This supplementary is organized as follows:

- In Sec. A, we first introduce our **open-vocabulary** user interaction application, highlighting its significance in enhancing user engagement.
- Next, in Sec. B, we discuss the trade-off of selecting adjustable coefficient w introduced in Sec. 3 of the main body.
- To address the restoration of real-world images affected by multiple unknown degradations, we present further results in Sec. C.1.
- To further validate the effectiveness of our proposed approach, we present additional qualitative results in Sec. C.2, including seven image restoration tasks: denoising, super-resolution, deblurring, deraining, dehazing, low light enhancement, and deraindrop.
- In Sec. C.3, we conduct additional ablation studies focusing on the semantic-agnostic constraint and the structural-correction module, providing deeper insights into their contributions.
- In Sec. D, we provide full implementation details of our methods and corresponding experiments to for reproducing our results.
- In Sec. E, we conduct a user study specifically targeting the real-world multiple degradation image restoration task. This study aims to provide additional evidence regarding the perceptual quality and effectiveness of AutoDIR in comparison to alternative methods.
- In Sec. F, we demonstrate the effectiveness of AutoDIR on real-world unseen Super-resolution and Denoise datasets.
- In Sec. G, we evaluate SA-BIQA on real-world unseen datasets to illustrate the robustness of SA-BIQA.

A Open-Vocabulary User Interaction

As shown in Fig. 1, AutoDIR provides a customizable approach to tailor the result outputs based on user preferences. Users can effectively modify the input image by providing corresponding **open-vocabulary** text prompts. The support of user interaction highlights the flexibility and adaptability of our proposed approach, allowing for a highly customizable image enhancement experience.

B Trade-off of adjustable coefficient w

The value of w for the weight of the structural correction module introduced in Sec. 3 of the main body determines the extent to which contextual information is utilized to recover the final result. Fig. 2 demonstrates that a larger value of w helps to recover the complex structures e.g. human face, of the original image. On the opposite, for tasks like super-resolution, a smaller value of w is required to maintain the generation capability of the latent diffusion model.

C Extensive Experimental Results

This section provides additional visualization and experimental details on datasets and training settings in the main text's Sec. 4.

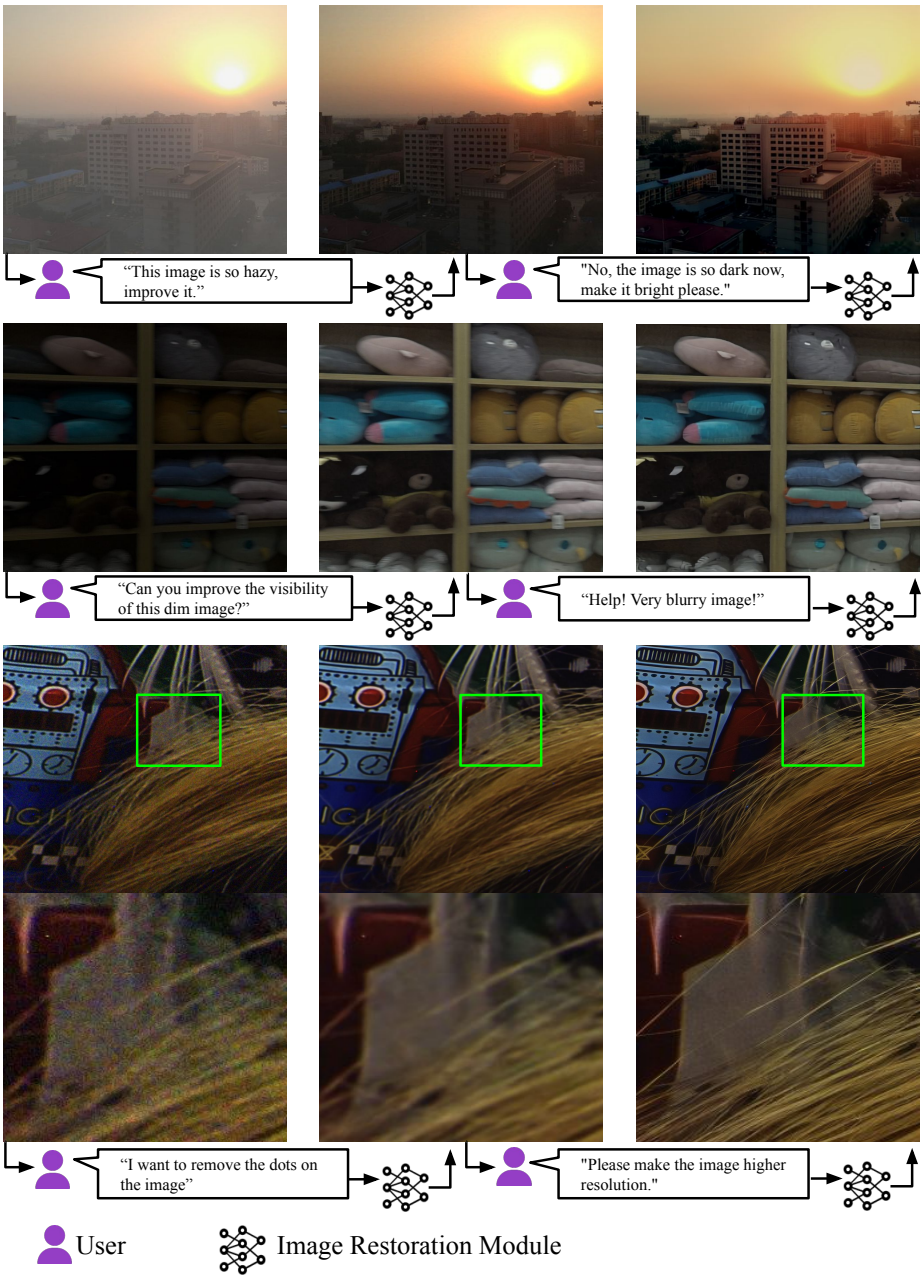

Fig. 1: User specified results. Users can edit the image according to their preference via **open-vocabulary** text instructions. The first column: original image I_0. The second column: the edited image I_1 after the user's first instruction given image I_0. The third column: the second edited image I_2 after the user's second instruction given the image I_1.

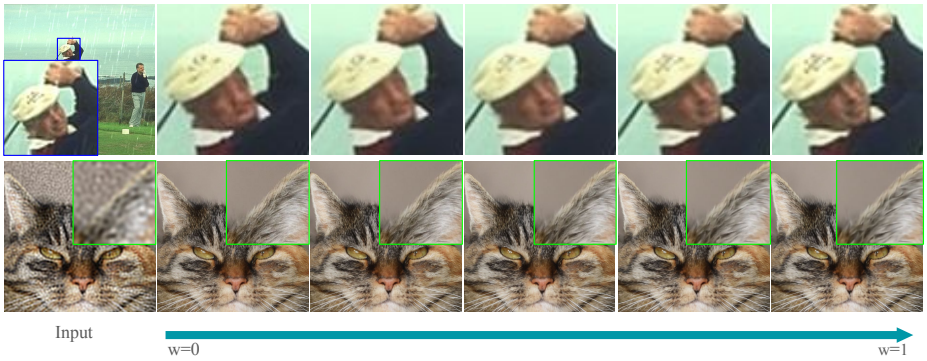

Fig. 2: Trade-off of the adjustable coefficient w for structural-correction model. The first row demonstrates that large w can recover the structural details of the original image. Conversely, the second row shows that a smaller w can maintain the generation capability of the generative latent diffusion model.

C.1 Results on Images with Multiple Unknown Degradations in Unseen Real-world Datasets

We present additional visualization results in Fig. 3 and Fig. 4 on images with multiple unknown degradations on **unseen real-world** UCD [24], EVUP [6], LOL-Blur [23] and RainDS [15] datasets, to further demonstrate the performance of our method in handling such complex scenarios.

