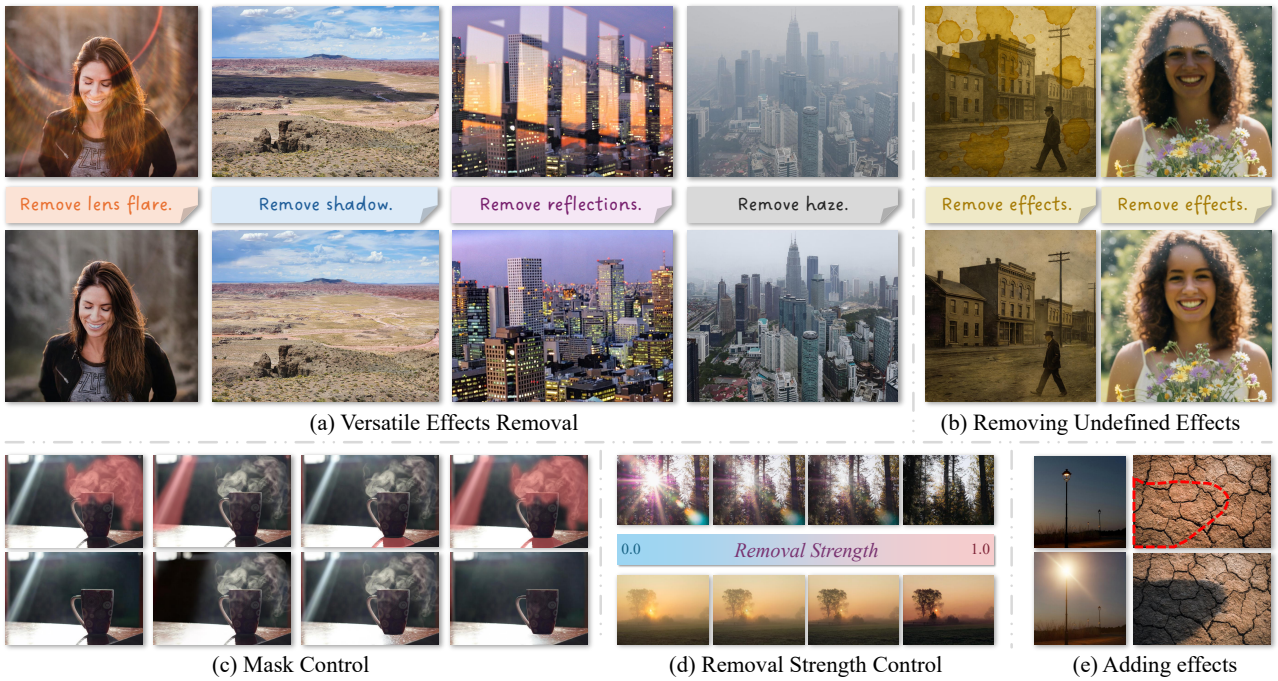
# UniSER: A Foundation Model for Unified Soft Effects Removal

Jingdong Zhang[1,2] Lingzhi Zhang[2] Qing Liu[2] Mang Tik Chiu[2] Connelly Barnes[2]
Yizhou Wang[2] Haoran You[2] Xiaoyang Liu[2] Yuqian Zhou[2] Zhe Lin[2]
Eli Shechtman[2] Sohrab Amirghodsi[2] Xin Li[1] Wenping Wang[1] Xiaohang Zhan[2,†]
[1]Texas A&M University [2]Adobe Research
{jdzhang, xinli, wenping}@tamu.edu, {lingzzha, xzhan}@adobe.com




(a) Versatile Effects Removal (b) Removing Undefined Effects

(c) Mask Control (d) Removal Strength Control (e) Adding effects

Figure 1. Our UniSER eliminates multiple challenging (a) and even undefined (b) soft effects from in-the-wild images while preserving background identities. Besides, UniSER supports precise pixel mask control (c), and removal strength control (d), allowing for intuitive and fine-grained restoration tailored to specific user needs. The framework is also capable of adding effects in the given region (e). Masks are global by default if not shown.

## Abstract

*Digital images are often degraded by soft effects such as lens flare, haze, shadows, and reflections, which reduce aesthetics even though the underlying pixels remain partially visible. The prevailing works address these degradations in isolation, developing highly specialized, specialist models that lack scalability and fail to exploit the shared underlying essences of these restoration problems. Meanwhile, although recent large-scale generalist models (e.g., GPT-4o, Flux Kontext, Nano Banana) offer powerful text-driven editing capabilities, they heavily rely on detailed prompts and often fail to achieve robust removal on such fine-grained tasks while preserving the scene's identity. Leveraging the common essence of soft effects, i.e., semi-transparent occlusions, we introduce a foundational versatile model UniSER, capable of addressing diverse degradations caused by soft effects within a single framework. Our methodology centers on curating a massive 3.8M-pair dataset to ensure robustness and generalization, which includes novel, physically-plausible data to fill critical gaps in public benchmarks, and a tailored training pipeline that fine-tunes a Diffusion Transformer to learn robust restoration priors from this diverse data, integrating fine-grained mask and strength controls. This synergistic approach allows UniSER to significantly outperform both specialist and generalist models,*

† indicates project leader.

*achieving robust, high-fidelity restoration in the wild.*

## 1. Introduction

Images captured in real-world environments inevitably suffer from degradations. A common class of such "soft" effects includes optical phenomena (e.g., lens flare, reflections) and atmospheric conditions (e.g., haze, fog). These effects corrupt scene radiance additively or multiplicatively, degrading contrast, color fidelity, and fine details [57, 85]. Consequently, image quality and visibility are compromised, and in severe cases, occlusions cause irreversible information loss, rendering recovery fundamentally ill-posed [37, 57, 91].

To restore image structures, most existing works address each degradation type separately. For instance, dehazing has progressed from prior-based methods such as the Dark Channel Prior (DCP) [37] to deep networks estimating scattering parameters or directly predicting clean images [16, 19, 29, 60, 77]. Similarly, shadow, flare, and reflection removal adopt task-specific designs [27, 57, 85, 91, 94, 118], relying on physical modeling, layer decomposition, or elaborate data and network strategies to mitigate ill-posedness. While such methods achieve strong task-specific performance, recent works [17, 63, 70] attempt to unify multiple degradations within one framework. Yet these models remain limited in scalability and robustness when facing extreme, diverse real-world conditions. This motivates the development of foundation models trained on large-scale data to achieve stronger generalization and resilience in the wild.

Concurrently, the rise of powerful foundation models like GPT-4o [46] and Nano Banana (Gemini 2.5 Flash Image) [21, 33] has introduced general-purpose, text-driven image generation/editing based on Multi-modal Large Language Models (MLLMs). These models can interpret complex prompts and perform realistic edits. However, for fine-grained tasks like soft effect removal, they exhibit significant limitations. Their performance is often unstable and heavily reliant on meticulously crafted text prompts. More critically, they lack the precise, pixel-wise control required for high-fidelity restoration and identity preservation. Treating soft effect removal as a general inpainting task leads to the alteration of local image structures or the identity of objects in the scene, which makes them unreliable for professional photo editing and critical computer vision pipelines.

Despite their diverse appearances, effects such as lens flare, haze, reflections, and shadows share the same intrinsic property: they are semi-transparent occlusions that degrade the image, but do not fully destroy the underlying scene identity. This shared property unifies them as a challenging decomposition problem (e.g. [32, 114–116] use dehaze models as strong comparative baselines for lens flares removal). To this end, we define a unified and extensible task, termed Soft Effects Removal (SER) to invert all these diverse degradation processes. This task is highly challenging. First, these effects are typically entangled with the scene itself, rather than merely superimposed as simple overlays. Second, the local image structures, and even pixel-level identities, should be precisely preserved. Third, regions that are fully occluded or invisible (*e.g.*, overexposed areas in lens flare or areas covered by extremely dense haze) must be plausibly reconstructed.

To effectively tackle these challenges, we introduce **UniSER** (Fig. 1 (a) & (b)), a data-centric versatile model for Soft Effects Removal. Our method is built upon two key points. First, we curated a large-scale dataset of approximately 3.8M balanced, high-quality, pixel-aligned image pairs. By unifying existing open-source datasets and augmenting them with extra real-world and synthetic data, we provide the precise supervision our model needs to learn content invariance. Second, as shown in Fig. 1 (c) & (d), we implemented fine-grained user controls, including pixel-level masks to define the removal area and strength levels to modulate the removal strength, making the process highly controllable. Beyond restoration, UniSER can also perform aesthetic edits, such as enhancing existing effects or generating new, realistic ones on clean images (Fig. 1 (e)). Our method achieves state-of-the-art results on multiple public benchmarks and demonstrates significantly better generalization on in-the-wild testing data.

In summary, our main contributions can be summarized as follows:

- **A Large-Scale Dataset for Generalization:** We curated a large-scale dataset of$\sim$3.8M image pairs, providing vast data distribution for strong generalization on challenging in-the-wild data.
- **A Versatile SER Model:** Trained on the curated dataset, a foundational versatile model UniSER achieves removing multiple challenging soft effects in the wild with state-of-the-art performance and surpasses much larger general-purpose models such as Nano Banana.
- **Controllable Editing:** Developed fine-grained user controls for SER tasks, including spatial masks and strength levels, to enable precise and controllable effect removal.

## 2. Related Work

### 2.1. Isolated Effects Removal

**Lens flare removal.** Previous learning-based methods improved data synthesis by considering camera ISP to enhance realism and generalization [115, 116]. Concurrently, architectural innovations emerged, including self-supervised methods to disentangle co-occurring flares [39], while others explicitly separated light source preservation from flare removal using dedicated detection modules [32], and networks leveraging both spatial and frequency domains [83].

More recently, large pretrained Latent Diffusion Models (LDMs) are adpated to leverage their powerful generative priors [114]. The development of these methods has also been heavily reliant on specialized datasets, from semi-synthetic ones [91], Flare7K [24], to real-world paired datasets [26].

**Reflection removal.** Early methods for single-image reflection removal (SIRR) focused on iterative refinement using edge maps [30] or recurrent networks [62, 96]. Subsequent research shifted towards improving training data realism by learning non-linear blending [90], employing physically-based rendering [52], and modeling glass absorption [113]. Architectural innovations followed, introducing location-aware modules [28] and advanced attention mechanisms [45, 106] to better distinguish between layers. More recent paradigms reduce reliance on paired data through unsupervised deep image priors [72] or by using Diffusion Models to generate guiding prompts [88]. This progress has been underpinned by the creation of key real-world benchmarks like $SIR^2$ [85] and the large-scale RRW dataset [118].

**Shadow removal.** Initial approaches to shadow removal relied on traditional physical priors and optimization frameworks [36, 105]. The advent of deep learning introduced end-to-end models like DeshadowNet [71] and methods that decomposed images into shadow-free and matte layers [57]. Subsequent architectural advancements included using Generative Adversarial Networks (GANs) for joint detection and removal [86], fusing synthetic exposure pairs [31], and learning via shadow generation [67]. More recent trends focus on eliminating the dependency on explicit shadow masks, utilizing mask-free transformers [27] or reformulating the problem as a dense prediction task [64]. The progress in this field has been propelled by benchmarks like SRD [71], ISTD [86], and the newer high-resolution WSRD dataset [84].

**Haze removal.** Single-image dehazing evolved from early methods based on statistical priors like the Dark Channel Prior (DCP) [37] to data-driven deep learning. Initial deep learning works included lightweight end-to-end networks [60], hybrid models that learned priors for traditional optimization [95], and unpaired training with GANs to address data scarcity [29]. Architectural innovations, such as gated context aggregation [16] and Vision Transformers [77], were later introduced to better handle non-uniform haze. Recent efforts focus on closing the synthetic-to-real domain gap by generating more physically plausible training data [19] or leveraging diffusion models for realistic haze synthesis [87]. This progress has been consistently driven by the development of comprehensive benchmarks [47, 61, 107].

