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.

1. 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 [19] for nighttime optical artifacts. For shadow removal, our dataset combines several widely-used benchmarks, including SRD [40], ISTD+ [49], and the high-resolution WSRD+ [48], 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 [62], multi-level haze dataset LM-Haze [61]. Large-scale synthetic datasets that provide broad coverage of different haze conditions like RESIDE [33] and HAZESPACE2M [26] are also included. Finally, for reflection removal, we integrated datasets that capture various scenarios, such as general real-world reflections RRW [64], polarization-based captures POLAR-RR [32], and flash-induced reflections RFC [31], and synthetic by overlaying dataset BDN [53]. 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.

2. 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.

2.1. 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:

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

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, $\omega_0 < 1.0$.
- κ 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.
- η 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 η increases the brightness of the veil, simulating the appearance of denser media where light scatters multiple times before reaching the camera.

2.1.1. 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:

$$T_c(x) = e^{-\tau_c(x)} \quad (2)$$

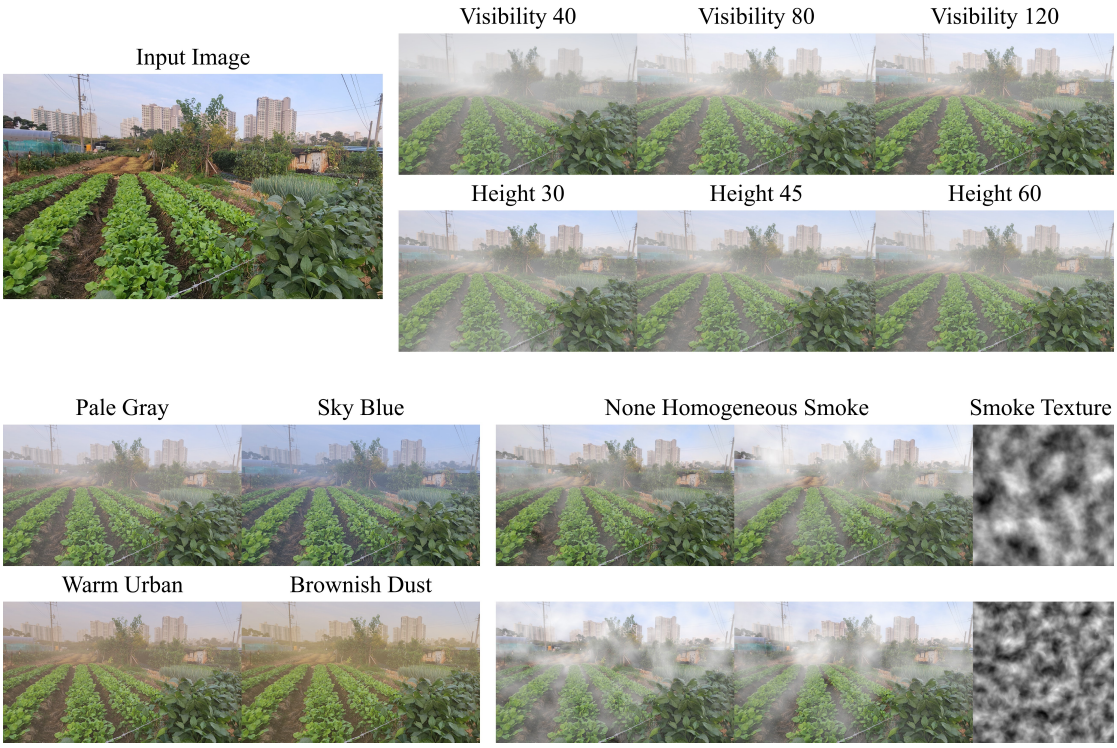

Figure 1. Visualization of our synthetic haze generated by our the proposed pipeline. Our method is capable of synthesizing multiple essences of haze, fog and smoke, within different colors, morphologies and optical properties.

The optical depth is the integral of the extinction coefficient $\beta_{t,c}$ over the distance $d(x)$ from the camera to the object at pixel x. To model realistic atmospheres, we assume an exponential decay of particle density with height h:

$$\beta_{t,c}(h) = \beta_{t0,c} \cdot e^{-h/H} \tag{3}$$

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:

$$\tau_c(x) \approx \beta_{t0,c} \cdot e^{-h(x)/H} \cdot d(x) \tag{4}$$

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

2.1.2. 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 [28]. 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.

2.1.3. 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 τ in each render to add further variety.