C.2 Results on Seven Joint-learned Tasks

We provide more visualization results on the seven image restoration tasks in Fig. 5, 6, 7, 8, 9 and we also provide the zoom-in visualization results in the main text in Fig. 10, 11, 12, 13, 14, 15.

C.3 Ablation Studies

Fine-tuning with Semantic-agnostic Constraint for BIQA and Mechanism of AutoDIR In order to further analyze the effectiveness of our semantic-agnostic constraint, we visualize the attention map of the image encoder in Fig. 16. As depicted in the visualization, before fine-tuning, the attention map primarily focuses on the pronominal object, which is consistent with the behavior of the original CLIP model [16] that was pre-trained on image classification tasks.

After fine-tuning with semantic-agnostic constraint, we observe that the attention maps expand to highlight background areas that may contain potential artifacts which shows that the BIQA model has successfully learned to prioritize and focus on artifacts, leading to more accurate BIQA results.

We further show the t-SNE visualization of image embeddings $\mathcal{E}_{\mathcal{I}}(I)$ with seven types of degradations to demonstrate the mechanism of AutoDIR handling images with unknown degradations. For example, the images captured by

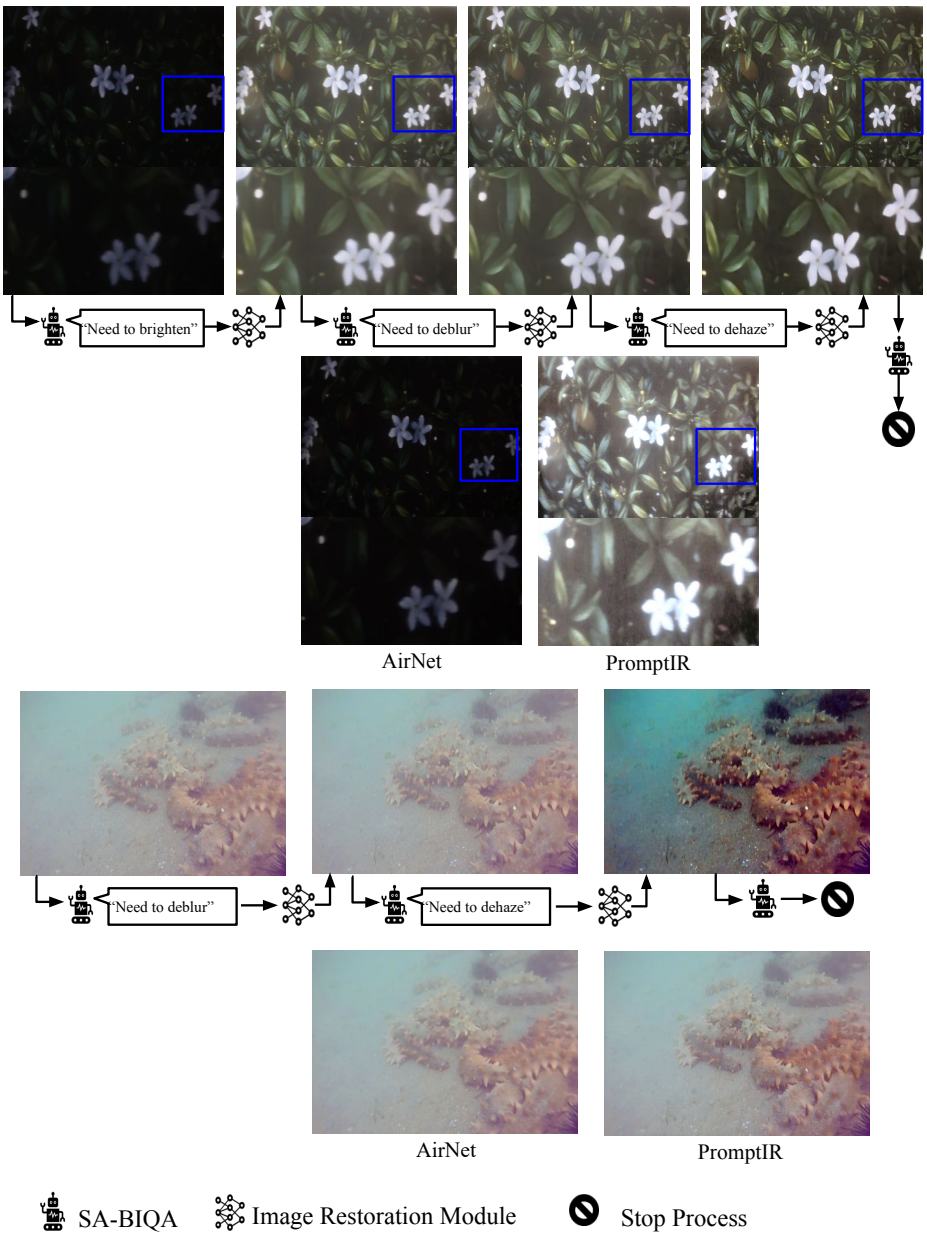

Fig. 3: Handling images of multiple **unknown** artifacts in the **unseen real-world** datasets.

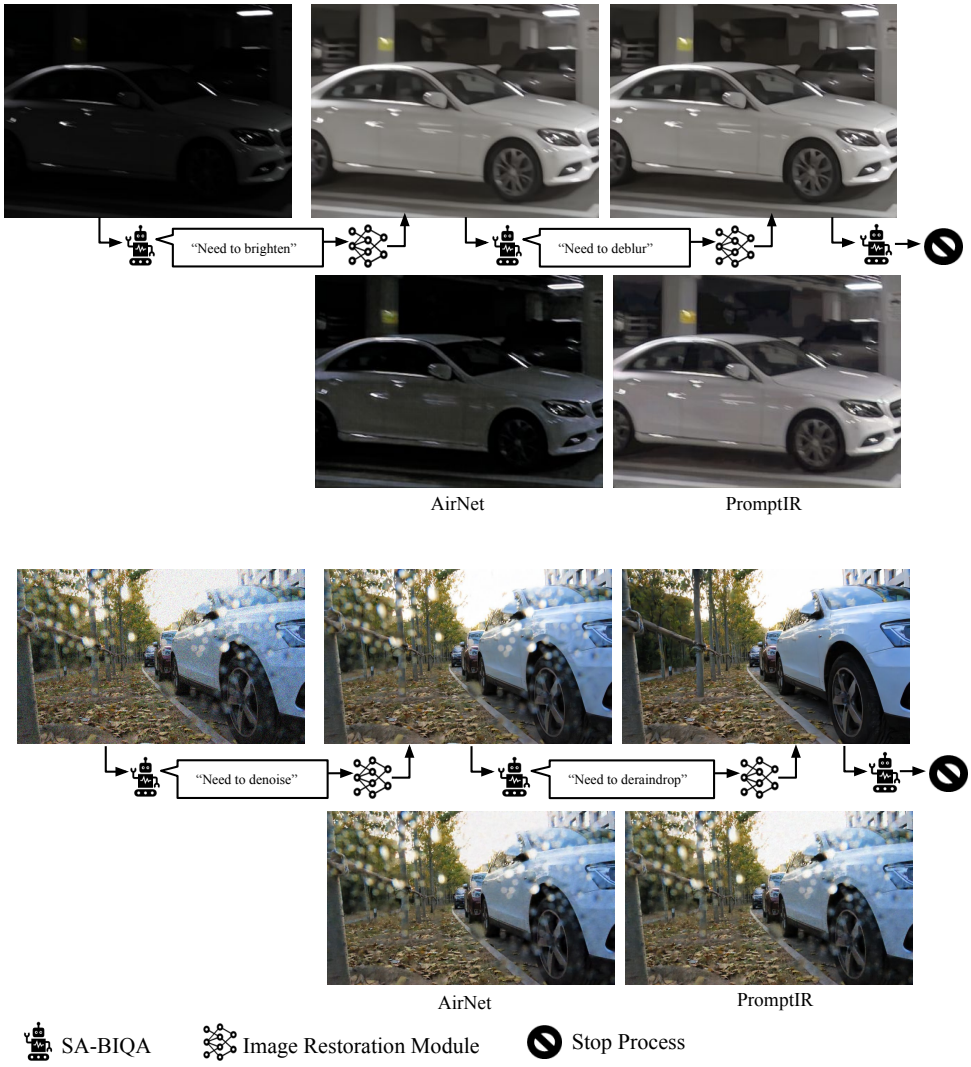

Fig. 4: Handling images of multiple **unknown** artifacts in the **unseen real-world** datasets.