Apart from them, some works delve into All-In-One (AIO) methods to restore image quality from multiple degradations within a multi-task model [17, 23, 48, 63, 66, 70, 73, 80, 111]. Despite the achievements from all these methods, key challenges persist including the limited diversity in datasets, while current methods still struggle with scalable training with robust generalization abilities, as well as handling more challenging types of challenging soft effects requiring semantic-awareness.

### 2.2. Prompt-based Image Editing

Prompt-based image editing originated from diffusion models, enabled by deterministic inversion techniques like DDIM [76] that map real images to an editable latent space. Initial methods controlled edits by manipulating internal model structures, such as altering cross-attention maps to preserve layout [40] or fine-tuning the entire model on a single image for complex, non-rigid changes [50]. The field has since evolved towards more direct user control, with models trained to follow natural language instructions [12] or allow for interactive, point-based spatial adjustments [74]. This shift towards more precise, semantic editing is increasingly powered by the advanced contextual understanding of Multimodal Large Language Models (MLLMs) [10, 21, 46]. However, current approaches still often lack fine-grained pixel control and can struggle to perfectly preserve the subject's identity during transformation.

## 3. Methodology

### 3.1. Data Curation

A powerful foundation model requires large-scale, high-quality, and diverse training data. To equip UniSER with robust generalization, we curated a comprehensive dataset by unifying pixel-aligned image pairs from four representative tasks: lens flare, shadow, haze, and reflection removal. This integration enables the model to learn a broad restoration representation while preserving content identity.

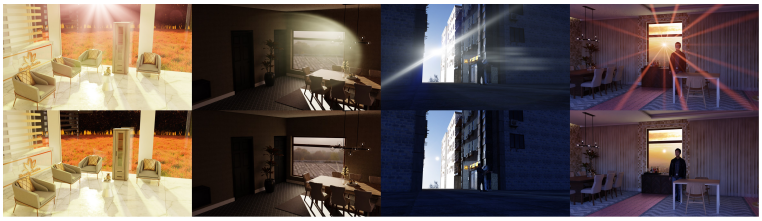
**Public datasets.** We incorporate multiple benchmark datasets spanning the four domains (see Table 1 and supplementary materials for details). Despite their usefulness, these datasets exhibit imbalance, such as the scarcity of large-scale flare removal data and limited diversity in haze scenarios.

**Data expansion.** To remedy these gaps and increase data volume, we expand training data through three sources: real-world captures, 2D synthesis, and 3D rendering.

- *Lens flare.* The key bottleneck lies in insufficient data. We therefore construct 78 indoor and outdoor 3D scenes in Blender [11], rendering about 70K paired images, named HALO dataset. Unlike Flare7K [24], which overlays flare layers on clean images, our rendered data produces geometrically consistent and realistic flare effects. The dataset covers diverse flare patterns, including reflective flare, glare, shimmer, and streaks.

Table 1. Summary of datasets curated for UniSER training. "†" represents the datasets curated by us, "*" represents the datasets which we re-synthesis effects with our own algorithm.

| Task | Dataset | Type | Description | Pairs |
|---|---|---|---|---|
| Lens flare | FlareReal600 [26] | Real-World | Nighttime flares, Streetview, Cityscapes, Outdoor | 0.6k |
| | HALO† | 3D Synthetic | Rendered, Various flares and scenes, Indoor & Outdoor | 70k |
| Shadow | WSRD+ [84] | Real-World | Object-level, Close-view, Rich texture, Complex shadows | 1k |
| | ISTD+ [86] | Real-World | Simple-shaped shadows, Monotonous scenes, Outdoor | 1.3k |
| | SRD [71] | Real-World | Various scenes, Outdoor | 2.6k |
| | LR-SRD† | Real-World | Object-level, Close-view, Hard & Soft shadow, Indoor & Outdoor | 26k |
| Haze | Haze-R [1–7] | Real-World | Collection including: I-HAZE, O-HAZE, Dense-Haze, NH-Haze, etc., Homogeneous & Non-Homogeneous, Indoor & Outdoor | 0.3k |
| | REVIDE [110] | Real-World | Video Frames, Indoor | 1.9k |
| | LM-Haze [107] | Real-World | Multi-level haze, Homogeneous, Indoor | 5k |
| | HAZESPACE* [47] | 2D Synthetic | Multi-level haze, Vast range of scenes, Outdoor | 24×70k |
| | RESIDE* [61] | 2D Synthetic | Multi-level haze, Indoor & Outdoor | 290k |
| | SYN-HAZE* | 2D Synthetic | Multi-level haze, Synthetic scenes, Include extremely dense haze, Indoor & Outdoor | 24×70k |
| Reflection | RRW [118] | Real-World | Various scenes, Diverse glass and reflection types | 14.9k |
| | POLAR-RR [59] | Real-World | Polarization-based, Indoor | 0.8k |
| | RFC [58] | Real-World | Flash-induced reflections | 5k |
| | BDN [96] | 2D Synthetic | Linearly Blended, Public Image Sources | 50k |



HALO

LR-SRD

Synthetic Haze

Figure 2. Visualization of our curated data samples. For LR-SRD we apply instance-level shadow mask to each shadow region.

- *Shadow.* While public datasets cover both indoor and outdoor scenes, they contain only ∼5K pairs. To scale up, we add an additional 26K photo pairs. Specifically, we repurpose internal object-effect removal data: by stitching objects without shadows into background images, we synthesize corresponding shadow-free versions to form the Large Real-world Shadow Removal Dataset (LR-SRD).
- *Haze.* Existing synthetic datasets (RESIDE, HAZESPACE) often appear uniform or algorithmically simplistic. To generate more realistic and challenging cases, we use their clean ground-truth images with monocular depth [51], and apply a physically motivated atmospheric rendering pipeline. This allows precise control of parameters such as visibility, airlight color, scatter, and optical thickness. To simulate non-homogeneous haze or fog, we introduce procedural noise fields and path blurring, yielding realistic textures of haze, smoke, and fog. More synthesis details are provided in the supplementary material.

These expanded datasets extend coverage to underrepresented scenarios, enhancing UniSER's robustness in the wild. A detailed breakdown is given in Table 1, with representative samples in Fig. 2.

### 3.2. Framework

As shown in Fig. 3, UniSER is a unified framework designed to tackle multiple soft effect removal tasks. Inspired by UniReal [18], the core architecture reformulates these diverse tasks as a problem of *discontinuous frame generation* within a latent diffusion model. The process begins with a Variational Autoencoder (VAE) [53] encoding the input image into a compact latent space, while a text encoder processes a task-specific prompt (*e.g.*, "*remove haze*") to generate instructive embeddings. These conditional inputs (image latent and textual embeddings) are then concatenated

The datasets will be released at: https://github.com/Evergreen0929/UniSER-Datasets

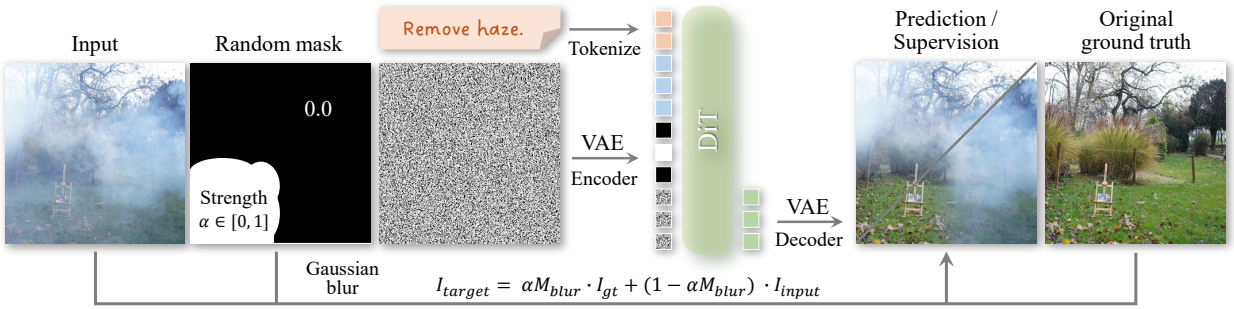



Figure 3. The architecture of UniSER. During training, the mask is randomly synthesized along with a scalar strength, and the supervision is composed by the input image and the original ground truth via the mask and the strength.

with the noisy target latent and fed as a sequence to a Diffusion Transformer (DiT). The DiT's full attention mechanism operates on this sequence, allowing it to iteratively predict and remove noise from the target latent by conditioning on both the visual context and the textual instructions. Finally, the fully denoised latent is passed through the VAE decoder to reconstruct the final, effect-free image. The model is trained using a mean squared error (MSE) loss between the predicted noise and the ground truth noise, with a timestep-dependent weighting scheme to balance the contributions of different noise levels.

**Random Masking Strategy.** As established in the framework, a mask can be supplied as a condition to guide the denoising process toward a specific spatial region. However, most of the training sets do not contain the mask of effects. To ensure the model can robustly handle any user-provided mask shape, we adopt a random masking strategy. During training, following [78, 112] we synthesize a wide variety of binary masks $M$ by randomly combining geometric primitives like rectangles with free-form, stroke-like patterns that simulate user brush strokes. Afterwards, providing pairs $\{I_{input}, I_{gt}\}$, we generate the corresponding training supervision $I_{target}$ where the effect is removed only within the masked region via simply compositing $I_{input}$ and $I_{gt}$ with the mask, as shown in Equation 1. Note that the regions of effects in the input image are unavailable, hence the masks do not necessarily cover them. In this way the behaviors the model to learn is summarized as following:

- Region inside the mask w/ effects: remove effects based on the strength;
- Region inside the mask w/o effects: keep identical;
- Region outside the mask: keep identical.

Additionally, to make the supervision natural-looking, we blur the mask boundary via dilation and Gaussian blur. This strategy exposes the model to a vast distribution of possible mask shapes, enhancing its generalization capability for arbitrary user edits, and removing the sepcific effect regions.

**Removal Strength Control.** Beyond specifying *where* to remove an effect, UniSER allows users to control *how much* of the effect is removed. This is achieved by training the model to interpret continuous values in the conditional mask as an indicator of removal intensity. During the training process, for each sample, we uniformly sample a floating-point scalar value to represent "strength", denoted as $\alpha \in [0, 1]$. Instead of conditioning the model on a binary mask $M$, we provide a soft value mask $\alpha M$. The model thus learns to associate a mask value of 1.0 with complete removal, 0.0 with no change, and intermediate values with partial removal. On the other hand, along with the aforementioned blurred mask, the training target is generated by linearly interpolating between the clean ground truth ($I_{gt}$) and the input with effects ($I_{input}$) using the randomly sampled $\alpha$. Formally, the supervision during training is computed as following:

$$I_{target} = \alpha M_{blur} \cdot I_{gt} + (1 - \alpha M_{blur}) \cdot I_{input} \quad (1)$$

This joint strategy of conditioning on a soft mask while generating a correspondingly blended target enables the model to learn a continuous and intuitive mapping from the control signal to the desired degree of effect removal.

**Handling Undefined Effects.** Our framework also extends to zero-shot generalization on unseen soft effects through two complementary fine-tuning strategies. First, we randomly replace task-specific prompts with a generic prompt *"remove effects"*, encouraging the model to capture a shared notion of removal across tasks. Second, we introduce an auxiliary task using clean images: random masks are generated and overlaid with semi-transparent or opaque regions to synthesize degraded inputs, which are trained exclusively with the generic prompt. This prevents overfitting to predefined effect categories and compels the model to learn the broader concept of removing arbitrary occlusions, thereby enabling generalized restoration.

**Adding & Enhancing Effects.** We can easily invert the removal task to adding or enhancing effects by swapping the roles of the input and the target. Similarly, the adding or enhancing ability is controlled by the mask and strength given by users. We demonstrate this ability in Fig. 7.

