeman, W.T.: Reflection removal using ghosting cues. In: Proc. of Computer Vision and Pattern Recognition (2015)
52. Simon, C., Kyu Park, I.: Reflection removal for in-vehicle black box videos. In: Proc. of Computer Vision and Pattern Recognition (2015)
53. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
54. Sun, H., Li, W., Liu, J., Chen, H., Pei, R., Zou, X., Yan, Y., Yang, Y.: CoSeR: Bridging image and language for cognitive super-resolution. arXiv preprint arXiv:2311.16512 (2023)
55. Sun, J., Weng, S., Chang, Z., Li, S., Shi, B.: UniCoRN: A unified conditional image repainting network. In: Proc. of Computer Vision and Pattern Recognition (2022)

56. Tang, J., Zhong, H., Weng, S., Shi, B.: LuminAIRe: Illumination-aware conditional image repainting for lighting-realistic generation. In: Proc. of Advances in Neural Information Processing Systems (2023)
57. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Gao, W., Kot, A.C.: Region-aware reflection removal with unified content and gradient priors. IEEE Transactions on Image Processing (2018)
58. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: Proc. of International Conference on Computer Vision (2017)
59. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: CRRN: Multi-scale guided concurrent reflection removal network. In: Proc. of Computer Vision and Pattern Recognition (2018)
60. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. International Journal of Computer Vision (2021)
61. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: CoRRN: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
62. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
63. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: CRIS: CLIP-driven referring image segmentation. In: Proc. of Computer Vision and Pattern Recognition (2022)
64. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers (2003)
65. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proc. of Computer Vision and Pattern Recognition (2019)
66. Wen, Q., Tan, Y., Qin, J., Liu, W., Han, G., He, S.: Single image reflection removal beyond linearity. In: Proc. of Computer Vision and Pattern Recognition (2019)
67. Weng, S., Li, W., Li, D., Jin, H., Shi, B.: MISC: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In: Proc. of Computer Vision and Pattern Recognition (2020)
68. Weng, S., Shi, B.: Conditional image repainting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
69. Weng, S., Wu, H., Chang, Z., Tang, J., Li, S., Shi, B.: L-CoDe: Language-based colorization using color-object decoupled conditions. In: Proc. of the AAAI Conference on Artificial Intelligence (2022)
70. Yang, J., Gong, D., Liu, L., Shi, Q.: Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In: Proc. of European Conference on Computer Vision (2018)
71. Yang, Y., Ma, W., Zheng, Y., Cai, J.F., Xu, W.: Fast single image reflection suppression via convex optimization. In: Proc. of Computer Vision and Pattern Recognition (2019)
72. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware vision transformer for referring image segmentation. In: Proc. of Computer Vision and Pattern Recognition (2022)
73. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (2014)

74. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. of International Conference on Computer Vision (2023)
75. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of Computer Vision and Pattern Recognition (2018)
76. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proc. of Computer Vision and Pattern Recognition (2018)
77. Zhang, Y.N., Shen, L., Li, Q.: Content and gradient model-driven deep network for single image reflection removal. In: Proc. of ACM International Conference on Multimedia (2022)
78. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Proc. of Advances in Neural Information Processing Systems (2024)
79. Zheng, Q., Chen, J., Lu, Z., Shi, B., Jiang, X., Yap, K.H., Duan, L.Y., Kot, A.C.: What does plate glass reveal about camera calibration? In: Proc. of Computer Vision and Pattern Recognition (2020)
80. Zheng, Q., Shi, B., Chen, J., Jiang, X., Duan, L.Y., Kot, A.C.: Single image reflection removal with absorption effect. In: Proc. of Computer Vision and Pattern Recognition (2021)
81. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. In: Proc. of Computer Vision and Pattern Recognition (2024)

L-DiffER: Single Image Reflection Removal with Language-based Diffusion Model (Supplementary Material)

Yuchen Hong[1,2 #] Haofeng Zhong[1,2,3 #] Shuchen Weng[1,2]
Jinxiu Liang[1,2] Boxin Shi[1,2,3 *]

[1] State Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology,
School of Computer Science, Peking University
[3] AI Innovation Center, School of Computer Science, Peking University
{hfzhong, shuchenweng, cssherryliang, shiboxin}@pku.edu.cn,
yuchenhong.cn@gmail.com

In the supplementary material, we introduce details about the time-variant coefficients in the proposed iterative condition refinement strategy, show evaluations on reflection recovery, and provide ablation studies on the network design, the saturation-aware structure constraint, loss functions, and language descriptions. We further conduct comparisons on the model size and inference time, alongside additional qualitative comparisons with state-of-the-art reflection removal methods.