2.2. 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 (θ_1, θ_2), which simulates a turbulent flow field. The resulting vectors are normalized to create a unit vector field $\hat{V}(\vec{p})$:

$$\vec{V}(\vec{p}) = \begin{bmatrix} P(\vec{p}; \theta_1) \\ P(\vec{p}; \theta_2) \end{bmatrix}, \quad \hat{V}(\vec{p}) = \frac{\vec{V}(\vec{p})}{\|\vec{V}(\vec{p})\| + \epsilon} \tag{5}$$

   where ϵ 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:

$$M_{k+1}(\vec{p}) = (1 - \alpha) \cdot M_k(\vec{p}) + \alpha \cdot M_k(\vec{p}') \tag{6}$$

   where $\vec{p}' = \vec{p} + \hat{V}(\vec{p}) \cdot \delta_s$. Here, δ_s is the step length, α 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:

$$\tau_c(x) \approx (\beta_{t0,c} \cdot M(x)) \cdot e^{-h(x)/H} \cdot d(x) \tag{7}$$

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. 1.

3. 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 [17], 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.

4. 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.

4.1. 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 [51]. 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 $\vec{p}$.

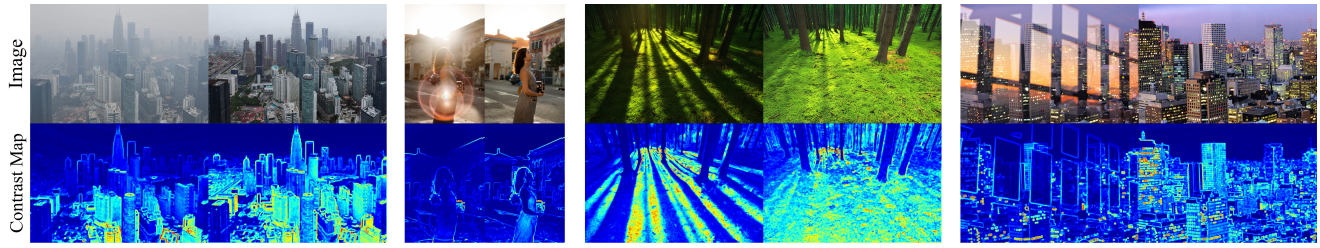

Figure 2. Contrast maps of image before and after edit by UniSER. Significant enhancements of contrast inside effect regions are observed, indicating our method successfully enhances the degraded image details.

This operation is performed for both the input and output images, yielding local contrast maps C_{in} and C_{out}.

3. **Gain Computation.** The final Residual Contrast Gain, ΔC_{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:

$$\Delta C_{res} = \text{mean}_{\vec{p}|M_{edit}(\vec{p})=1} \left(C_{out}(\vec{p}) - C_{in}(\vec{p}) \right) \quad (8)$$

A positive ΔC_{res} value indicates a net increase in detail and texture within the restored regions.

4.2. QwenQA: VLM-based Assessment

With the rapid development of foundation models [11, 22, 29, 37, 42, 54, 55, 57, 60] and VLMs [8–10, 12, 14, 30, 44, 46], 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.

5. Implementation Details

5.1. 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 ω_0 to (1.0, 1.0, 1.0) and using a neutral