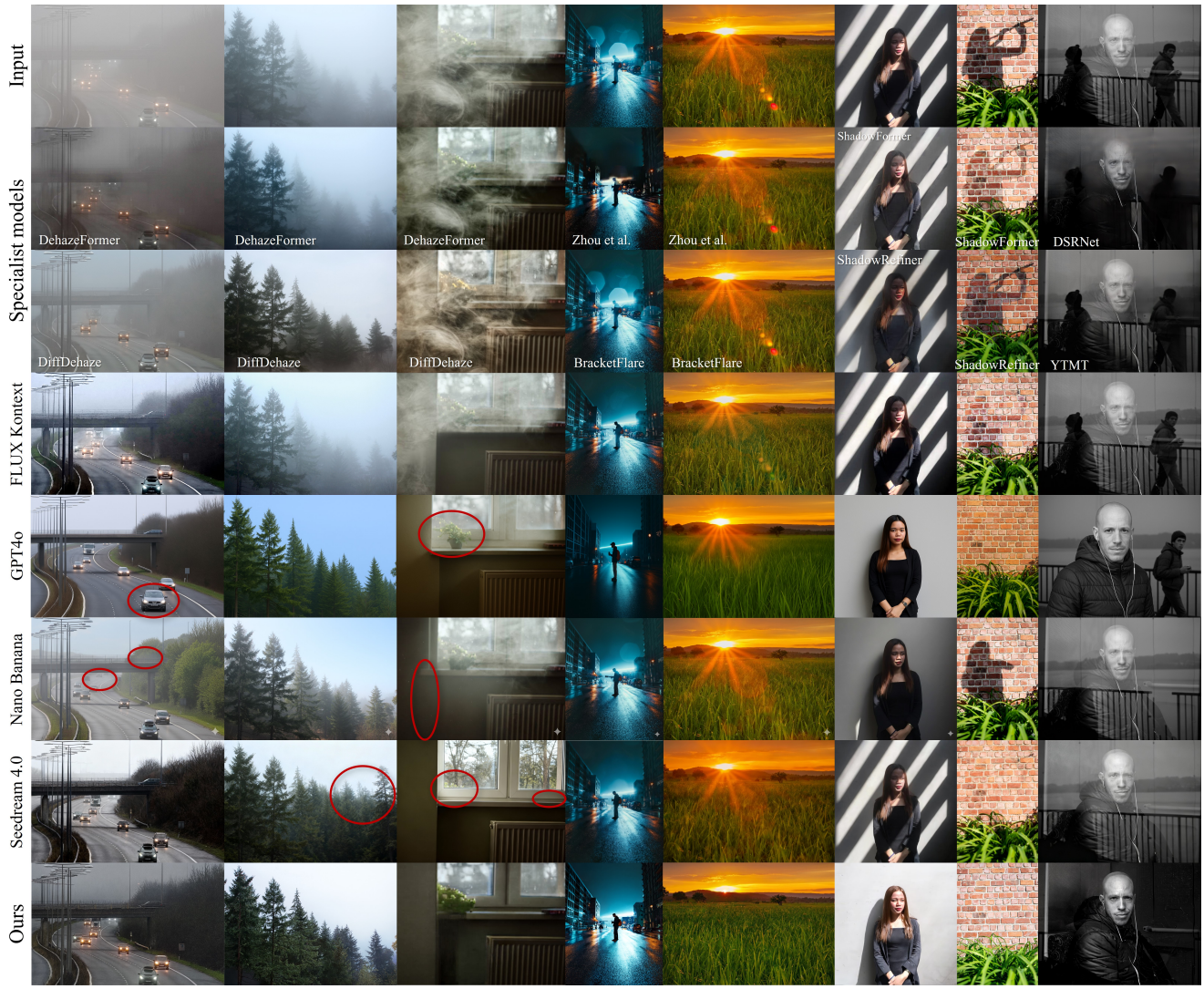



Figure 4. Comparisons with state-of-the-art specialist and generalist models on in-the-wild testing data. For effect removal, our method significantly outperforms these baselines. Moreover, generalist models fail to preserve the identity of background objects, some of the discrepancies are circled, better view by zooming in.

# 4. Experiments

## 4.1. Benchmarks and Baselines

**Benchmarks.** We evaluate UniSER across four soft-effect tasks on widely used benchmarks. For *lens flare removal*, we adopt the Flare7K real-world test set [24]. For *shadow removal*, we test on SRD [71], ISTD+ [86], and the high-resolution WSRD+ [84]. For *haze removal*, we use the SOTS and HSTS subsets of RESIDE [61]. For *reflection removal*, we employ $SIR^2$ [85] and the Nature test set [62]. UniSER is fine-tuned on the training splits of these datasets for domain adaptation. Evaluation uses standard full-reference metrics: PSNR and SSIM.

To assess real-world robustness, we collected 39 in-the-wild images containing haze, fog, flare, reflection, and shadow. As no ground truth is available, we report reference-free metrics (LIQE [108], contrast gain [89]), and a reference-based evaluation with Qwen2.5-VL-72B [10], a vision-language model instructed to judge the percentage of effect removal. We will further discuss these metrics in the supplementary material.

**Baselines.** We compare against both generalist and specialist methods. Generalist baselines include GPT-4o [46], FLUX Kontext [55], Nano Banana [33], and Seedream 4.0 [13]. Specialist baselines cover: 1) *Lens flare*: [24, 100, 115], BracketFlare [25], Difflare [114]; 2) *Dehazing*: DCP [37], AOD-Net [60], GCANett [16], PSD [19], Dehazeformer [77], MSF-Net [117], UCL-Dehazet [89], DiffDehaze [87]; 3) *Shadow removal*: ShadowFormer [34], ShadowRefiner [27], DCShadowNet [49], ShadowDiffusion [35], StableShadowDiff [93]; 4) *Reflection removal*: [109], YTMT [43], DSRNet [44], PromptRR [88],

Table 2. No-reference quantitative comparison on in-the-wild images for four SER tasks. We report results from multiple image quality assessment metrics.

| **Haze** | | | | **Shadow** | | | |
|---|---|---|---|---|---|---|---|
| **Method** | **LIQE↑** | **Contrast↑** | **QwenQA↑** | **Method** | **LIQE↑** | **Contrast↑** | **QwenQA↑** |
| Dehazeformer | 1.9999 | +0.74 | 0.0 | ShadowFormer | 3.3704 | +3.09 | 18.8 |
| DiffDehaze | 1.5624 | +0.03 | 9.1 | ShadowRefiner | 3.5179 | -2.30 | 26.3 |
| Flux Kontext | 2.2584 | +3.85 | 22.7 | Flux Kontext | 3.3184 | +0.73 | 36.3 |
| Nano Banana | 2.6864 | +0.26 | 27.3 | Nano Banana | 3.6399 | -4.93 | 35.0 |
| Seedream 4.0 | 2.1253 | +2.60 | 52.7 | Seedream 4.0 | 2.7640 | -3.58 | 36.3 |
| **Ours** | **2.8225** | **+5.57** | **60.0** | **Ours** | **3.7764** | **+3.61** | **65.0** |
| **Lens Flares** | | | | **Reflections** | | | |
| **Method** | **LIQE↑** | **Contrast↑** | **QwenQA↑** | **Method** | **LIQE↑** | **Contrast↑** | **QwenQA↑** |
| Uformer | 1.3832 | -4.39 | 30.9 | YTMT | 1.1187 | -2.30 | 14.4 |
| BracketFlare | 3.3377 | -10.72 | 13.6 | DSRNet | 1.6975 | -4.17 | 17.8 |
| Flux Kontext | 3.0574 | -0.31 | 62.7 | Flux Kontext | 1.7009 | +0.71 | 8.9 |
| Nano Banana | 3.0358 | -4.05 | 71.8 | Nano Banana | 2.0935 | -1.25 | 56.7 |
| Seedream 4.0 | 2.1643 | -4.41 | 73.6 | Seedream 4.0 | 1.6145 | -0.96 | 53.5 |
| **Ours** | **3.5186** | **+2.33** | **92.7** | **Ours** | **2.2257** | **+1.83** | **75.6** |

Table 3. Quantitative comparison with state-of-the-art methods across four soft effect removal tasks. We report PSNR (↑) and SSIM (↑) on eight benchmarks. Our unified model is compared against specialist methods in each respective category.

| **Lens Flares** | | | **Haze** | | | | |
|---|---|---|---|---|---|---|---|
| **Method** | Flare7k | | **Method** | HSTS | | SOTS | |
| | PSNR | SSIM | | PSNR | SSIM | PSNR | SSIM |
| *Zhang et al.* [100] | 21.02 | 0.784 | DCP | 17.01 | 0.803 | 18.38 | 0.819 |
| *Zhou et al.* [115] | 25.18 | 0.872 | AOD-Net | 19.68 | 0.835 | 20.08 | 0.861 |
| UNet [24] | 26.11 | 0.879 | GCANet | 21.37 | 0.874 | 21.66 | 0.867 |
| Restormer [24] | 26.28 | 0.883 | PSD | 19.37 | 0.824 | 20.49 | 0.844 |
| Uformer [24] | 26.98 | 0.890 | MSFNet | 31.03 | 0.931 | **30.07** | 0.939 |
| Difflare | 26.06 | **0.898** | UCL-Dehaze | 26.87 | 0.933 | 25.21 | 0.927 |
| **Ours** | **27.34** | 0.891 | **Ours** | **32.17** | **0.962** | 29.52 | **0.955** |

| **Shadow** | | | | | | | **Reflections** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | WSRD+ | | ISTD+ | | SRD | | **Method** | SIR2 | | Nature20 | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | | PSNR | SSIM | PSNR | SSIM |
| ShadowFormer | 25.44 | 0.820 | 32.78 | 0.934 | 30.58 | 0.958 | *Zhang et al.* [109] | 22.45 | 0.872 | 20.37 | 0.772 |
| ShadowRefiner | 26.04 | 0.827 | 31.03 | 0.928 | - | - | YTMT | 23.05 | 0.886 | 21.03 | 0.802 |
| DCShadowNet | 21.62 | 0.593 | 25.50 | 0.694 | - | - | DSRNet | 24.97 | 0.907 | 21.70 | 0.820 |
| ShadowDiffusion | - | - | 31.08 | 0.950 | 31.91 | 0.968 | PromptRR | 24.22 | 0.876 | 21.00 | 0.814 |
| StableShadowDiff | 26.26 | 0.827 | 35.19 | **0.970** | 33.63 | 0.968 | L-DiffER | 25.18 | 0.911 | 23.95 | **0.831** |
| **Ours** | **26.91** | **0.829** | **35.59** | 0.964 | **34.16** | **0.971** | **Ours** | **25.98** | **0.911** | **24.17** | 0.812 |

L-DiffER [42].

## 4.2. Comparisons with State-of-The-Art

**Qualitative Comparisons.** Fig. 4 visually compares UniSER with state-of-the-art models on challenging in-the-wild images. Specialist models generalize poorly to out-of-domain data, often resulting in incomplete removal or new artifacts. Meanwhile, powerful generalist models like Nano Banana and FLUX Kontext suffer from instability and fail to preserve scene details, leading to significant content drift (highlighted by red circles). In contrast, UniSER effectively removes a wide range of soft effects while remaining highly faithful to the original image content, producing clean and content-consistent results.

**Quantitative Comparisons.** To assess real-world generalization,we first conduct a comparison on a challenging in-the-wild test set using no-reference metrics, shown in Table 2. In this more difficult setting, UniSER significantly outperforms both specialist and generalist baselines in terms of perceptual quality and removal efficacy, achieving the highest LIQE, Contrast gain, and QwenQA scores across nearly all tasks, which highlights its robust generalization. We then evaluate UniSER against specialists on eight standard benchmarks using full-reference metrics (Table 3). The results show our unified model achieves state-of-the-art performance, consistently outperforming or matching specialist models by obtaining top scores across all four tasks, including the highest PSNR on multiple benchmarks.

## 4.3. Ablations and Applications

**Joint effect removal.** We conduct an ablation study to validate the effectiveness of our joint-task learning strategy.