6 Details about the time-variant coefficients

In this section, we provide details of time-variant coefficients β_t and γ_t ($t \in \{1, ..., T\}$) in the proposed iterative condition refinement strategy, controlling the refinement of the structure and color condition in Eq. (4) (corresponding to footnote 2 in the main paper). We set β_t and γ_t to the same value for synchronously updating the structure and color condition, which are defined as follows:

$$\beta_t = \gamma_t \coloneqq \prod_{i=1}^{t}(1 - \psi_i), \tag{17}$$

where the variants $\psi_i \in \Psi \coloneqq \{\psi_1, ..., \psi_T\}$ are constants increasing linearly from $\psi_1 = 10^{-4}$ to $\psi_T = 0.02$ inspired by [3], which is reasonable since these constants are small enough related to data scaled to $[-1, 1]$, guaranteeing that the forward and reverse processes exhibit nearly the same functional form.

7 Evaluations on reflection recovery

In this section, we conduct experiments to evaluate the recovery of reflection layers (corresponding to footnote 4 in the main paper). Quantitative comparisons

[#] Equal contributions. [*] Corresponding author.

Table 3: Comparison of quantitative results on real datasets for evaluating the recovery of reflection layers, compared with several state-of-the-art single-image reflection removal methods [2,9,10,12,17,21,23]. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

Dataset (size)	Metric	Method							
		Zhang *et al.*	CoRRN	ERRNet	IBCLN	Dong *et al.*	YTMT	DSRNet	Ours
SIR^2	PSNR↑	24.64	25.94	25.37	25.76	26.07	24.48	26.24	**26.45**
(500)	SSIM↑	0.486	0.513	0.506	0.502	0.519	0.479	0.521	**0.532**
Real20	PSNR↑	25.16	24.83	25.41	25.29	25.21	25.09	25.36	**25.51**
(20)	SSIM↑	0.506	0.487	0.519	0.508	0.532	0.515	0.529	**0.548**
Nature	PSNR↑	24.97	24.73	24.25	25.55	25.65	24.51	24.89	**25.88**
(20)	SSIM↑	0.511	0.497	0.542	0.552	0.568	0.458	0.561	**0.572**
Average	PSNR↑	24.67	25.85	25.33	25.73	26.02	24.50	26.16	**26.39**
(540)	SSIM↑	0.488	0.511	0.508	0.504	0.521	0.480	0.523	**0.534**