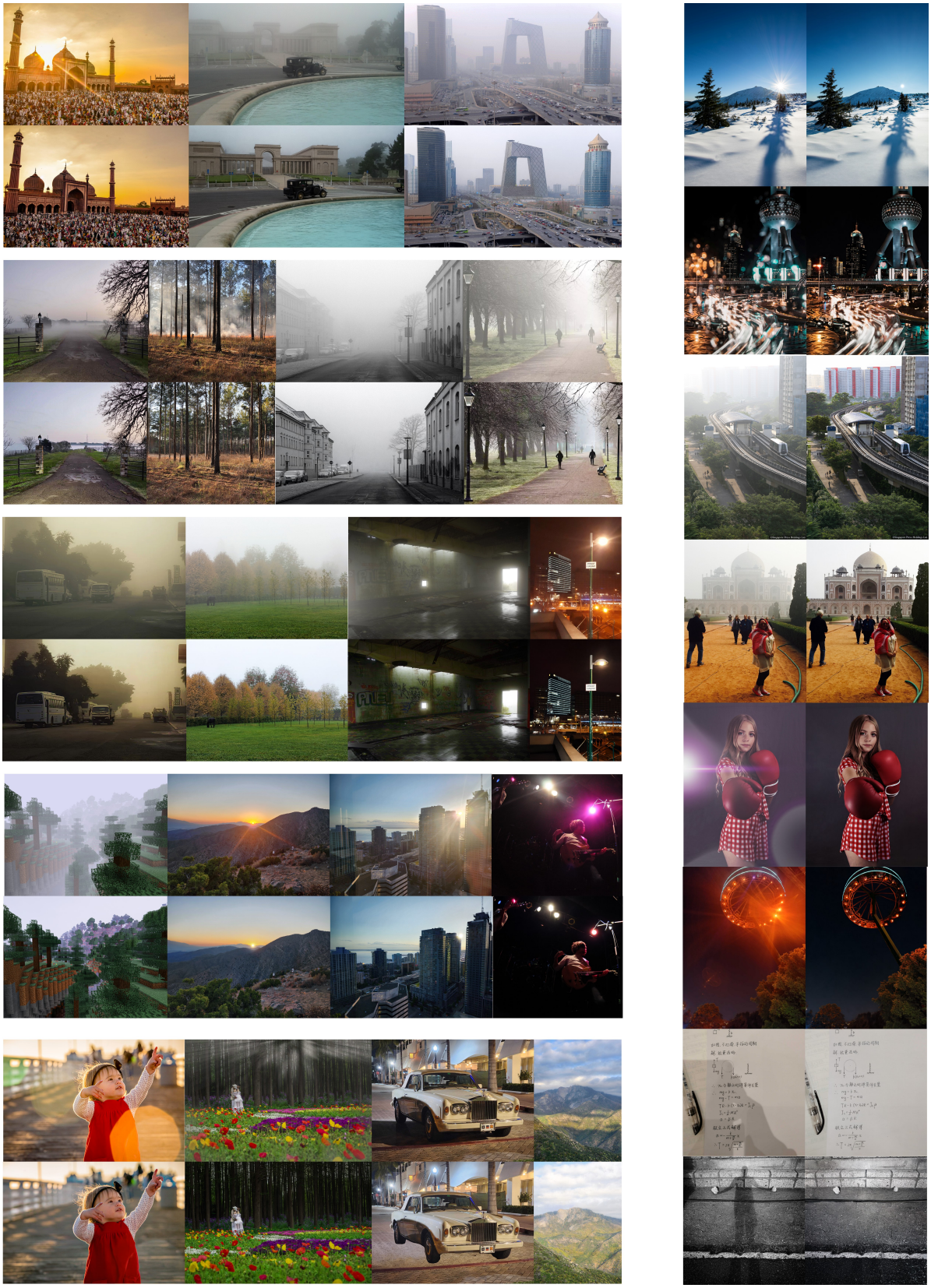

Figure 3. Gallery: Removing effects with UniSER.