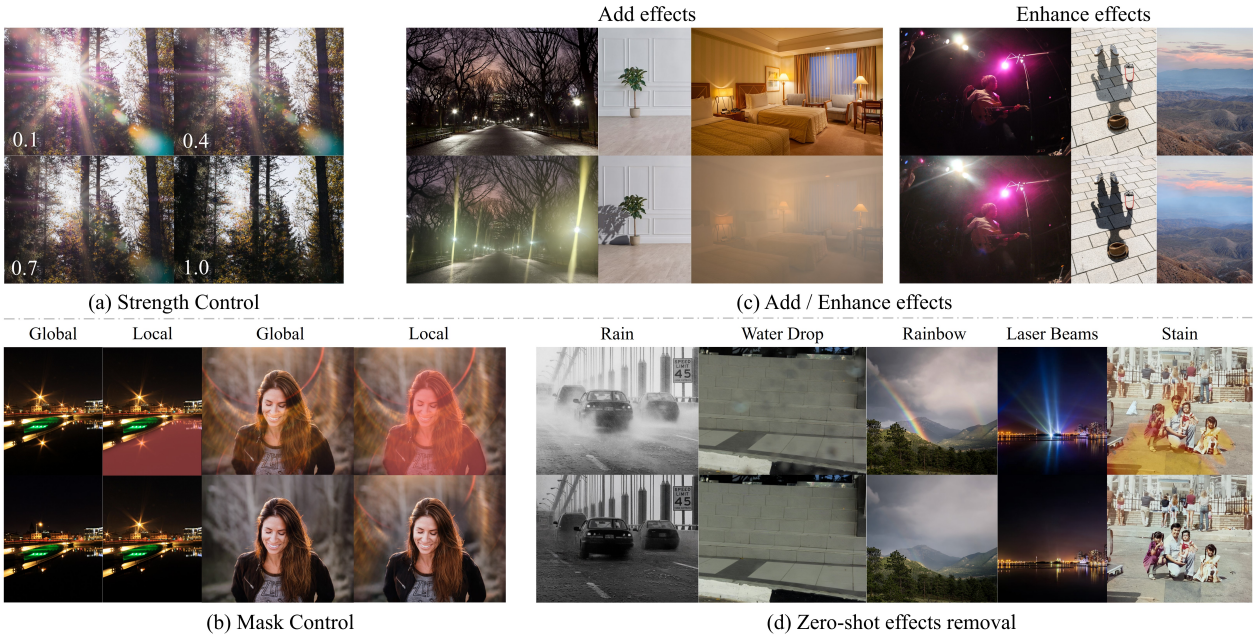



Figure 5. (a) Illustration of Strength Control for effect removal. (b) Illustration of Mask Control for accurate user regional editing. (c) Adding realistic effects to clean image, or enhance current effects for flexible editing purpose. (d) Zero-shot generalization ability on multiple unseen degradations like rain, stain, etc.

As shown in Table 4, we compare our full model, trained with Joint-Task Learning (JTL), against four same models trained independently using Single-Task Learning (STL). The results clearly indicate that the JTL model consistently outperforms the STL models across all four tasks on their respective benchmarks. This superiority suggests that by learning a unified representation from diverse soft effects, UniSER develops a more robust and generalizable feature space that benefits all individual tasks.

**Strength control.** As illustrated in Figure 7(a), UniSER provides fine-grained control over the intensity of the effect removal. Users can specify a continuous strength value, allowing for a smooth transition from partial reduction to complete effect removal. This feature offers greater flexibility for users to achieve their desired level of restoration.

**Mask control.** UniSER supports precise, localized editing through mask-based control, as shown in Figure 7(b). By providing a binary mask, users can designate specific spatial regions for effect removal while leaving the rest of the image untouched. This allows for targeted and accurate edits tailored to user needs.

**Effects addition and enhancement.** Beyond removal, the UniSER framework is also capable of generative tasks. As demonstrated in Figure 7(c), by inverting the process, our model can realistically add new soft effects to clean images or enhance existing ones. This versatility makes it a valuable tool for creative editing and data augmentation.

**Zero-shot removal.** UniSER exhibits strong generalization capabilities to novel degradations not seen during training. As shown in Figure 7(d), the model can perform zero-shot removal of unseen artifacts such as rain and stains. This ability underscores the robustness of the learned features and the model's potential to handle a wider range of image restorations beyond its core training tasks.

Table 4. Ablation study on training strategies. JTL (Joint-Task Learning) represents our full UniSER, while STL (Single-Task Learning) denotes models trained separately for each task.

| Method | Lens Flares | | Haze | | Shadow | | Reflections | |
|---|---|---|---|---|---|---|---|---|
| | Flare7k | | HSTS | | ISTD+ | | SIR2-wild | |
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| STL | 27.18 | 0.890 | 31.91 | **0.963** | 35.43 | 0.963 | 26.40 | 0.876 |
| JTL | **27.34** | **0.891** | **32.17** | 0.962 | **35.59** | **0.964** | **27.44** | **0.918** |

## 5. Conclusion and Limitations

We introduced UniSER, a unified foundation model that validates a data-centric methodology for Soft Effects Removal (SER) task, which effectively handles diverse degradations including lens flare, haze, shadows, and reflections. By curating a large-scale dataset with high-quality pairs and training with dedicated controls, UniSER overcomes the poor generalization of specialist models and the content inconsistency of generalist approaches. Extensive experiments demonstrate that our model achieves state-of-the-art performance on standard benchmarks and superior perceptual quality on in-the-wild images while providing fine-grained user controls, supports creative effect generation, and shows strong zero-shot generalization capabilities. Key limitations include its high computational cost and the extensive resources required for training. Nevertheless,

UniSER represents a significant step towards a universal and controllable solution for high-fidelity image restoration.

## References


[1] Cosmin Ancuti, Codruta O Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: A dehazing benchmark with real hazy and haze-free indoor images. In *International conference on advanced concepts for intelligent vision systems*, pages 620–631. Springer, 2018. 4, 1

[2] Codruta O Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 754–762, 2018. 1

[3] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *2019 IEEE international conference on image processing (ICIP)*, pages 1014–1018. IEEE, 2019. 1

[4] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 444–445, 2020. 1

[5] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2021 nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 627–646, 2021.

[6] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Han Zhou, Wei Dong, Yangyi Liu, Jun Chen, Huan Liu, Liangyan Li, et al. Ntire 2023 hr nonhomogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1808–1825, 2023.

[7] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, Yidi Liu, Xingbo Wang, Yurui Zhu, Gege Shi, Xin Lu, Xueyang Fu, et al. Ntire 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6453–6468, 2024. 4, 1

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 4

[9] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[10] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 4

[11] D Blender Online Community. Blender—a 3d modelling and rendering package. *Blender Foundation*, 2018. 3

[12] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[13] ByteDance. Seedream-4: A large-scale text-to-image generation model, 2024. 6

[14] Yancheng Cai, Fei Yin, Dounia Hammou, and Rafal Mantiuk. Do computer vision foundation models learn the low-level characteristics of the human visual system? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20039–20048, 2025. 4

[15] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. 4

[16] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019. 2, 3, 6

[17] I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17969–17979, 2025. 2, 3, 8, 9

[18] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 4

[19] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7180–7189, 2021. 2, 3, 6

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 4

[21] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 3

[22] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 8

[23] Yuning Cui, Wenqi Ren, and Alois Knoll. Bio-inspired im-

age restoration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3, 8, 9

[24] Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. *Advances in Neural Information Processing Systems*, 35:3926–3937, 2022. 3, 6, 7, 9

[25] Yuekun Dai, Yihang Luo, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Nighttime smartphone reflective flare removal using optical center symmetry prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20783–20791, 2023. 6, 9

[26] Yuekun Dai, Dafeng Zhang, Xiaoming Li, Zongsheng Yue, Chongyi Li, Shangchen Zhou, Ruicheng Feng, et al. Mipi 2024 challenge on nighttime flare removal: Methods and results. *arXiv preprint arXiv:2404.19534*, 2024. 3, 4, 1

[27] Wei Dong, Han Zhou, Yuqiong Tian, Jingke Sun, Xiaohong Liu, Guangtao Zhai, and Jun Chen. Shadowrefiner: Towards mask-free shadow removal via fast fourier transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6208–6217, 2024. 2, 3, 6, 9

[28] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2021. 3

[29] Deniz Engin, Anil Genç, and Hazim Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 825–833, 2018. 2, 3

[30] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017. 3

[31] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2021. 3

[32] Allabakash Ghodesawar, Vinod Patil, Ankit Raichur, Swaroop Adrashyappanamath, Sampada Malagi, Nikhil Akalwadi, Chaitra Desai, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudenagudi. Deflare-net: Flare detection and removal network. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 465–472. Springer, 2023. 2

[33] Google. Introducing gemini 2.5 flash image, our state-of-the-art image model, 2025. 2, 6

[34] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps shadow removal. In *Proceedings of the AAAI conference on artificial intelligence*, pages 710–718, 2023. 6, 9

[35] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14049–14058, 2023. 6

[36] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR 2011*, pages 2033–2040. IEEE, 2011. 3

[37] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 2, 3, 6

[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4

[39] Yuwen He, Wei Wang, Wanyu Wu, and Kui Jiang. Disentangle nighttime lens flares: self-supervised generation-based lens flare removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3464–3472, 2025. 2

[40] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626*, 3, 2022. 3

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 8

[42] Yuchen Hong, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. L-differ: Single image reflection removal with language-based diffusion model. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 7

[43] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. *Advances in Neural Information Processing Systems*, 34:24683–24694, 2021. 6, 9

[44] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13138–13147, 2023. 6, 9

[45] Yue Huang, Zi'ang Li, Tianle Hu, Jie Wen, Guanbin Li, Jinglin Zhang, Guoxu Zhou, and Xiaozhao Fang. Single image reflection removal via inter-layer complementarity. *arXiv preprint arXiv:2505.12641*, 2025. 3

[46] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 6

[47] Md Tanvir Islam, Nasir Rahim, Saeed Anwar, Muhammad Saqib, Sambit Bakshi, and Khan Muhammad. Hazespace2m: A dataset for haze aware single image dehazing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9155–9164, 2024. 3, 4, 1, 8

[48] Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *European Conference on Computer Vision*, pages 340–359. Springer, 2024. 3, 8

[49] Yeying Jin, Aashish Sharma, and Robby T Tan. Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5027–5036, 2021. 6

[50] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023. 3

[51] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025. 4, 2, 8

[52] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5164–5173, 2020. 3

[53] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[54] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4

[55] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 6

[56] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024. 4

[57] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2019. 2, 3

[58] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14811–14820, 2021. 4, 1

[59] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1750–1758, 2020. 4, 1

[60] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017. 2, 3, 6

[61] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1):492–505, 2018. 3, 4, 6, 1, 8

[62] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3565–3574, 2020. 3, 6

[63] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020. 2, 3, 8

[64] Yu-Fan Lin, Chia-Ming Lee, and Chih-Chung Hsu. Densesr: Image shadow removal as dense prediction. *arXiv preprint arXiv:2507.16472*, 2025. 3

[65] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *CVPR*, pages 1871–1880, 2019. 8

[66] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson WH Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4208, 2024. 3, 8

[67] Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4927–4936, 2021. 3

[68] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[69] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 8

[70] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36:71275–71293, 2023. 2, 3, 8

[71] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4067–4075, 2017. 3, 4, 6, 1

[72] Hamed RahmaniKhezri, Suhong Kim, and Mohamed Hefeeda. Unsupervised single-image reflection removal. *IEEE Transactions on Multimedia*, 25:4958–4971, 2022. 3

[73] Sudarshan Rajagopalan and Vishal M Patel. Awracle: All-weather image restoration using visual in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6675–6683, 2025. 3, 8

[74] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 3

[75] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4

[76] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[77] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 2, 3, 6, 9

[78] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 5

[79] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 4

[80] Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28165–28175, 2025. 3, 8

[81] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[82] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, pages 527–543. Springer, 2020. 8