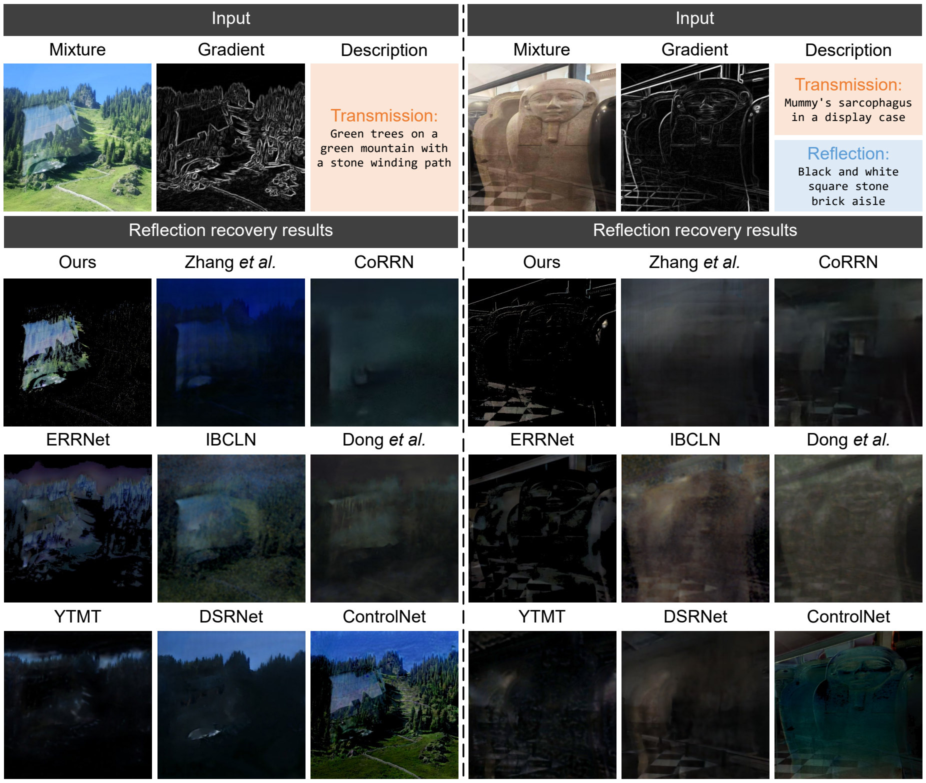

Fig. 8: Qualitative comparison of estimated reflection layers on real mixture images collected from the Internet, compared with several single-image methods [2,9,10,12,17,21,23] and a diffusion-based method ControlNet [22]. Please zoom in for details.

Table 4: Ablation studies on the network design and the loss function. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing transmission recovery results.

Metric	Abl. on network design		
	Both cond. enc.	Single trainable	Ours
PSNR↑	24.39	24.86	**25.08**
SSIM↑	0.889	0.896	**0.905**
LPIPS↓	0.141	0.135	**0.123**
NIQE↓	4.696	4.678	**4.322**
FID↓	57.12	52.48	**46.58**

are conducted on existing reflection removal datasets (*i.e.*, SIR^2 [18], Real20 [23], and Nature [12]), and ground truths of reflection layers are obtained as in [10]. We employ PSNR [11] and SSIM [20] as error metrics following [4–8,24]. Qualitative comparisons are conducted on images from the Internet. We compare the proposed method with state-of-the-art single-image methods (including Zhang *et al.* [23], CoRRN [17], ERRNet [21], IBCLN [12], Dong *et al.* [2], YTMT [9], and DSRNet [10]) and a language-based diffusion model ControlNet [22]. As ERRNet [21], ControlNet [22], and the proposed method only estimate the transmission layers $\widetilde{\mathbf{T}}$ from the mixture images $\mathbf{M}$ by networks, we obtain reflection layers $\widetilde{\mathbf{R}}$ by $\widetilde{\mathbf{R}} = \mathbf{M} - \widetilde{\mathbf{T}}$ following [15,17]. As quantitative and qualitative results shown in Table 3 and Fig. 8 (corresponding to Fig. 5 in the main paper), contributing to the high-fidelity and faithful recovery of transmission layers, the proposed method achieves the state-of-the-art performance on reflection recovery, which indicates the efficacy of the proposed iterative condition strategy and multi-condition constraint mechanism for leveraging the language-based diffusion model.

8 Additional ablation studies

In this section, we conduct ablation studies to investigate the effectiveness of the network design, the saturation-aware structure constraint, loss functions, and language descriptions (corresponding to footnote 5 in the main paper).

Ablation studies on the network design. As mentioned in Sec. 3.1 of the main paper, the proposed L-diffER utilizes a color encoder $\mathcal{E}^{\mathrm{c}}$ and a structure encoder $\mathcal{E}^{\mathrm{s}}$ to transform the color and structure condition into latent space, respectively. In practice, the color encoder $\mathcal{E}^{\mathrm{c}}$ uses the compression encoder of Stable Diffusion (SD) [13] and the structure encoder $\mathcal{E}^{\mathrm{s}}$ uses the condition encoder of ControlNet [1,22]. We conduct an ablation study by replacing $\mathcal{E}^{\mathrm{c}}$ with the condition encoder of ControlNet [22] (denoted as 'Both cond. enc.') that is also the network architecture of $\mathcal{E}^{\mathrm{s}}$ to investigate the influence of encoders. Furthermore, following ControlNet [22], two trainable copied modules (denoted by the pink blocks in Fig. 2 of the main paper) of SD [13] are employed to separately extract color and structure features from color and structure latents, so we conduct another ablation study by only using a single trainable copied module to jointly extract color and structure features (denoted as 'Single trainable'). As

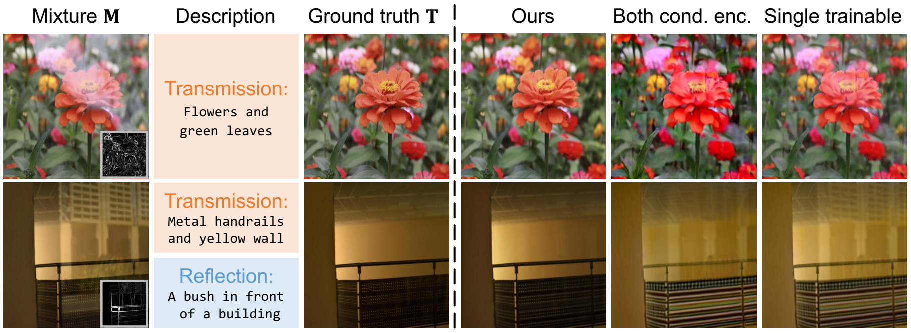

Fig. 9: Ablation study on the network design. We show the gradient map (*i.e.*, the original structure condition) at the lower right of each mixture image (*i.e.*, the original color condition). Please zoom in for details.