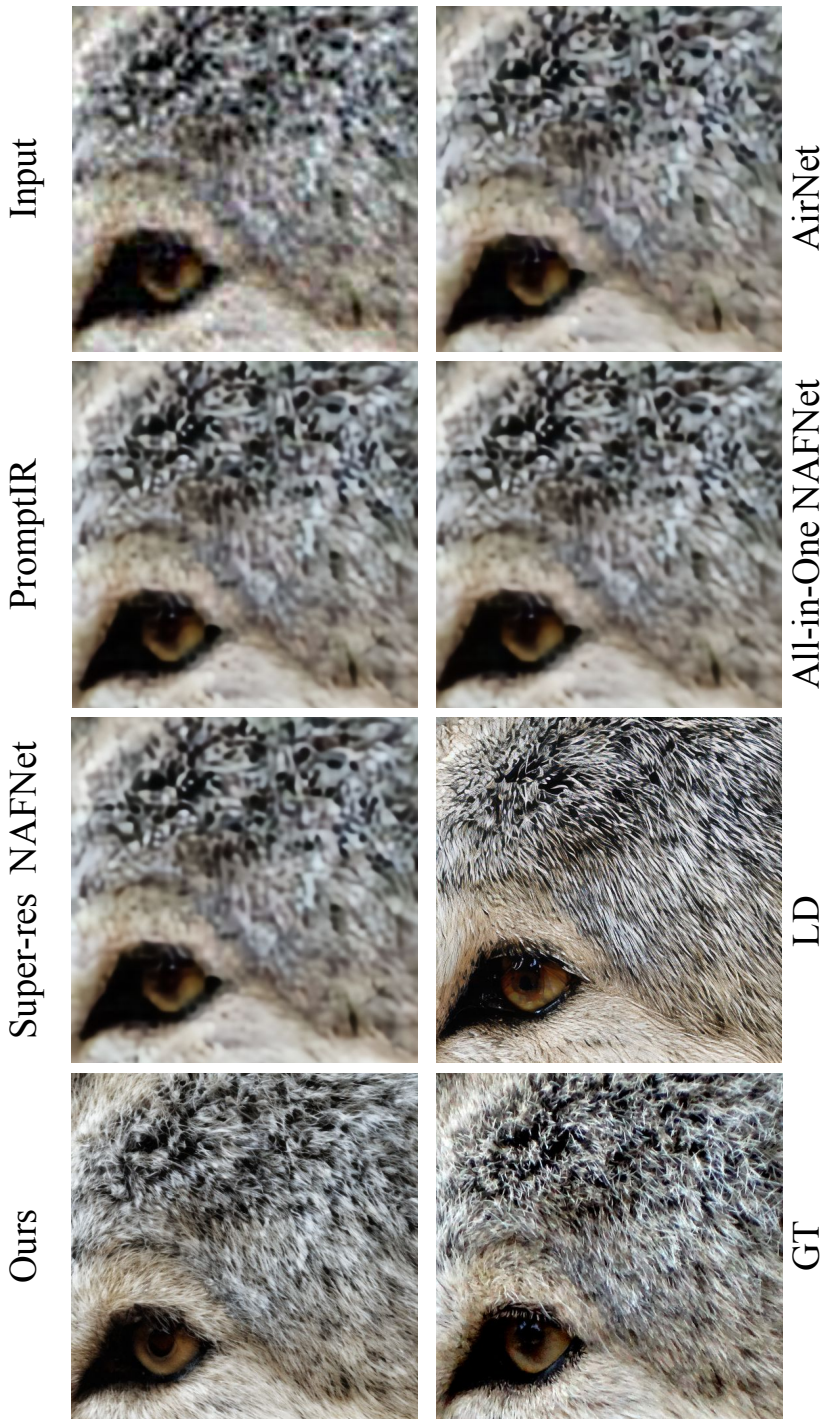

Fig. 5: Qualitative comparisons on Super-Resolution.

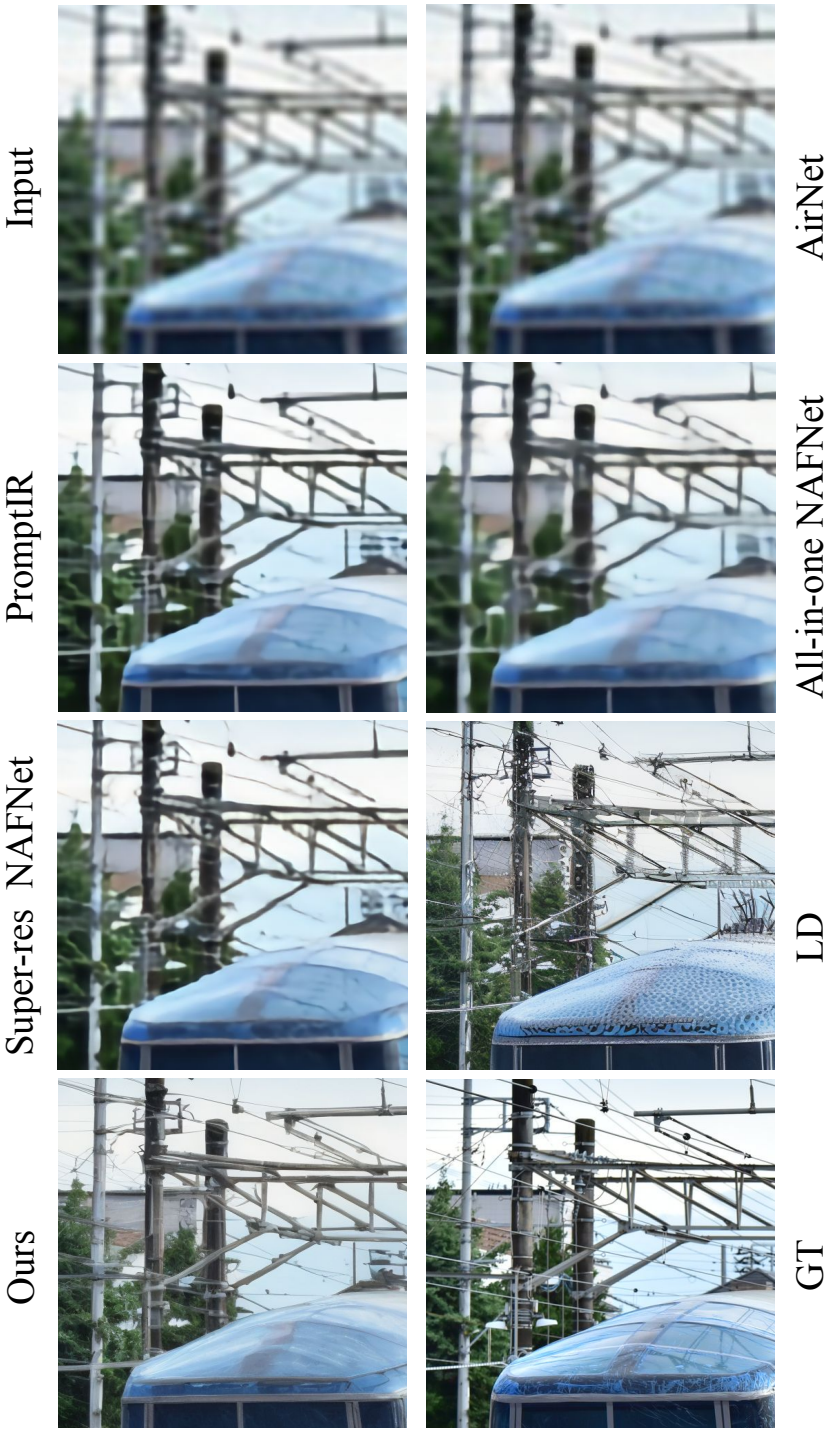

Fig. 6: Qualitative comparisons on Super-Resolution.