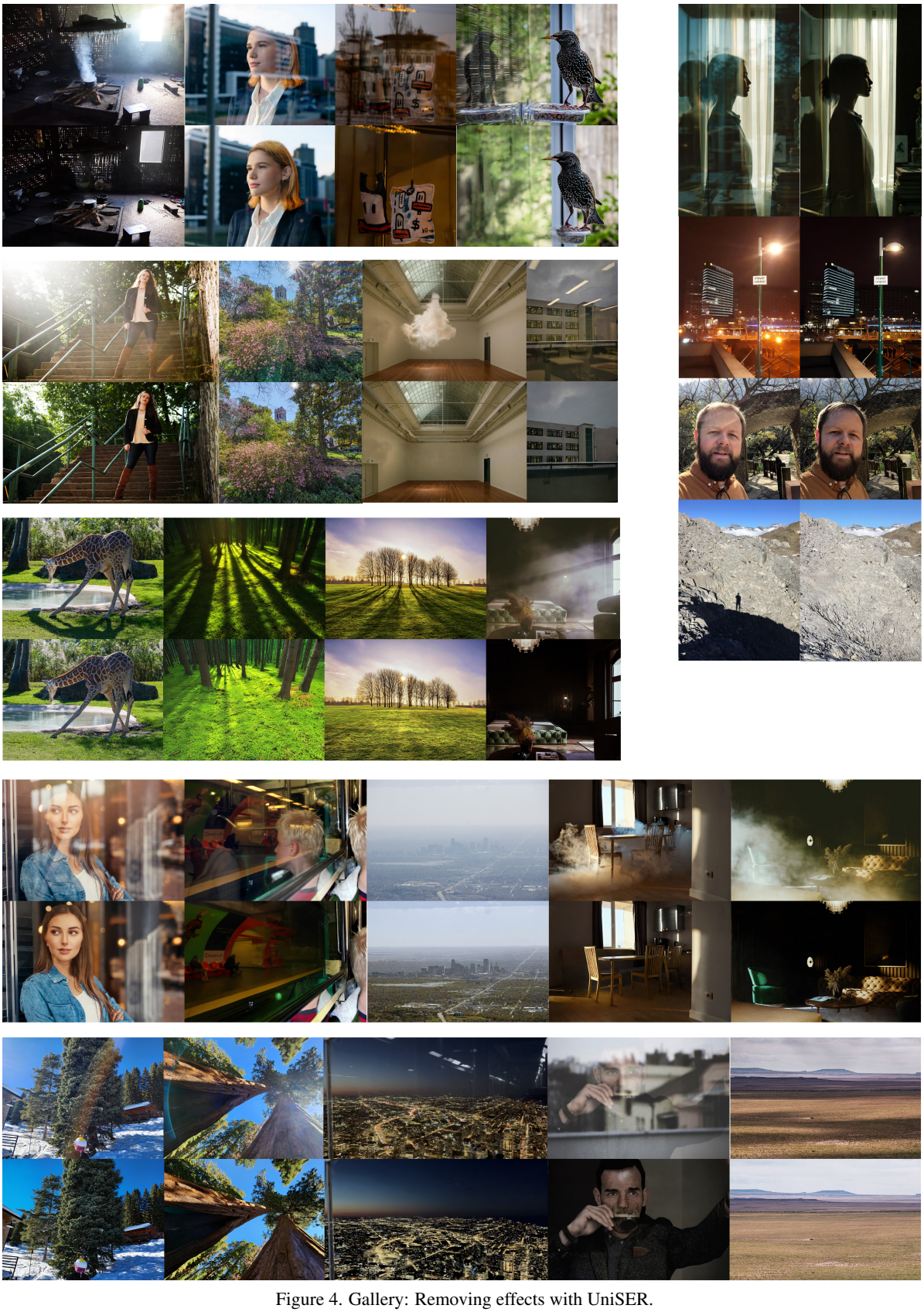

Figure 4. Gallery: Removing effects with UniSER.