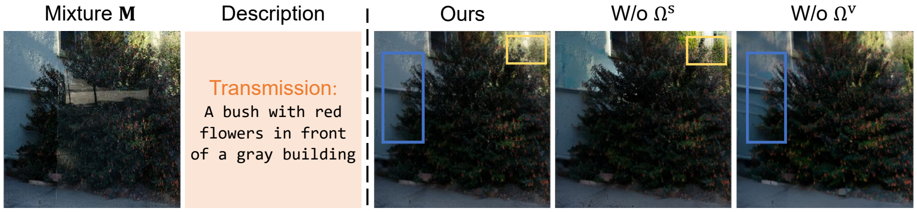

Fig. 10: The effect of $\mathbf{\Omega}^{\mathrm{s}}$ and $\mathbf{\Omega}^{\mathrm{v}}$. Please zoom in for details.

Table 5: Ablation studies on pixel loss functions. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing transmission recovery results.

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
W/o $\mathcal{L}_{\mathrm{RGB}}$	24.87	0.896	0.137	4.702	52.32
W/o $\mathcal{L}_{\mathrm{grad}}$	24.68	0.894	0.140	4.813	53.47
W/o $\mathcal{L}_{\mathrm{num}}$	24.97	0.900	0.131	4.593	48.64
W/o $\mathcal{L}_{\mathrm{pix}}$	24.22	0.884	0.149	4.751	58.21
Ours	**25.08**	**0.905**	**0.123**	**4.322**	**46.58**

results shown in Table 4 and Fig. 9, modifications on the encoder or the trainable copied module lead to performance degradation, indicating the effectiveness of our network design for condition injection.

Ablation studies on the saturation-aware structure constraint. As mentioned in the main paper, $\mathbf{\Omega}^{\mathrm{s}}$ in Eq. (7) is designed to indicate saturated regions for retaining the generative capability of the diffusion model in these regions, and $\mathbf{\Omega}^{\mathrm{v}}$ in Eq. (8) is to indicate valid edges for preventing overly generation in non-saturated regions. We conduct ablation studies shown in Fig. 10 by removing $\mathbf{\Omega}^{\mathrm{s}}$ and $\mathbf{\Omega}^{\mathrm{v}}$, respectively, which verifies the effectiveness of the two components.

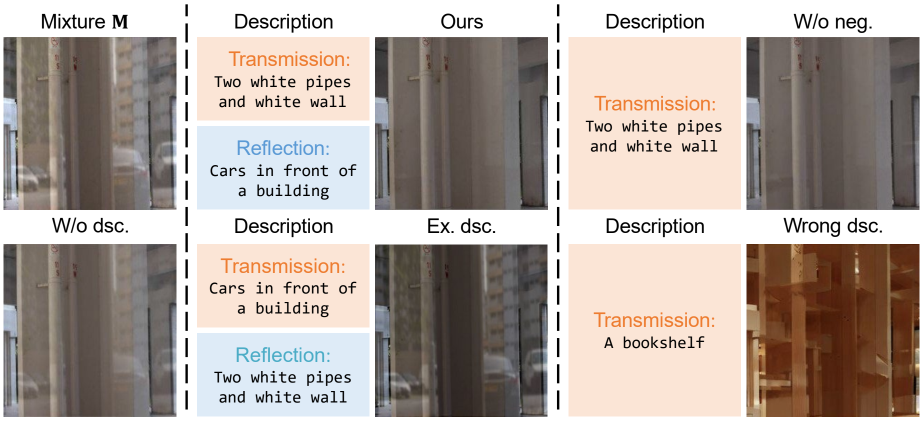

Fig. 11: The effect of input language descriptions. Please zoom in for details.