Fig. 7: Qualitative comparisons on Super-Resolution with state-of-the-art all-in-one methods.

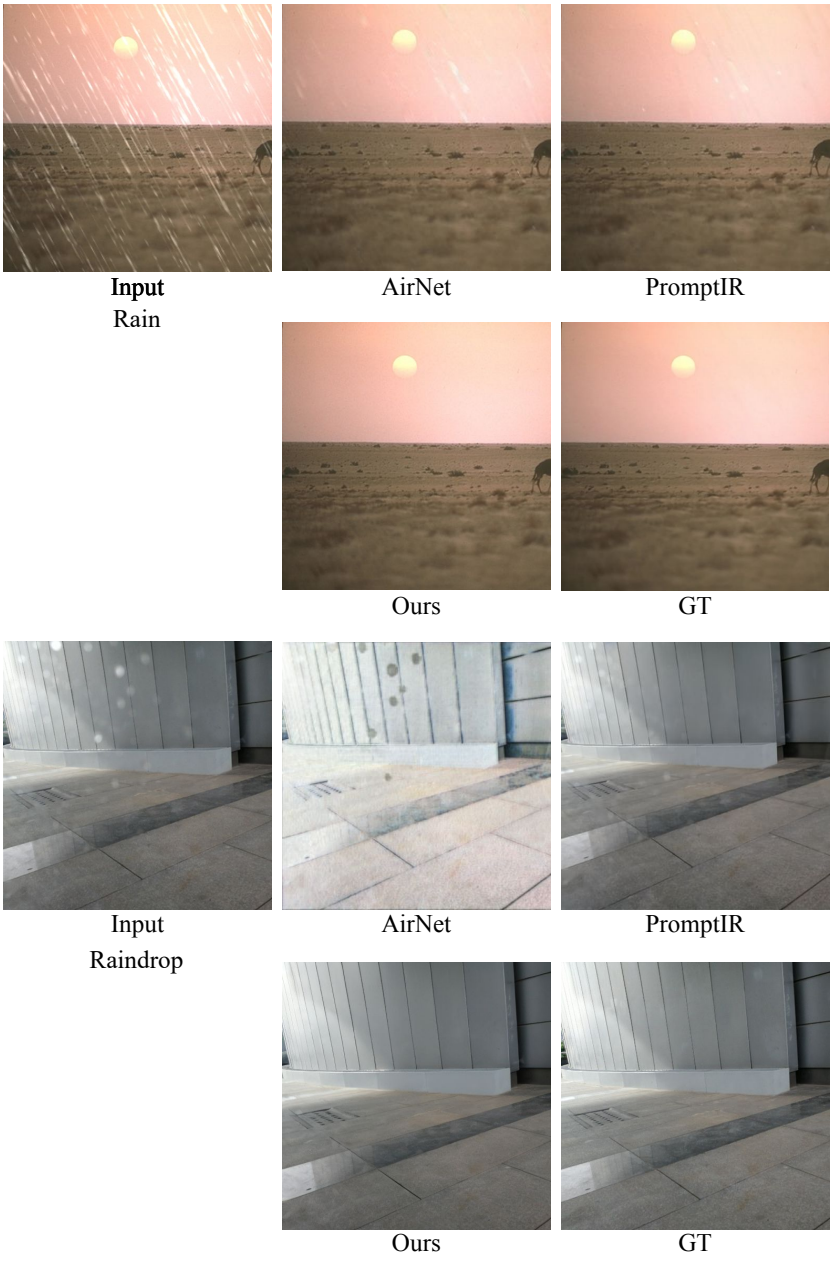

Fig. 8: Qualitative comparisons on derain and deraindrop with state-of-the-art all-in-one methods.

Fig. 9: Qualitative comparisons on low light enhancement and dehazing with state-of-the-art all-in-one methods.

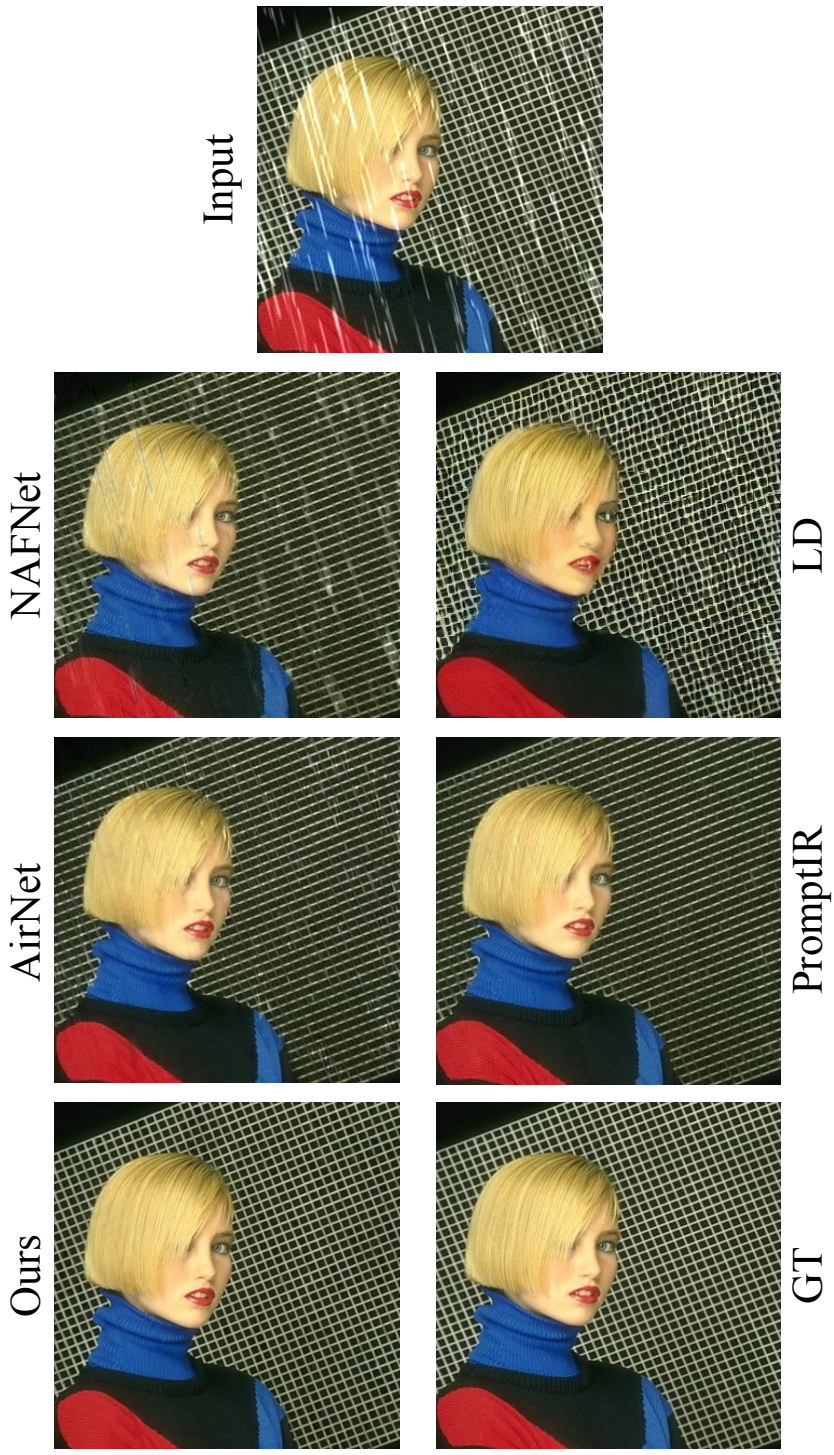

Fig. 10: Zoomed-in deraining results in the main text.