[83] Florin Vasluianu, Zongwei Wu, and Radu Timofte. Sfnet-a spatial-frequency domain neural network for image lens flare removal. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1711–1717. IEEE, 2024. 2

[84] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1826–1835, 2023. 3, 4, 6, 1

[85] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE international conference on computer vision*, pages 3922–3930, 2017. 2, 3, 6

[86] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1788–1797, 2018. 3, 4, 6, 1

[87] Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zhai, and Xiaohong Liu. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23091–23100, 2025. 3, 6, 9

[88] Tao Wang, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Tae-Kyun Kim, Tong Lu, Hongdong Li, and Ming-Hsuan Yang. Promptrr: Diffusion models as prompt generators for single image reflection removal. *arXiv preprint arXiv:2402.02374*, 2024. 3, 6

[89] Yongzhen Wang, Xuefeng Yan, Fu Lee Wang, Haoran Xie, Wenhan Yang, Xiao-Ping Zhang, Jing Qin, and Mingqiang Wei. Ucl-dehaze: Toward real-world image dehazing via unsupervised contrastive learning. *IEEE Transactions on Image Processing*, 33:1361–1374, 2024. 6, 3

[90] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019. 3

[91] Yicheng Wu, Qiurui He, Tianfan Xue, Rahul Garg, Jiawen Chen, Ashok Veeraraghavan, and Jonathan T Barron. How to train neural networks for flare removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2239–2247, 2021. 2, 3

[92] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, pages 675–684, 2018. 8

[93] Jiamin Xu, Yuxin Zheng, Zelong Li, Chi Wang, Renshu Gu, Weiwei Xu, and Gang Xu. Detail-preserving latent diffusion for stable shadow removal. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7592–7602, 2025. 6

[94] Minglong Xue, Aoxiang Ning, Shivakumara Palaiahnakote, and Mingliang Zhou. Dfdnet: Dynamic frequency-guided de-flare network. *arXiv preprint arXiv:2507.17489*, 2025. 2

[95] Dong Yang and Jian Sun. Proximal dehaze-net: A prior learning-based deep network for single image dehazing. In *Proceedings of the european conference on computer vision (ECCV)*, pages 702–717, 2018. 3

[96] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, pages 654–669, 2018. 3, 4, 1

[97] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 4

[98] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 4

[99] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. *ECCV*, 2022. 8

[100] Jing Zhang, Yang Cao, Zheng-Jun Zha, and Dacheng Tao. Nighttime dehazing with a synthetic benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2355–2363, 2020. 6, 7

[101] Jingdong Zhang, Weikai Chen, Yuan Liu, Jionghao Wang, Zhengming Yu, Zhuowen Shen, Bo Yang, Wenping Wang, and Xin Li. Spgen: Spherical projection as consistent and flexible representation for single image 3d shape generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025. 4

[102] Jingdong Zhang, Jiayuan Fan, Peng Ye, Bo Zhang, Hancheng Ye, Baopu Li, Yancheng Cai, and Tao Chen. Bridgenet: Comprehensive and effective feature interactions via bridge feature for multi-task dense predictions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3657–3672, 2025. 8

[103] Jingdong Zhang, Hanrong Ye, Xin Li, Wenping Wang, and Dan Xu. Multi-task label discovery via hierarchical task tokens for partially annotated dense predictions. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 719–728, 2025.

[104] Jingdong Zhang, Xiaohang Zhan, Lingzhi Zhang, Yizhou Wang, Zhengming Yu, Jionghao Wang, Wenping Wang, and Xin Li. Mtpano: Multi-task panoramic scene understanding via label-free integration of dense prediction priors. *arXiv preprint arXiv:2602.05330*, 2026. 4, 8

[105] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 3

[106] Qing Zhang, Yizhong Zhang, Xu Kuang, Yuanbo Zhou, and Tong Tong. Pa-nafnet: An improved nonlinear activation free network with pyramid attention for single image reflection removal. *Digital Signal Processing*, page 105474, 2025. 3

[107] Ruikun Zhang, Hao Yang, Yan Yang, Ying Fu, and Liyuan Pan. Lmhaze: intensity-aware image dehazing with a large-scale multi-intensity real haze dataset. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024. 3, 4, 1

[108] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 6

[109] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 6, 7

[110] Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9239–9248, 2021. 4, 1

[111] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25445–25455, 2024. 3, 8, 9

[112] H Zheng, Z Lin, J Lu, S Cohen, E Shechtman, C Barnes, J Zhang, N Xu, S Amirghodsi, and J Luo. Cm-gan: Image inpainting with cascaded modulation gan and object-aware training. arxiv 2022. *arXiv preprint arXiv:2203.11947*, 2, 2022. 5

[113] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Single image reflection removal with absorption effect. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13395–13404, 2021. 3

[114] Tianwen Zhou, Qihao Duan, and Zitong Yu. Difflare: Removing image lens flare with latent diffusion model. *arXiv preprint arXiv:2407.14746*, 2024. 2, 3, 6

[115] Yuyan Zhou, Dong Liang, Songcan Chen, Sheng-Jun Huang, Shuo Yang, and Chongyi Li. Improving lens flare removal with general-purpose pipeline and multiple light sources recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12969–12979, 2023. 2, 6, 7

[116] Yuyan Zhou, Dong Liang, Songcan Chen, and Sheng-Jun Huang. Image lens flare removal using adversarial curve learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[117] Xinshan Zhu, Shuoshi Li, Yongdong Gan, Yun Zhang, and Biao Sun. Multi-stream fusion network with generalized smooth l 1 loss for single image dehazing. *IEEE Transactions on Image Processing*, 30:7620–7635, 2021. 6

[118] Yurui Zhu, Xueyang Fu, Peng-Tao Jiang, Hao Zhang, Qibin Sun, Jinwei Chen, Zheng-Jun Zha, and Bo Li. Revisiting single image reflection removal in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25468–25478, 2024. 2, 3, 4, 1

# UniSER: A Foundation Model for Unified Soft Effects Removal
Supplementary Material

In this supplementary material, we are going to illustrate i) more details of our data curation details; ii) more details of the data synthesis pipelines; iii) the detailed design of non-reference metrics; iv) more implementation details and v) more experimental results and quality analysis.

## 6. Data Curation on Public Datasets.

Our data collection process aggregates established benchmarks from each domain. For lens flare removal, we incorporate the real-world paired dataset FlareReal600 [26] for nighttime optical artifacts. For shadow removal, our dataset combines several widely-used benchmarks, including SRD [71], ISTD+ [86], and the high-resolution WSRD+ [84], to cover a wide variety of shadow types and complexities. The most extensive category is haze removal, for which we collected a diverse range of datasets. This includes smaller, real-world datasets captured under controlled conditions, we name this set as Haze-R, including: I-HAZE [1], O-HAZE [2], Dense-Haze [3], NH-Haze [4–7], and video dehaze dataset REVIDE [110], multi-level haze dataset LM-Haze [107]. Large-scale synthetic datasets that provide broad coverage of different haze conditions like RESIDE [61] and HAZESPACE2M [47] are also included. Finally, for reflection removal, we integrated datasets that capture various scenarios, such as general real-world reflections RRW [118], polarization-based captures POLAR-RR [59], and flash-induced reflections RFC [58], and synthetic by overlaying dataset BDN [96]. However, these publicly available datasets were originally collected for specific tasks. As a result, their overall distribution is imbalanced, including discrepancies across different tasks, between real and synthetic data, as well as between indoor and outdoor scenes, and day and night conditions.

## 7. Details of the Haze Synthesis Pipeline

A significant portion of our training dataset, particularly for atmospheric effects like haze, fog, and smoke, was generated using a custom synthesis pipeline. This pipeline was designed to overcome the limitations of existing synthetic datasets, which often lack physical realism and diversity. Our methodology is built upon two core components: (1) a physically-motivated atmospheric rendering engine that applies uniform atmospheric effects based on scene geometry, and (2) a procedural texture generator that creates complex, non-homogeneous patterns to simulate phenomena like patchy fog or smoke plumes.

### 7.1. Physically-Motivated Atmospheric Rendering Model

The foundation of our synthesis pipeline is a unified rendering model inspired by the Radiative Transfer Equation (RTE). This model mathematically describes how light interacts with a participating medium (like haze or fog) as it travels from a scene object to the camera. The final color at a pixel $x$, denoted $I_{out,c}(x)$ for a color channel $c$, is a composite of the attenuated scene radiance and the in-scattered light from the atmosphere, known as airlight.

The image formation model is expressed as:

$$I_{out,c}(x) = I_{in,c}(x) \cdot T_c(x) + A_c \cdot (\omega_{0,c} \cdot \kappa) \cdot (1 - T_c(x)^{\eta}) \quad (2)$$

where:

- $I_{in,c}(x)$ is the original, effect-free color of the scene at pixel $x$.
- $T_c(x)$ is the **transmittance**, representing the fraction of light that successfully travels from the object to the camera without being scattered or absorbed.
- $A_c$ is the color of the **airlight**, which is the ambient environmental light scattered towards the camera by the atmospheric particles. This parameter is crucial for defining the hue of the haze (e.g., white for fog, sky-tinted for haze, warm gray for smoke).
- $\omega_{0,c}$ is the **single-scattering albedo**, a value in $[0, 1]$ indicating the proportion of light extinction that is due to scattering versus absorption. For non-absorptive media like fog and haze, $\omega_0 \approx 1.0$. For absorptive media like smoke, $\omega_0 < 1.0$.
- $\kappa$ is an **anisotropy gain factor**, derived from the Henyey-Greenstein phase function. It accounts for directionality of scattering (i.e., whether particles scatter light more strongly forward or backward). For simplicity in our large-scale synthesis, we set $\kappa = 1$, modeling isotropic scattering.
- $\eta$ is a **multiple-scattering boost exponent** $(0 < \eta \leq 1)$. This term provides a compact approximation for the effects of multiple scattering events. A lower value of $\eta$ increases the brightness of the veil, simulating the appearance of denser media where light scatters multiple times before reaching the camera.

#### 7.1.1. Optical Depth and Transmittance

The transmittance $T_c(x)$ is determined by the optical depth $\tau_c(x)$ of the medium along the line of sight, following the Beer-Lambert law:

$$T_c(x) = e^{-\tau_c(x)} \quad (3)$$

The optical depth is the integral of the extinction coefficient $\beta_{t,c}$ over the distance $d(x)$ from the camera to the object

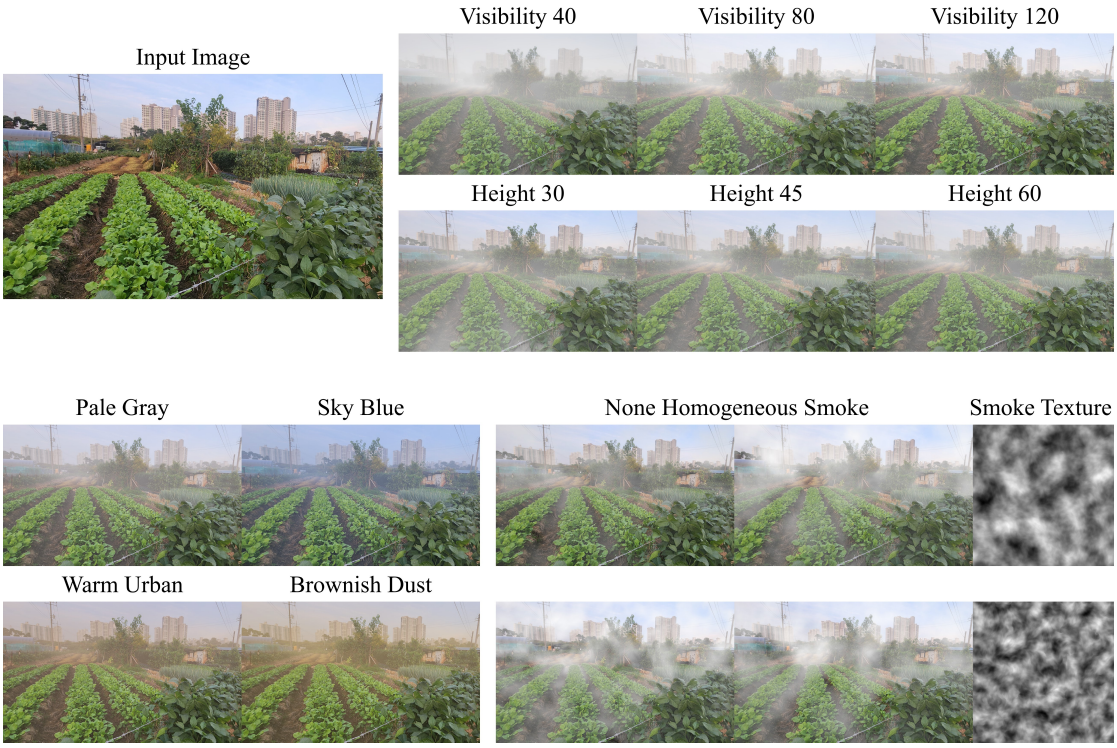



Figure 6. Visualization of our synthetic haze generated by our the proposed pipeline. Our method is capable of synthesizing multiple essences of haze, fog and smoke, within different colors, morphologies and optical properties.