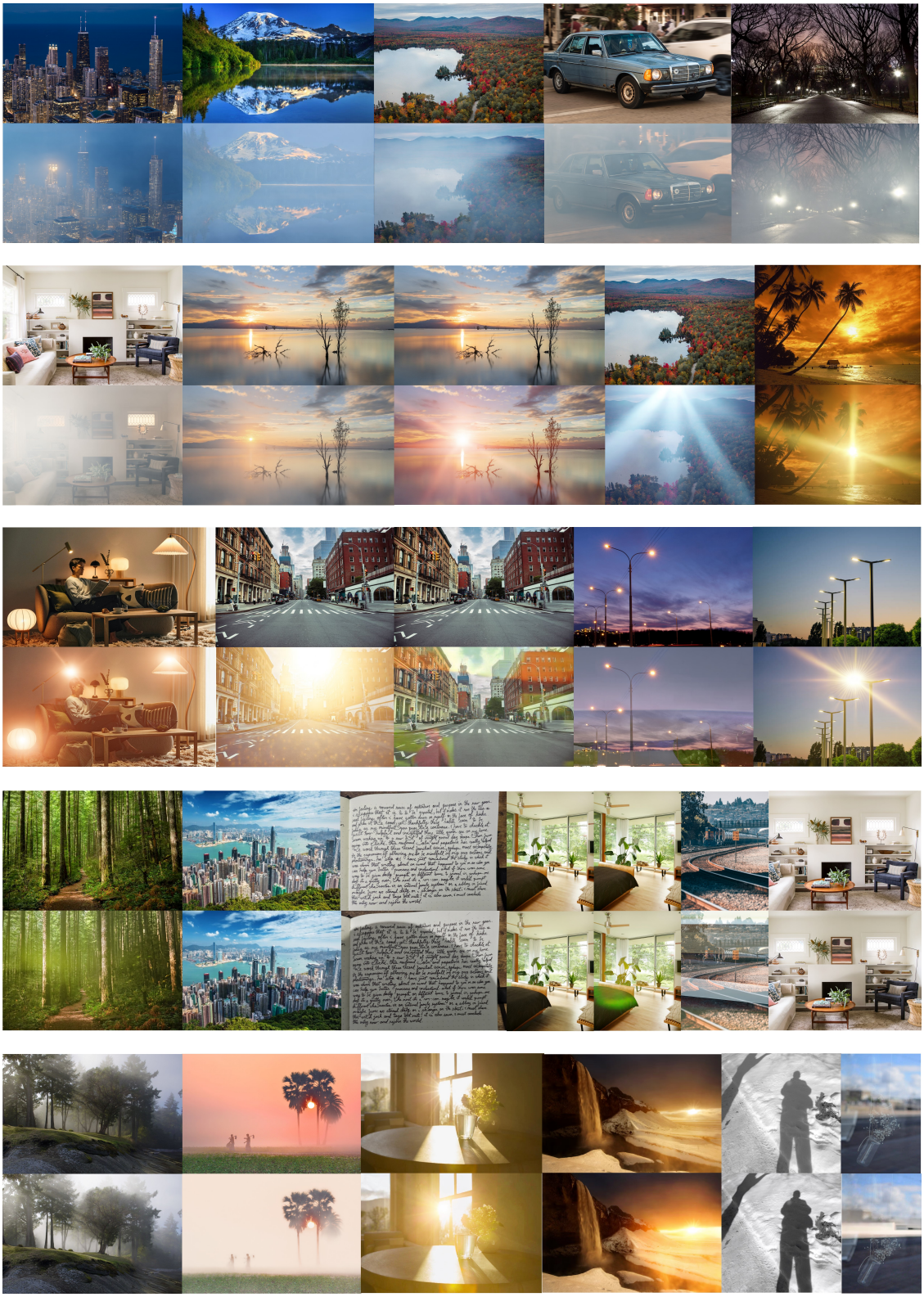

Figure 5. Gallery: Adding or enhancing effects with UniSER.

white airlight. Fog density was controlled by varying *visibility* (from 30m to 1000m) and the multiple-scattering boost exponent η (typically between 0.5 and 1.0). To simulate 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 ω_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 [38] 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 [33] and HAZESPACE [26]. For each clean image in these datasets, we first estimated a monocular depth map [28] 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.

5.2. 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 [23] framework, which is adapted to use continuous timesteps for increased flexibility. Notably, we employ v-parameterization instead of the standard ϵ-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 1.2×10^{-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.

5.3. 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.

6. More Results and Ablations

6.1. More Quantitative Results

Comparison with All-In-One Models. With the rapid evolution in multi-task learning [15, 35, 47, 52, 56, 58–60], aiming at jointly learning multiple tasks within one model, All-In-One (AIO) models [13, 16, 27, 34, 36, 39, 41, 45, 63] are explored for image restorations. We provide additional quantitative comparisons with the most recent All-

Table 1. 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 [63]	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 [13]	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 [16]	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 6. Qualitative comparisons with AIO models on restoring degradations like rain, haze, raindrops, and stains.

Table 2. 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% [18]	0.0% [43]	14.6% [20]	15.9% [24]
Specialist B	11.8% [17]	25.0% [50]	20.5% [21]	18.9% [25]
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%**

In-One (AIO) image restoration methods, including DiffUIR [63], Unirestore [13], and BioIR [16]. As shown in Tab. 1, existing AIO models struggle significantly 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. 6 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

Figure 7. 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.

comprehensive user study, summarized in Tab. 2. 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.

6.2. More Qualitative Results

We provide more visual results in Fig. 3, Fig. 4 and Fig. 5, 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.

6.3. 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 2. 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 7, 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.

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. 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 non-homogeneous 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. 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. 4

[11] 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

[12] 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

[13] 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. 8, 9

[14] 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

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

[16] Yuning Cui, Wenqi Ren, and Alois Knoll. Bio-inspired image restoration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 8, 9

[17] 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, 9

[18] 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. 9

[19] 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. 1

[20] 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. 9

[21] 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. 9

[22] 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

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

[24] 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. 9

[25] 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. 9

[26] 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. 1, 8

[27] 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. 8

[28] 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. 2, 8

[29] 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

[30] 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

[31] 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. 1

[32] 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. 1

[33] 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. 1, 8

[34] 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. 8

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

[36] 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. 8

[37] 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

[38] 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

[39] 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. 8

[40] 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. 1

[41] 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. 8

[42] 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

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

[44] 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

[45] 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. 8

[46] 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

[47] 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

[48] 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. 1

[49] 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. 1

[50] 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. 9

[51] 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. 3

[52] 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

[53] 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. 1

[54] 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

[55] 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

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

[57] 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

[58] 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

[59] 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.

[60] 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

[61] 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. 1

[62] 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. 1

[63] 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. 8, 9

[64] 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. 1