Fig. 11: Zoomed-in dehazing results in the main text.

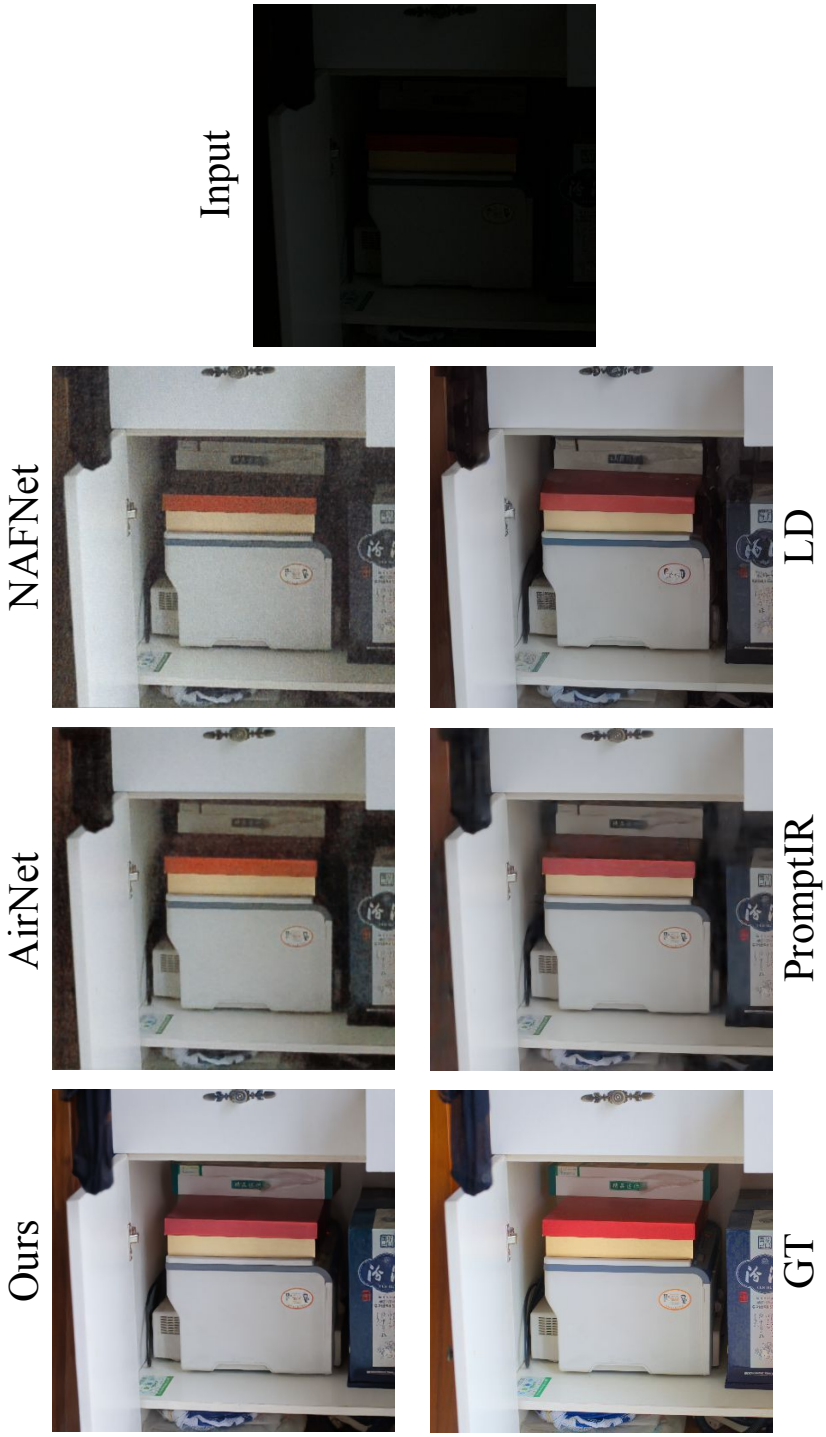

Fig. 12: Zoomed-in low light enhancement results in the main text.

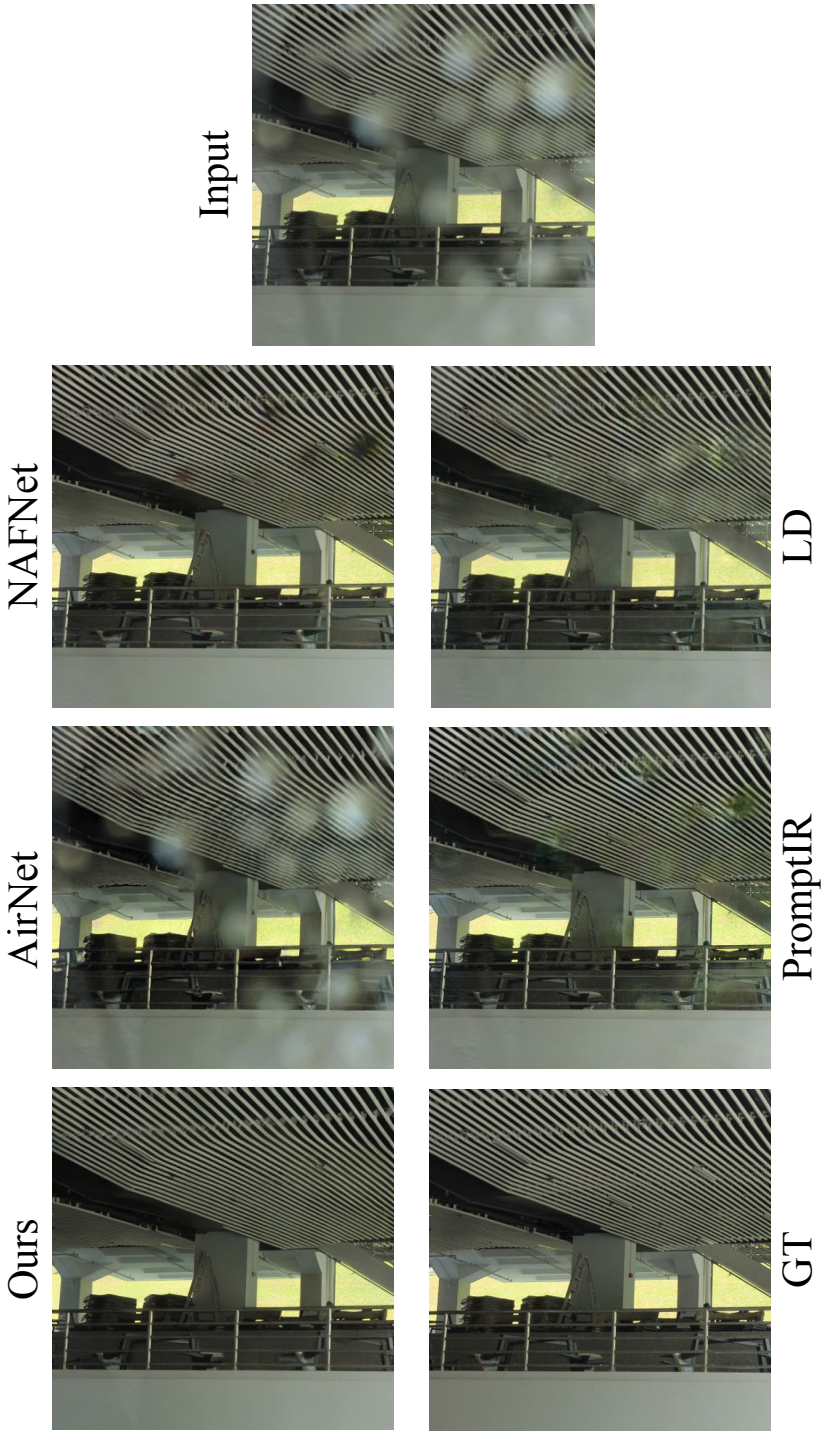

Fig. 13: Zoomed-in deraindrop results in the main text.