Ablation studies on loss functions. We further investigate the effect of the loss functions (introduced in Sec. 3.4 of the main paper) as shown in Table 5. Compared with the complete model, disabling the RGB loss (‘W/o $\mathcal{L}_{\text{RGB}}$’) and gradient loss (‘W/o $\mathcal{L}_{\text{grad}}$’) degrade performance significantly since inaccurate conditions mislead transmission recovery. Disabling the numerical loss (‘W/o $\mathcal{L}_{\text{num}}$’) also causes performance decline, suggesting the efficacy of the supervision on images’ mean and variance. Besides, the performance reduction of the variant ‘W/o $\mathcal{L}_{\text{pix}}$’ demonstrates the necessity of conducting supervision on the pseudo transmission layers $\mathbf{T}_{0|t}$ at the pixel level. Note that we do not disable $\mathcal{L}_{\text{ldm}}$ since diffusion models can not train without it.

Ablation studies on language descriptions. We conduct ablation studies on the numbers and the order of input language descriptions to verify the effectiveness of language guidance. As shown in Fig. 11, abandoning the description of the reflection layer (W/o neg.) causes a few reflection residuals in the example, and more reflections remain if directly using an empty description (W/o dsc.). If exchanging the order of prompts (Ex. dsc.), results will degrade due to different statistics of transmission and reflection layers [14,16], but the contents of recovered results still conform to the prompt. Similarly, using the wrong description of the transmission layer (Wrong dsc.) will cause a completely different recovered result. By utilizing both language descriptions of two layers, the proposed method achieves high-fidelity reflection removal, which indicates the efficacy of introducing language descriptions.

9 Comparisons of the model size and inference time

In this section, we show the comparisons of the model size and inference time in Table 6. The image size is 384 × 384, and we run the inference on an NVIDIA GeForce RTX 3090. Though owning more parameters and more inference time

Table 6: Comparisons of the model size and inference time. We show both trainable and total parameters for diffusion-based methods.

Metric	Single-image				Diffusion-based	
	CoRRN [57]	IBCLN [30]	YTMT [20]	DSRNet [21]	ControlNet [66]	Ours
Params	59.5M	21.6M	73.4M	137.6M	364.4M/1.4B	728.8M/1.8B
Time (s)	0.079	0.127	0.206	0.361	11.547	12.685

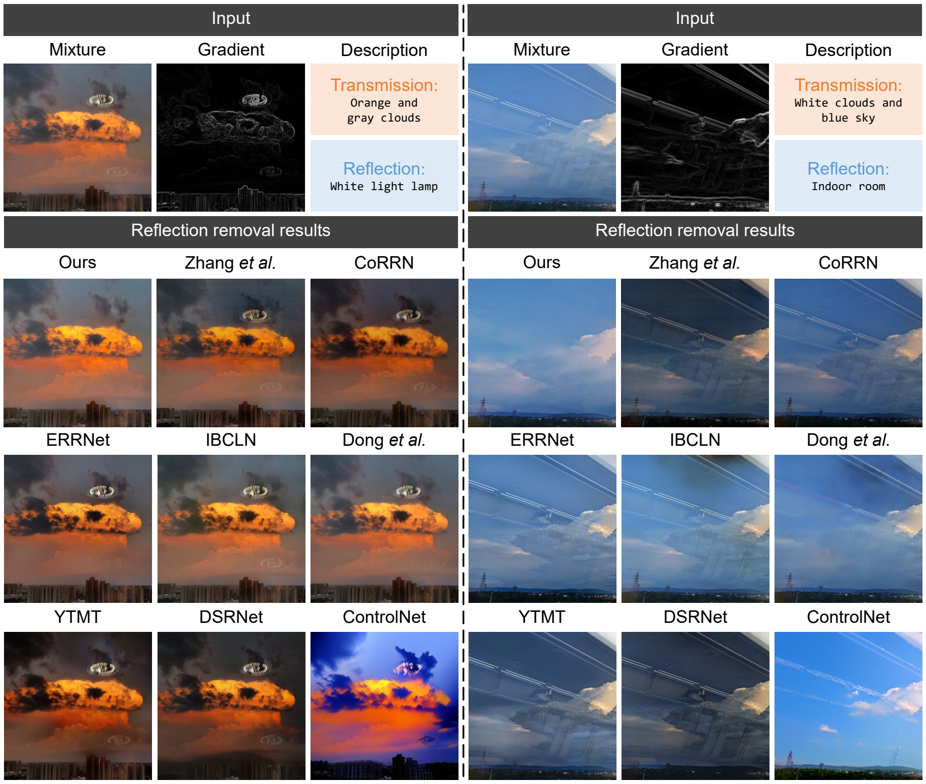

Fig. 12: Qualitative comparison of estimated transmission layers on real mixture images collected from the Internet, compared with several single-image methods [2,9,10,12,17,23] and a diffusion-based method ControlNet [22]. Please zoom in for details.

than traditional single-image methods, it is worth noting that the proposed method pioneers in introducing large language-based diffusion models for reflection removal to solve the most concerned problems, *i.e.*, relieving the ill-posedness and tackling low-transmitted or saturated reflections, which facilitates future research in a new perspective.