at pixel $x$. To model realistic atmospheres, we assume an exponential decay of particle density with height $h$:

$$\beta_{t,c}(h) = \beta_{t0,c} \cdot e^{-h/H} \tag{4}$$

where $\beta_{t0,c}$ is the base extinction coefficient at a reference height (e.g., sea level), and $H$ is the **scale height**, which defines how rapidly the atmosphere thins out. For a near-horizontal viewing angle, the optical depth can be approximated as:

$$\tau_c(x) \approx \beta_{t0,c} \cdot e^{-h(x)/H} \cdot d(x) \tag{5}$$

The base extinction coefficient $\beta_{t0,c}$ is directly related to the meteorological visibility $V$ by the Koschmieder formula, $\beta_{t0} \approx 3.912/V$.

#### 7.1.2. Geometric Inputs: Depth and Height

Our rendering pipeline requires per-pixel geometric information.

- **Depth:** We use monocular depth maps estimated from the clean input images by Marigold [51]. These normalized depth maps are converted to distance in meters, $d(x)$, using a scene-specific maximum distance $d_{max}$.
- **Height:** When a true height map is unavailable, we utilize a **screen-space height proxy**: $h(x) = h_{max} \cdot (1 - y_{norm})$, where $y_{norm}$ is the normalized vertical coordinate of the pixel (0 at the top, 1 at the bottom). This proxy effectively treats pixels near the horizon as being at a higher altitude, enabling the synthesis of effects like low-lying valley fog that is denser at the bottom of the image.

#### 7.1.3. Color Space and Parameterization

All physical calculations are performed in a linear RGB color space to ensure correctness. Input images, which are typically encoded in sRGB, are first decoded to linear space. After the atmospheric effects are composed, the resulting linear image is encoded back to sRGB. For our large-scale data generation, we programmatically varied all key parameters—including *visibility*, *airlight* color, *eta*, and *H*—across wide, physically plausible ranges to generate a diverse set of training pairs. We also introduced a random baseline value to the optical thickness $\tau$ in each render to add further variety.

### 7.2. Procedural Generation of Non-Homogeneous Media

To simulate complex, turbulent atmospheric effects like patchy fog or smoke, we integrated a procedural texture generator into our pipeline. This process creates realistic, wispy patterns that are used to spatially modulate the density of the rendered haze.

The generation process involves two main steps:

1. **Vector Field Generation:** We first generate a 2D vector field $\vec{V}(\vec{p})$ for each pixel coordinate $\vec{p} = (x, y)$. The components of this field are determined by two independent layers of Perlin noise, $P(\cdot)$, distinguished by unique seeds $(\theta_1, \theta_2)$, which simulates a turbulent flow field. The resulting vectors are normalized to create a unit vector field $\hat{V}(\vec{p})$:

$$\vec{V}(\vec{p}) = \begin{bmatrix} P(\vec{p}; \theta_1) \\ P(\vec{p}; \theta_2) \end{bmatrix}, \quad \hat{V}(\vec{p}) = \frac{\vec{V}(\vec{p})}{\|\vec{V}(\vec{p})\| + \epsilon} \tag{6}$$

   where $\epsilon$ is a small constant to prevent division by zero.
2. **Path Blurring (Advection):** A base noise texture, $M_0(\vec{p})$, is iteratively advected along the vector field $\hat{V}(\vec{p})$ for $N$ steps. In each step $k$, the new texture $M_{k+1}(\vec{p})$ is a blend of the previous texture $M_k(\vec{p})$ and a value sampled from a forward-projected position $\vec{p}'$. This technique smears the initial pattern, creating characteristic streaks. The update rule is:

$$M_{k+1}(\vec{p}) = (1 - \alpha) \cdot M_k(\vec{p}) + \alpha \cdot M_k(\vec{p}') \tag{7}$$

   where $\vec{p}' = \vec{p} + \hat{V}(\vec{p}) \cdot \delta_s$. Here, $\delta_s$ is the step length, $\alpha$ is a blending factor (we use $\alpha = 0.5$), and $M_k(\vec{p}')$ is obtained via bilinear interpolation as $\vec{p}'$ may have non-integer coordinates.

The resulting grayscale texture after $N$ iterations, $M_N(\vec{p})$, is then used as a spatial density modulator, $M(x)$, for the extinction coefficient. The final optical depth calculation is modified to incorporate this texture:

$$\tau_c(x) \approx (\beta_{t0,c} \cdot M(x)) \cdot e^{-h(x)/H} \cdot d(x) \tag{8}$$

This allows us to render haze that is not uniform but varies in density and structure across the image, greatly enhancing the realism and challenge of our synthetic dataset. We also illustrate a sample image synthesized with multiple different types of haze, fog or smoke in Fig. 6.

## 8. Details of Lens Flares & Shadow Synthesis

**HALO Dataset (Lens Flares).** To ensure generalization across complex and diverse real-world lens flares, including extreme cases such as highly blurry artifacts, we synthesize the HALO dataset using the Blender Cycles engine combined with the Flares Wizard add-on. We constructed 375 distinct scenes, comprising 300 distant views (200 outdoor, 100 indoor) and 75 close-up views (50 outdoor, 25 indoor). This yields an overall indoor-to-outdoor ratio of approximately 1:1, with 10% of the scenes utilizing nighttime HDR lighting. Within these scenes, we placed 1,200 unique objects (3 variants per distant scene, 4 per close-up), maintaining a roughly 1:1 ratio between human subjects and general objects. For the flare patterns, we predefined ∼110 diverse types categorized into Streak, Shimmer, Glare, and Reflective effects, with a distribution ratio of 4:2:2:4. During rendering, flare size, intensity, and color are randomly fine-tuned to ensure vast diversity. Notably, the flare generation utilizes physically-inspired 2D heuristics that dynamically respond to the relative 3D spatial positions of the light source and the camera. Compared to static 2D image blending [24], this dynamic simulation provides superior geometric realism and structural diversity.

**LR-SRD Dataset (Shadows).** The shadows in the LR-SRD dataset are entirely natural, derived from real photographs containing genuine objects and their cast shadows. To construct the *shadow-free ground-truth*, we capture a clean background separately and stitch it directly into the **masked** shadow region of the original source image. This composition strategy yields highly realistic paired data without synthetic artifacts. Furthermore, we extract and apply instance-level shadow masks during the training phase to provide precise spatial guidance for the network.

## 9. Non-Reference Evaluation Metrics

To rigorously assess the performance of our model on in-the-wild images where a ground-truth reference is unavailable, we employed specialized non-reference evaluation paradigms. These metrics are designed to provide both a quantitative measure of detail recovery and a qualitative score that emulates human perceptual judgment.

### 9.1. Residual Contrast Gain

While local contrast is a well-established indicator of image sharpness and detail, commonly iused in non-reference dehazing or similar tasks [89]. However since the measurements are averaged over the entire image, for localized effects like some types of lens flares or local shadows, the global evaluation is not significant. To overcome this limitation, we measure the **Residual Contrast Gain**, which quantifies the change in local contrast exclusively within the image regions modified by our model. This approach ensures that the evaluation focuses directly on the model's restoration efficacy. The computation is performed via the following steps:

1. **Identification of Edited Regions.** Given a grayscale input image $I_{in}$ and the model's grayscale output $I_{out}$, we first identify the edited regions by computing a pixel-wise absolute difference map, $D(\vec{p}) = |I_{in}(\vec{p}) - I_{out}(\vec{p})|$, for all pixel coordinates $\vec{p}$. A binary edit mask, $M_{edit}$, is then generated by applying a threshold to this difference map, isolating the set of modified pixels over which the analysis is performed.
2. **Local Contrast Calculation.** We define the local contrast at a pixel $\vec{p}$, denoted $C(\vec{p})$, as the standard deviation of pixel intensities within a $k \times k$ window centered at $\vec{p}$. This operation is performed for both the input and output images, yielding local contrast maps $C_{in}$ and $C_{out}$.

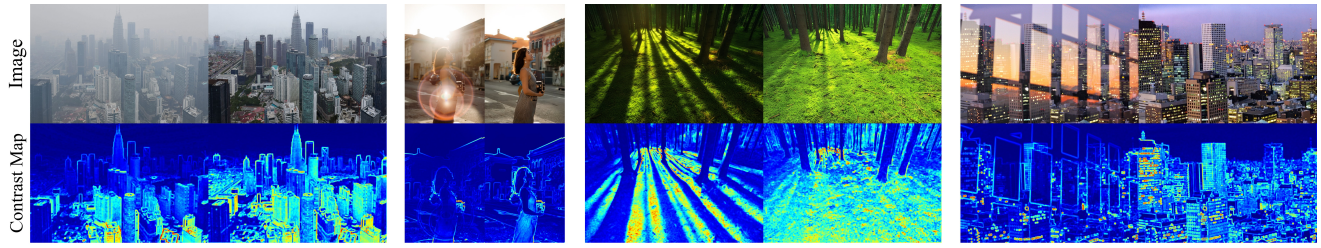



Figure 7. Contrast maps of image before and after edit by UniSER. Significant enhancements of contrast inside effect regions are observed, indicating our method successfully enhances the degraded image details.

3. **Gain Computation.** The final Residual Contrast Gain, $\Delta C_{res}$, is the difference between the average local contrast of the output and input images, computed exclusively over the set of edited pixels (where $M_{edit} = 1$). This is formulated as:

$$\Delta C_{res} = \text{mean}_{\vec{p}|M_{edit}(\vec{p})=1} \left( C_{out}(\vec{p}) - C_{in}(\vec{p}) \right) \quad (9)$$

A positive $\Delta C_{res}$ value indicates a net increase in detail and texture within the restored regions.

### 9.2. QwenQA: VLM-based Assessment

With the rapid development of foundation models [14, 38, 54, 68, 75, 97, 98, 101, 104] and VLMs [8–10, 15, 20, 56, 79, 81], more works introduce multi-modal large language models to assest the quality. We also developed the **QwenQA** evaluation metric, to leverage the powerful Vision-Language Model (VLM) for more human-like visual assessments. Our framework is built upon the **Qwen2.5-VL-72B-Instruct** model [10]. The evaluation protocol is designed for consistency and automated parsing, involving three key stages:

1. **Input Standardization.** To eliminate resolution as a confounding variable, the model's prediction image is first resampled to match the exact dimensions of the original input image, ensuring a fair comparison context for the VLM.
2. **Constrained Prompt Engineering.** The core of QwenQA lies in a meticulously engineered prompt designed to elicit a precise and quantitative response. The prompt structure includes:
   - *Role Assignment:* The VLM is instructed to act as a "top-tier image quality assessment expert," priming it to leverage its most relevant internal knowledge.
   - *Task Definition:* The prompt provides clear context, defining "Image A" as the original with a specific artifact (e.g., 'haze', 'shadow') and "Image B" as the processed result.
   - *Objective Quantization:* The VLM's objective is narrowly focused on a single quantitative task: "evaluate the percentage by which the '[artifact name]' is reduced in Image B compared to Image A". This transforms a descriptive task into a quantitative one.
   - *Strict Output Formatting:* The prompt strictly constrains the VLM's output to a specific format: "Score: [number]%". This instruction explicitly forbids any additional descriptive text, explanations, or conversational filler, which is critical for reliable automated parsing.
3. **Automated Score Parsing.** The final step is to parse the VLM's structured textual output. A regular expression is used to robustly extract the numerical percentage score from the response, yielding the final QwenQA score.