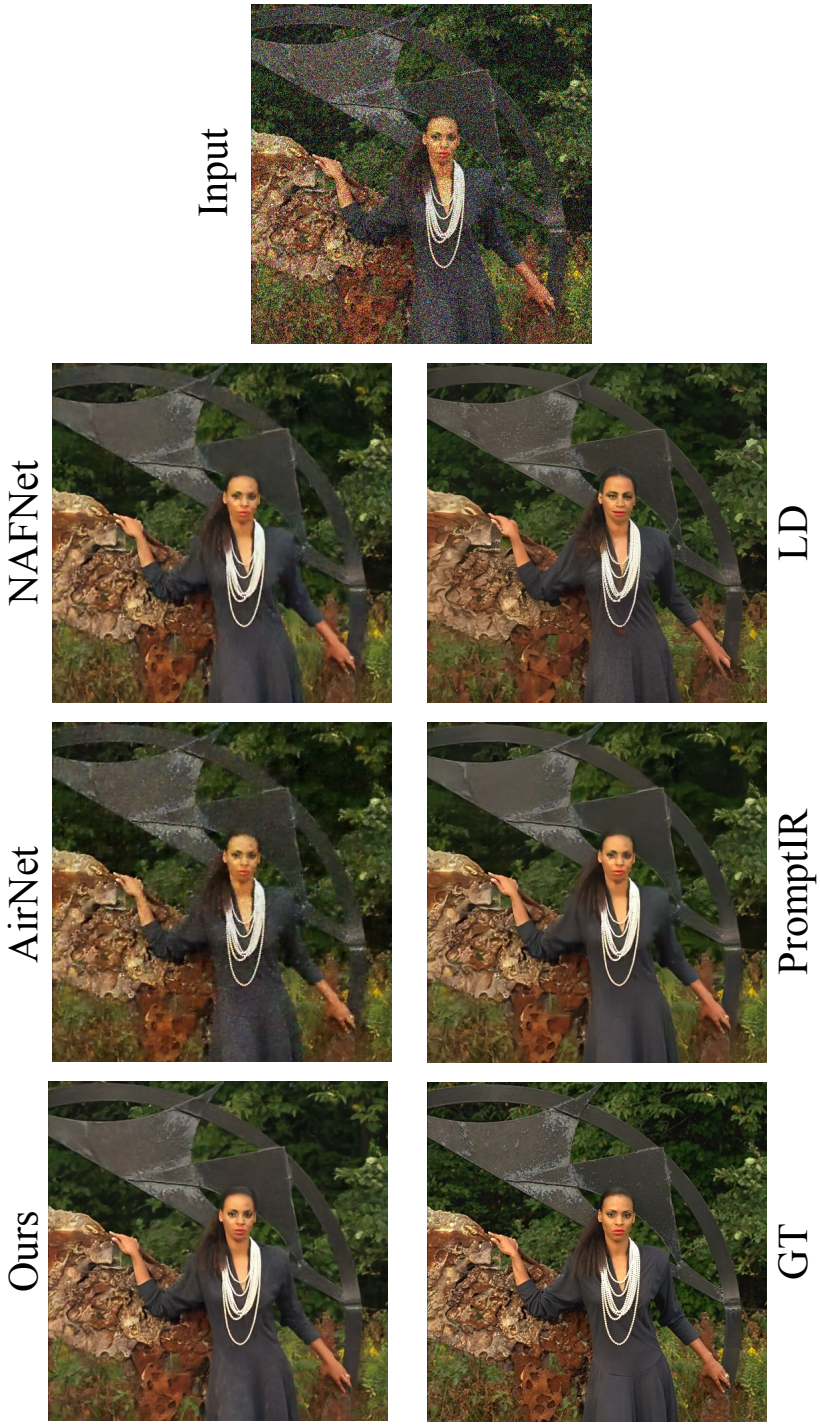

Fig. 14: Zoomed-in denoise results in the main text.

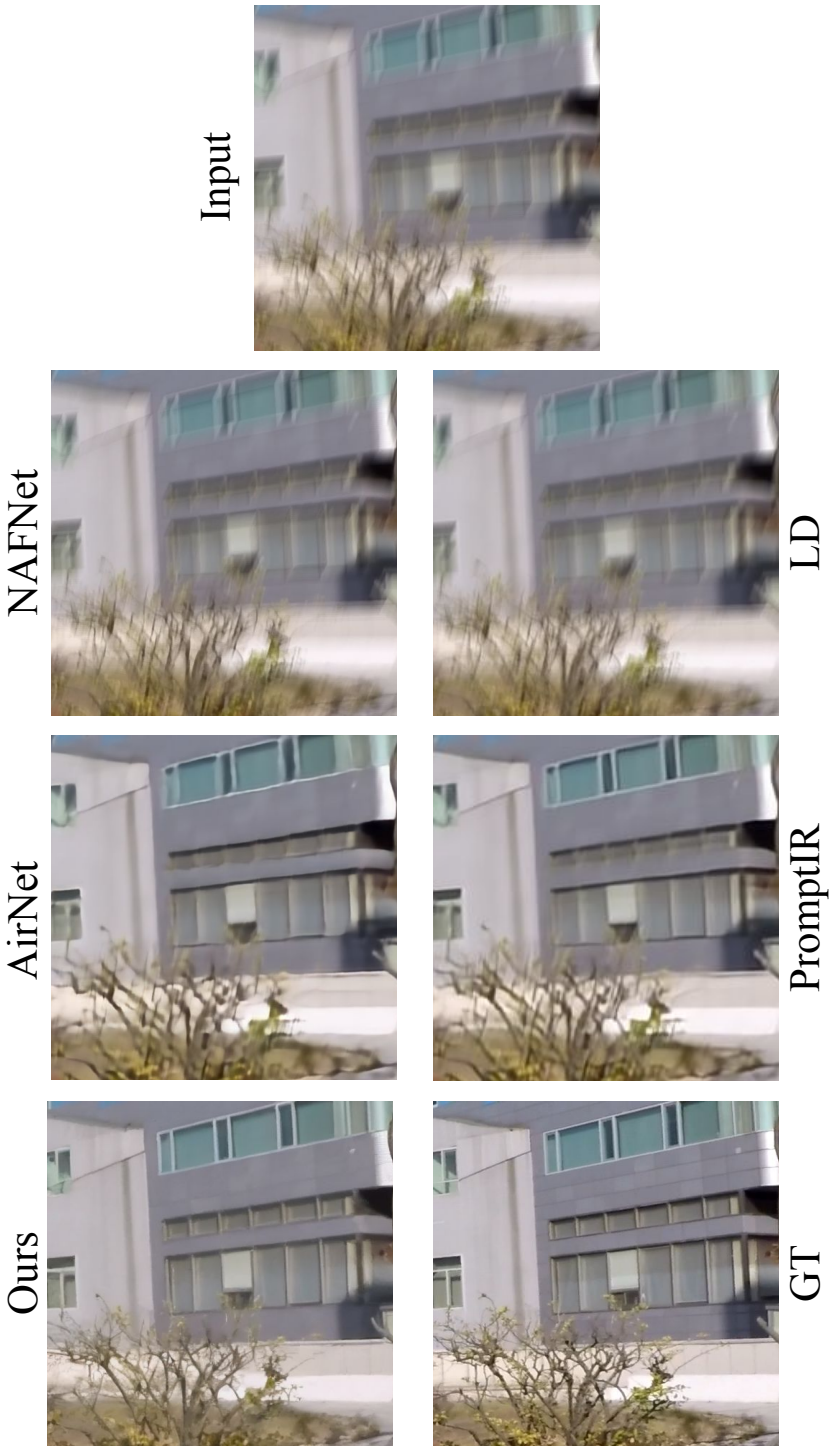

Fig. 15: Zoomed-in deblur results in the main text.