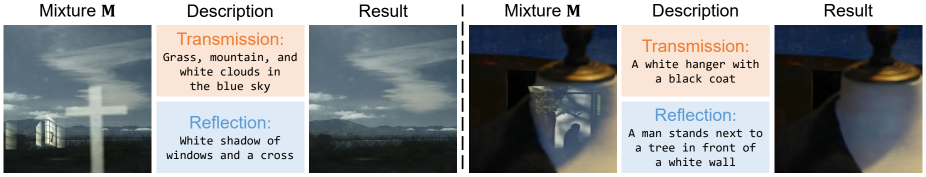
Mixture **M**
Description
Result
Mixture **M**
Description
Result
Transmission:
Grass, mountain, and white clouds in the blue sky
Reflection:
White shadow of windows and a cross
Transmission:
A white hanger with a black coat
Reflection:
A man stands next to a tree in front of a white wall

Fig. 13: High-resolution real examples with low-transmitted or saturated reflections. Please zoom in for details.

10 Additional qualitative results

In this section, additional qualitative experiments are conducted on real images to show the effectiveness of the proposed language-based diffusion model for reflection removal. Experimental settings are the same as in Sec. 4.1 of the main paper, and qualitative results are shown in Fig. 12. As can be observed, single-image reflection removal methods [2, 9, 10, 12, 17, 21, 23] fail in low-transmitted reflection regions, and ControlNet [22] generates results with color shifts and structure distortions. The proposed method thoroughly removes reflections and recovers clear transmission layers even in low-transmitted and saturated reflection regions (*e.g.*, the reflections of white lights in the left and right example of Fig. 12), which demonstrates the effectiveness of using language descriptions to provide auxiliary semantic information and the unique advantage of leveraging generative priors from diffusion models [13] by employing the proposed refinement strategy and constraint mechanism for conditions.

We further present two real examples with low-transmitted or saturated reflections. By adopting a patch aggregation method [19], we can obtain their corresponding high-resolution (*i.e.*, 1024 × 1024) results shown in Fig. 13, which verifies the robustness of the proposed method.

References

1. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-cad: Language-based colorization with any-level descriptions using diffusion priors. In: Proc. of Advances in Neural Information Processing Systems (2023)
2. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proc. of International Conference on Computer Vision (2021)
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proc. of Advances in Neural Information Processing Systems (2020)
4. Hong, Y., Chang, Y., Liang, J., Ma, L., Huang, T., Shi, B.: Light flickering guided reflection removal. International Journal of Computer Vision (2024)
5. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with nir and rgb image feature fusion. IEEE Transactions on Multimedia (2022)
6. Hong, Y., Lyu, Y., Li, S., Shi, B.: Near-infrared image guided reflection removal. In: Proc. of International Conference on Multimedia and Expo (2020)

7. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Panoramic image reflection removal. In: Proc. of Computer Vision and Pattern Recognition (2021)
8. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: PAR^{2}Net: End-to-end panoramic image reflection removal. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
9. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. Proc. of Advances in Neural Information Processing Systems (2021)
10. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proc. of International Conference on Computer Vision (2023)
11. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters (2008)
12. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proc. of Computer Vision and Pattern Recognition (2020)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of Computer Vision and Pattern Recognition (2022)
14. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Gao, W., Kot, A.C.: Region-aware reflection removal with unified content and gradient priors. IEEE Transactions on Image Processing (2018)
15. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: CRRN: Multi-scale guided concurrent reflection removal network. In: Proc. of Computer Vision and Pattern Recognition (2018)
16. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. International Journal of Computer Vision (2021)
17. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: CoRRN: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
18. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
19. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution (2024)
20. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers (2003)
21. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proc. of Computer Vision and Pattern Recognition (2019)
22. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. of International Conference on Computer Vision (2023)
23. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proc. of Computer Vision and Pattern Recognition (2018)
24. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. arXiv preprint arXiv:2402.11874 (2024)