## 10. Implementation Details

### 10.1. Haze Synthesis Details

Our primary objective in data expansion was to generate a challenging and realistic training set that surpasses the limitations of existing synthetic datasets. To achieve this, we developed a high-throughput synthesis pipeline to apply our physically-motivated atmospheric rendering model on a large scale. This section details the parameterization for various haze types, the batch processing architecture, and the datasets involved.

**Parameterization for Diverse Atmospheric Effects.** The versatility of our rendering model allows us to simulate a wide range of atmospheric conditions by adjusting a few key physical parameters. We defined distinct configurations for haze, fog, and smoke, which were systematically varied to ensure a broad data distribution.

- **Haze:** To simulate different environmental conditions, we primarily varied the *airlight* color and *visibility*. For instance, we used sky-tinted colors like (153, 174, 215) for typical haze, warmer tones such as (200, 180, 140) for urban pollution, and grayish colors like (210, 210, 220) for high-altitude conditions. Visibility was typically set in the range of 100m to 1000m to produce varying levels of haze density.
- **Fog:** Fog is characterized by its dense, non-absorptive particles. We simulated this by setting the single-scattering albedo $\omega_0$ to (1.0, 1.0, 1.0) and using a neutral white airlight. Fog density was controlled by varying *visibility* (from 30m to 1000m) and the multiple-scattering boost exponent $\eta$ (typically between 0.5 and 1.0). To sim-

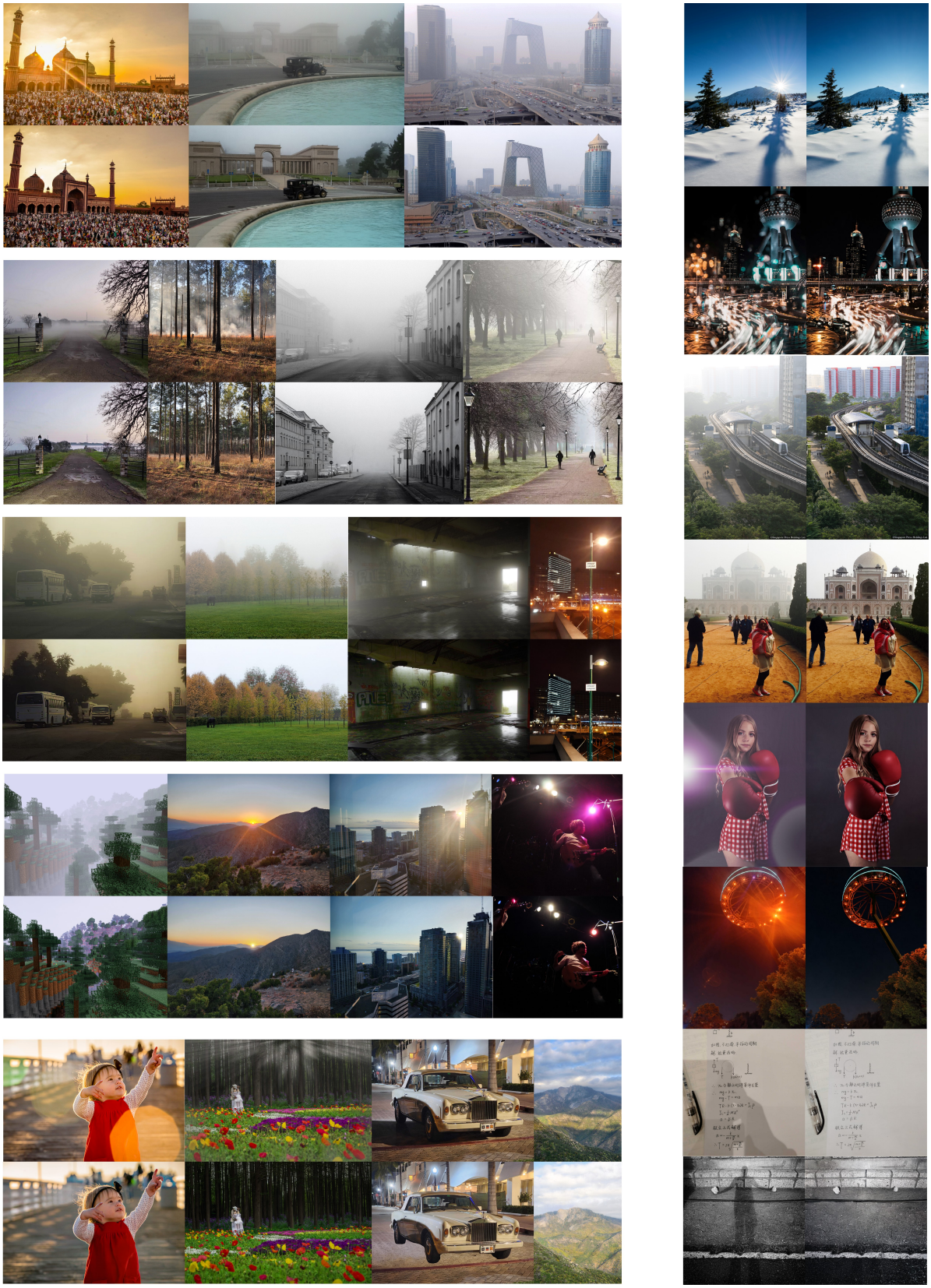


Figure 8. Gallery: Removing effects with UniSER.

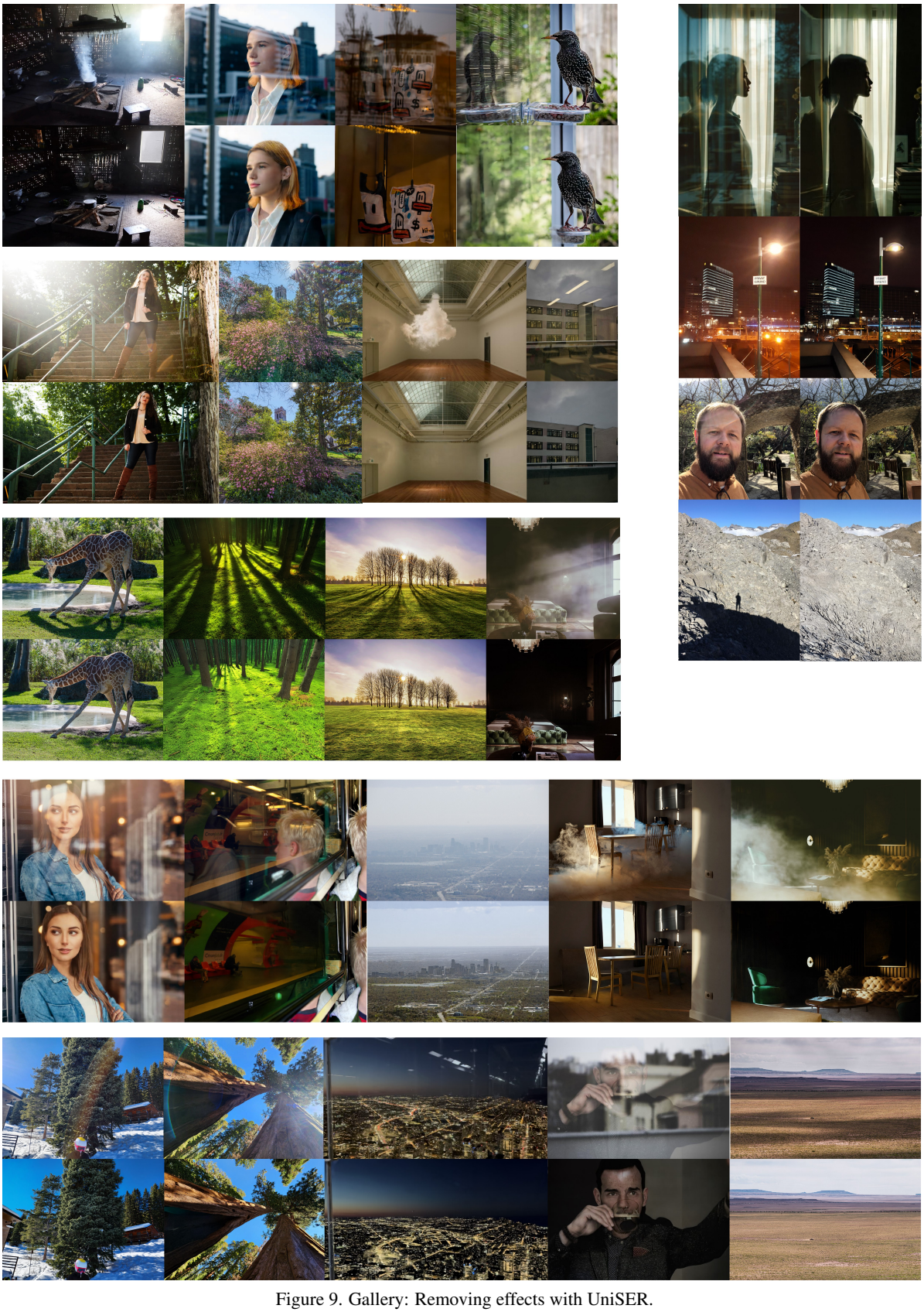


Figure 9. Gallery: Removing effects with UniSER.

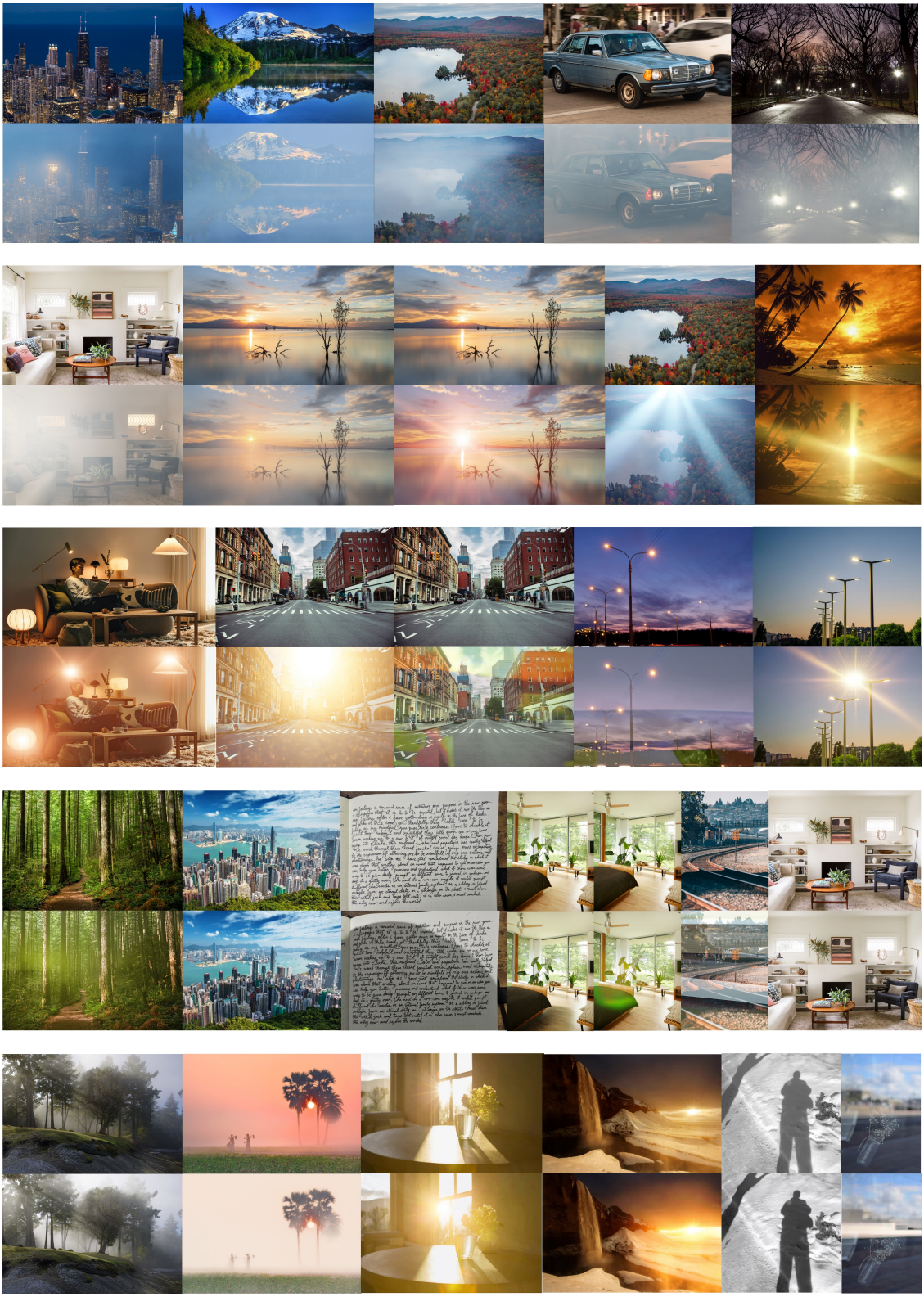


Figure 10. Gallery: Adding or enhancing effects with UniSER.