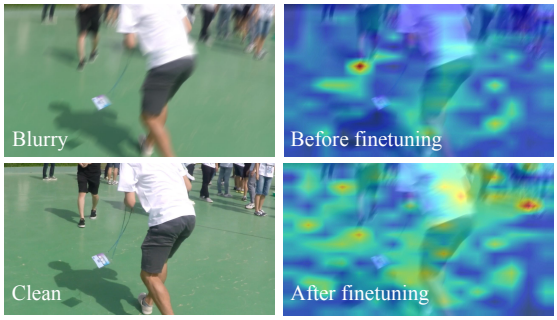

Fig. 16: Self-attention maps of SA-BIQA image encoder. Attention maps focus on (have the highest attention values) the foreground object in BIQA image encoder without fine-tuning, while attention maps focus on (have the highest attention values) both foreground and background objects in SA-BIQA image encoder fine-tuned with SA constraint.

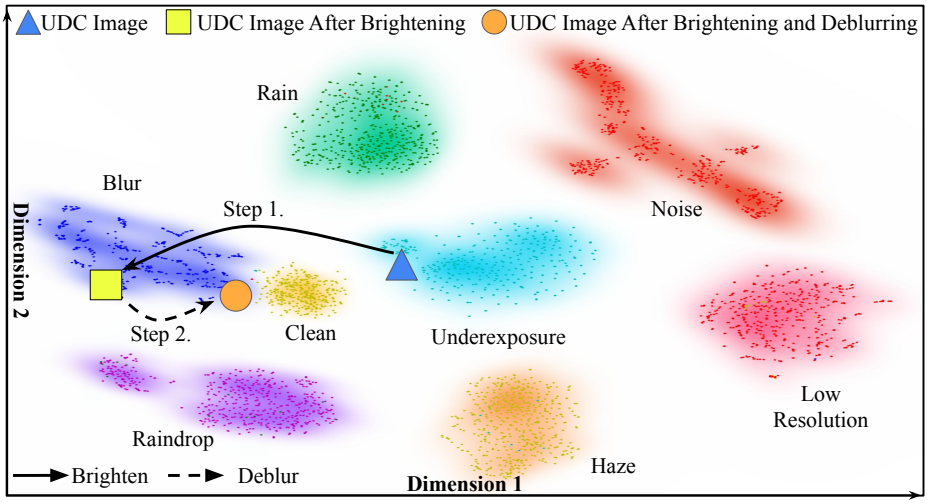

Fig. 17: t-SNE visualization of image embeddings $\mathcal{E}_{\mathcal{I}}(I)$ with seven types of degradations, which illustrates the space of common image degradations. Images captured by Under-Display Cameras (UDC) [24] suffer from both blur and underexposure. AutoDIR automatically decides (via SA-BIQA) that the first step is to improve "underexposure", and the second step is to remove "blur", moving the input images towards the region of clean images.

Under-Display Cameras (UDC) [24] suffer from both blur and underexposure. AutoDIR automatically decides (via BIQA) the first operator is to improve "underexposure", and the second operator is to remove "blur", moving the input images towards the region of clean images.

Importance of Structural-Correction Module (SCM) for latent diffusion The generative Latent Diffusion model [17] has demonstrated a strong ability to generate unseen features. However, it falls short when preserving the original structural information of the input image, which is crucial for image enhancement tasks.

As illustrated in Fig. 18, Structural-Correction Latent-Diffusion has high-quality results with fine details intact. On the other hand, latent diffusion exhibits significant distortion in faces and text. Moreover, Fig. 19 demonstrates that the structural-correction module also shows the ability to correct hallucinated undesire textures of the results of the latent diffusion.

BIQA performance comparison of ViT and CLIP-variants. As shown in Tab. 1, we report the F-scores of seven degradation tasks, our SA-CLIP outperforms ViT Classifier, pre-trained CLIP, and naively finetuned CLIP in all the seven tasks.

Table 1: F-Score of image degradation detection on seven degradation tasks and clean image.

Degradation	Haze	Blur	Rain	LOL	Raindrop	Noise	Low-Res
ViT-classifier	0.6666	0.7714	0.9603	0.7692	1.0000	0.6934	0.8450
Original CLIP	0.6738	0.7206	0.8889	0.8823	0.7568	0.8049	0.7229
Fine-tuned CLIP	0.6955	0.7744	0.9135	0.9091	0.9464	0.7463	0.9085
SA-CLIP (ours)	**1.0000**	**0.9444**	**0.9950**	**1.0000**	**1.0000**	**1.0000**	**0.9985**

D Implementation Details

Our experiments are conducted using PyTorch on a computational setup comprising eight NVIDIA V100 GPUs. The training process for the AutoDIR framework involves three distinct steps. Firstly, we initialize the training by freezing the text encoder and fine-tuning the image encoder within the Blind Image Quality Assessment (BIQA). We utilize the Adam optimizer with a batch size of 1024 and train for 20 epochs. The initial learning rate is set to 3×10^{-6} and follows a cosine annealing rule. Next, we proceed to fine-tune the All-in-One Image Restoration (AIR) backbone using the Adam optimizer. During this stage, we employ a learning rate of $1e^{-4}$ and a batch size of 256. The fine-tuning process is performed for 15 epochs. Finally, we freeze the previously trained pipeline components and focus on training the structural-correction module (SCM). For this stage, we employ the Adam optimizer and a cosine annealing rule. The initial learning rate is set to $1e^{-3}$, and the batch size is 256. We train the SCM for 8,000 iterations. In our experiments, the structural-correction module for Structural-Correction-Latent Diffusion is based on NAFNet architecture [3]. During inference, the coefficient w of the structural-correction module is set to be 1 as default

for denoising, deraining, dehazing, deraindrop, low light enhancement, and deblurring tasks and 0.1 for the super-resolution task to maintain the generation capability of the generative latent diffusion model.

Datasets: The seven image restoration tasks are denoising, deblurring, super-resolution, low-light enhancement, dehazing, deraining, and deraindrop. For denoising, we use SIDD [1] and a synthetic Gaussian and Poisson noise dataset with DIV2K [2] and Flickr2K [11]. For super-resolution, we follow previous practice and train AutoDIR with DIV2K [2] and Flickr2K [11] training sets, following [18] for degraded image generation. In addition, we use GoPro [12], LOL [20], RESIDE [9], Rain200L [22], and Raindrop [14] for deblurring, low-light enhancement, dehazing, deraining, and deraindrop, respectively. During inference, we evaluate multiple test sets. These include SIDD [1], Kodak24 [4], DIV2K [2], GoPro [12], LOL [20], SOTS-Outdoor [9], Rain100 [22], and Raindrop [14], each corresponding to their respective tasks. For experiments with unknown degradations, we use the Under-Display Camera (TOLED) dataset [24] and the Enhancing Underwater Visual Perception (EUVP) dataset [6].

E User Study

To further examine the effectiveness of AutoDIR, we conduct a user study on the images with unknown degradations in unseen real-world datasets or real-captured images. We compare AutoDIR with state-of-the-art all-in-one AirNet [10] and PromptIR [13]. As shown in Fig. 20, given the input and the restored results, the question is to ask which image has the best visual, and the choices are in random order. We collect 22 forms and there are $22 \times 28 = 616$ responses in total. Fig. 21 illustrates that AutoDIR gathers more than 96% of the votes for producing the best denoising results.