ulate low-lying or valley fog, we significantly reduced the scale height $H$ (e.g., to 30-60m) to confine the effect to the lower parts of the scene.

- **Smoke:** Unlike haze and fog, smoke is an absorptive medium. This was modeled by setting $\omega_0$ to values less than 1.0 (e.g., 0.75 to 0.85). The *airlight* was configured with warm, darker colors like (180, 150, 120) or (160, 120, 90) to represent the tint of the smoke particles. The scale height $H$ was generally kept low (e.g., 40-50m) to simulate ground-level smoke plumes.

**Large-Scale Batch Synthesis Architecture.** To apply these configurations across a massive number of images, we implemented an efficient, parallelized processing pipeline. The core rendering engine was ported to PyTorch [69] to leverage GPU acceleration. We utilized multiprocessing to create a pool of worker processes. In a multi-GPU environment, these workers were assigned to available GPUs in a round-robin fashion, enabling concurrent rendering of multiple image-configuration pairs. Each worker independently handled the data I/O, pre-processing (color space conversion, data normalization), GPU-based rendering, and post-processing of the synthesized hazy image. This architecture allowed us to generate our extensive dataset in a time-efficient manner.

**Datasets for Synthesis.** As stated in our methodology, our goal was to enhance existing large-scale datasets by generating more challenging and realistic haze effects. We leveraged the high-quality, clean ground truth images from public benchmarks, primarily RESIDE [61] and HAZESPACE [47]. For each clean image in these datasets, we first estimated a monocular depth map [51] and then applied our full suite of atmospheric rendering configurations, resulting in a significant expansion of the training data with diverse and physically plausible haze, fog, and smoke effects.

### 10.2. Training Details

Our work builds upon a pretrained DiT-based image editing model that has demonstrated strong capabilities in general inpainting tasks, such as object addition, removal, and modification. This provides a robust starting point for fine-tuning on our specialized soft-effects dataset. A key aspect of our training methodology is a hierarchical data sampling strategy designed to balance contributions from numerous datasets across multiple tasks. Our data pipeline first groups datasets by their primary task (e.g., shadow removal, dehazing, reflection removal, etc.). During each training step, a task is uniformly sampled, and then a specific dataset within that task group is selected based on a predefined sampling weight. This weighting ratio is configured for each dataset, allowing us to strategically oversample smaller, high-quality real-world datasets to learn the knowledge without domain gaps, while still benefiting from the diversity of larger-scale synthetic data sources to prevent overfitting and enhance the generalization ability. This ensures the model receives a balanced and comprehensive exposure to all types of soft effects.

For the fine-tuning process, our model operates within the DDPM [41] framework, which is adapted to use continuous timesteps for increased flexibility. Notably, we employ $v$-parameterization instead of the standard $\epsilon$-parameterization to improve training stability and sample quality. Our training objective is to predict the noise added to the clean image's latent representation at a given timestep. The loss function is the mean squared error (MSE) between the predicted noise and the ground truth noise, with a timestep-dependent weighting scheme applied to balance the contribution of different noise levels throughout the training. We train the model for 10k steps at a resolution of 1024x1024. We employ the AdamW optimizer with a learning rate of $1.2 \times 10^{-5}$, governed by a linear warmup of 2000 steps followed by a cosine decay schedule. Our UniSER is trained on all of the data mentioned above with 8 NVIDIA A100 80G for 10k iterations.

### 10.3. Evaluation Details for Baselines

When evaluating the generalist baselines, we provided detailed and specific text prompts to ensure they could achieve their optimal performance. These prompts explicitly described the effect to be removed and the relevant scene context, for instance: *"remove the atmosphere haze completely in this image"* or *"remove the shadow casted by the giraffe on the grass"*. Furthermore, to account for the stochastic nature of generative models, if a model performed poorly or failed to remove the effect on a particular sample, we conducted multiple attempts to ensures we are not using ambiguous or vague text prompts. This is a fair evaluation and mitigates biases arising from individual random outcomes. In contrast, our UniSER has minimal dependency on text prompts. In our framework, the text serves merely as a high-level task indicator (e.g., *"remove haze"*) without requiring a detailed description of the scene's content. Consequently, our approach achieves stable and robust results without the need for iterative prompt engineering.

## 11. More Results and Ablations

### 11.1. More Quantitative Results

**Comparison with All-In-One Models.** With the rapid evolution in multi-task learning [22, 65, 82, 92, 99, 102–104], aiming at jointly learning multiple tasks within one model, All-In-One (AIO) models [17, 23, 48, 63, 66, 70, 73, 80, 111] are explored for image restorations. We provide additional quantitative comparisons with the most recent All-In-One (AIO) image restoration methods, including DiffUIR [111], Unirestore [17], and BioIR [23]. As shown in Tab. 5, existing AIO models struggle significantly

Table 5. Quantitative comparison on in-the-wild images and standard task benchmarks with AIO models.

| Method | In-the-Wild | | | | | | | | SOTS | | Flare7k | | WSRD+ | | Nature20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Haze | | Lens Flares | | Shadow | | Reflections | | Haze | | Lens Flares | | Shadow | | Reflections | |
| | LIQE | QwenQA | LIQE | QwenQA | LIQE | QwenQA | LIQE | QwenQA | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| DiffUIR [111] | 2.0493 | 0.0 | 2.9079 | 0.0 | 3.3070 | 10.0 | 1.6983 | 0.0 | 26.38 | 0.922 | 23.14 | 0.853 | 18.43 | 0.782 | 21.08 | 0.768 |
| Unirestore [17] | 1.7477 | 0.9 | 2.4617 | 1.8 | 3.0792 | 17.5 | 1.7068 | 0.0 | 21.96 | 0.863 | 22.90 | 0.831 | 18.11 | 0.755 | 20.78 | 0.753 |
| BioIR [23] | 2.0810 | 0.0 | 2.8830 | 0.0 | 3.4091 | 12.5 | 1.6518 | 0.0 | 24.74 | 0.918 | 19.60 | 0.771 | 17.05 | 0.765 | 20.41 | 0.752 |
| **Ours (zero-shot)** | **2.8225** | **60.0** | **3.5186** | **92.7** | **3.7764** | **65.0** | **2.2257** | **75.6** | 27.43 | 0.928 | 26.98 | 0.890 | 25.13 | 0.820 | 23.99 | 0.808 |
| **Ours (adapt)** | - | - | - | - | - | - | - | - | **29.52** | **0.955** | **27.34** | **0.891** | **26.91** | **0.829** | **24.17** | **0.812** |


Figure 11. Qualitative comparisons with AIO models on restoring degradations like rain, haze, raindrops, and stains.

Table 6. We conduct a user study on in-the-wild test cases to validate our evaluation.

| **Method** | Lens Flare | Haze | Shadow | Reflections |
|---|---|---|---|---|
| Specialist A | 2.9% [25] | 0.0% [77] | 14.6% [27] | 15.9% [43] |
| Specialist B | 11.8% [24] | 25.0% [87] | 20.5% [34] | 18.9% [44] |
| Nano Banana | 38.3% | 34.1% | 3.0% | 60.0% |
| Flux Kontext | 45.2% | 52.5% | 21.6% | 14.0% |
| Seedream4.0 | 64.1% | 70.3% | 25.6% | 55.2% |
| UniSER (Ours) | **97.5%** | **97.1%** | **94.4%** | **89.2%** |

on challenging in-the-wild test cases, often failing to remove the degradations effectively (as reflected by near-zero QwenQA scores). In contrast, our UniSER demonstrates exceptional zero-shot generalization, achieving the highest LIQE and QwenQA scores across all four effect categories. Furthermore, when adapted to specific benchmark domains, our model consistently establishes state-of-the-art performance in full-reference metrics (PSNR and SSIM) across the SOTS, Flare7k, WSRD+, and Nature20 datasets. We additionally provide visual comparisons with AIO methods in Fig. 11 on unseen types of soft effects like rain, haze, raindrops, and stains.

**User Study.** To further validate our evaluation on real-world images without ground-truth labels, we conducted a comprehensive user study, summarized in Tab. 6. We compared UniSER against both task-specific specialist models and state-of-the-art generative foundation models. Participants were asked to evaluate the effect of the removal completeness and identity preservation of each model individually. UniSER overwhelmingly dominated user preference, securing 89.2% to 97.5% of the votes across all four tasks, further demonstrating its robustness in handling complex, in-the-wild soft effects.


Figure 12. Ablations on mask control and soft mask effects. UniSER precisely targets the user-specified red regions while leaving the unmasked background untouched. With the mask control incorporated in the training process, localized dense effects are easily and thoroughly removed. The blurred mask effectively eliminates hard boundary artifacts, ensuring a seamless and natural transition.

## 11.2. More Qualitative Results

We provide more visual results in Fig. 8, Fig. 9 and Fig. 10, by randomly pick in-the-wild photos degraded by soft effects, our UniSER shows perfect robustness on thoroughly removing the. Besides, UniSER is also capable of generating or enhancing multiple effects aesthetically.

## 11.3. More Ablations

**Contrast Analysis** To further investigate how UniSER improves image quality, we visualize the local contrast maps of images before and after editing, as shown in Figure 7. A significant enhancement in contrast is observed within the regions originally degraded by soft effects. This indicates that our method not only removes the obstructive artifacts but also successfully restores and enhances the underlying image details and textures that were suppressed by the effects, leading to a clearer and more vivid output.

**Mask Control and Boundary Smoothness.** As illustrated in Figure 12, executing a global removal without spatial guidance (w.o. Mask) alters the soft effects across the entire image. By introducing specific spatial masks (w. Mask), UniSER restricts the restoration strictly to the user-defined regions (e.g., targeted steam, localized shadows, or specific lens glares), accurately preserving the original lighting and atmospheric conditions in the unmasked areas, while applying mask control during the training process also helps the model tackle more challenging scenes with dense localized effects like smoke and shadow. Furthermore, the bottom-right panel demonstrates the necessity of our mask blurring strategy during training and inference. Applying a hard binary mask (w.o. Blur) forces an abrupt restoration, resulting in obvious structural inconsistencies and sharp boundary artifacts. In contrast, softening the mask via Gaussian blur (w. Blur) creates a smooth removal gradient, enabling the seamless integration of the restored area with the untouched background.