F Comparison on unseen real-world Super-res and Denoise dataset

We conduct experiments on the unseen real-world Super-Res dataset (RealSR test dataset [8]) without specific fine-tuning. As illustrated in Fig. 22, AutoDIR successfully reconstructs details such as eyebrows and beards, which other methods struggle to achieve. Additionally, we have included the quantitative results in Tab. 2.

As shown in Fig 23 and Tab. 3, we have also presented quantitative and qualitative results on the unseen real-world denoise dataset (PolyU-Denoise [21]), demonstrating the benefits of the AutoDIR approach.

G Evaluation of SA-BIQA on unseen real-world datasets.

We have conducted experiments on unseen real-world datasets, including Super-Res (RealSR [8]), Deblur (RealBlur [7]), and low-light enhancement (HuaWei [5]).

Fig. 18: Qualitative comparisons for latent diffusion model and LC-latent diffusion (ours) on dehazing and deraindrop tasks.

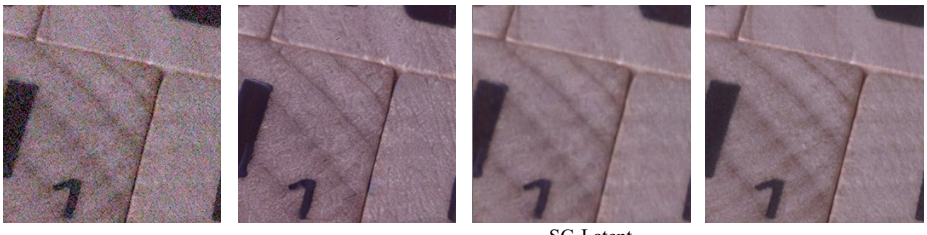

Fig. 19: Qualitative comparisons for latent diffusion model and LC-latent diffusion (ours) on denoising tasks.

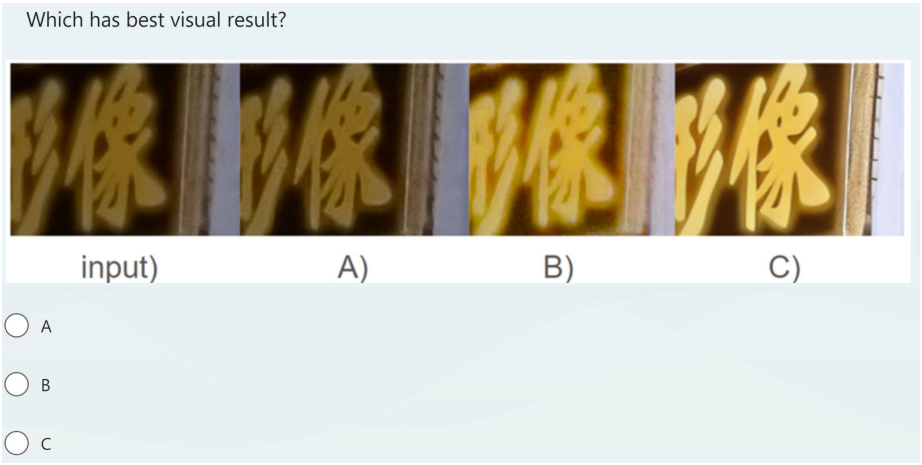

Fig. 20: Screenshot of the user interface in the user study.

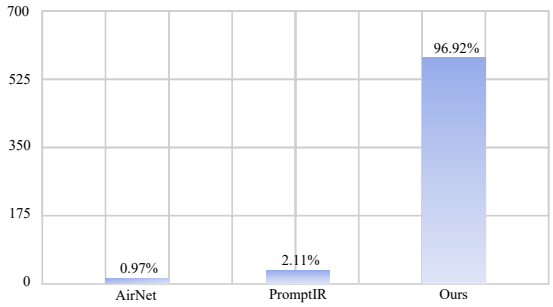

Fig. 21: Results of the user study on 28 real-world images with unknown artifacts on unseen datasets or real-captured images. We collected 22 forms and $28 \times 22 = 616$ responses in total. Among them, AutoDIR receives more than 96% of the votes as the best results.

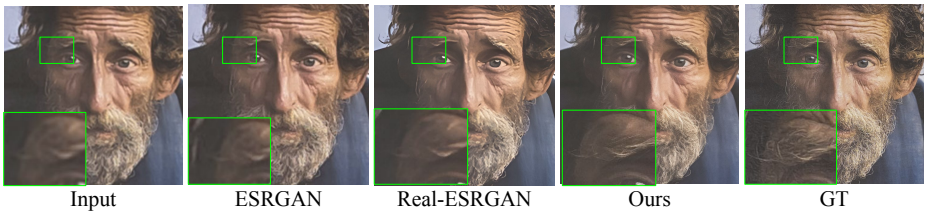

Fig. 22: Comparision on RealSR dataset with task-specific methods ESRGAN [19] and Real-ESRGAN [18].

The Tab. 4 demonstrates that SA-BIQA outperforms other methods by accu-

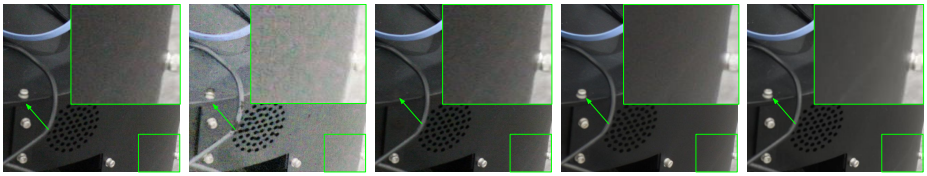

Fig. 23: Comparision on PolyU-Denoise dataset with all-in-one methods.

Table 2: Quantitative comparison on RealSR

Method	Real-World Super Resolution			
	MUSIQ ↑	CLIP-IQA ↑	NIQE ↓	NIMA ↑
NAFNet-SR	40.53	0.251	7.222	4.247
AirNet	21.73	0.239	11.839	3.773
PromptIR	24.85	0.248	8.526	4.082
ESRGAN	30.10	0.231	7.819	3.942
Real-ESRGAN+	<u>60.11</u>	<u>0.462</u>	**5.130**	<u>4.660</u>
Ours	**60.14**	**0.493**	<u>5.204</u>	**4.717**

Table 3: Quantitative comparison on PolyU-Denoise

Method	Real-world Denoise		
	SSIM↑	PSNR↑	LPIPS↓
NAFNet	0.924	34.16	<u>0.130</u>
AirNet	0.783	26.26	0.293
PromptIR	<u>0.933</u>	<u>34.43</u>	0.176
LD	0.910	32.85	0.196
Ours	**0.942**	**36.18**	**0.129**

rately predicting the dominant artifact in all tasks, achieving an accuracy of over 90%

Table 4: Quantitative comparison of SA-BIQA on real-world tasks.

Method	low-res	blur	low-light
ViT-classifier	0.9100	0.0357	<u>0.7666</u>
Original CLIP	0.8500	<u>0.6704</u>	0.2333
Fine-tuned CLIP	<u>0.9200</u>	0.4795	0.7000
SA-CLIP (ours)	**1.0000**	**0.9010**	**0.9333**

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
3. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Proceedings of European Conferences on Computer Vision (ECCV) (2022)
4. Franzen, R.: Lossless true color image suite. http://r0k.us/graphics/kodak/ (1999)
5. Hai, J., Xuan, Z., Yang, R., Hao, Y., Zou, F., Lin, F., Han, S.: R2rnet: Low-light image enhancement via real-low to real-normal network. Journal of Visual Communication and Image Representation (2023)
6. Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE Robotics and Automation Letters (RA-L) (2020)
7. Jaesung Rim, Haeyun Lee, J.W.S.C.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: Proceedings of European Conferences on Computer Vision (ECCV) (2020)
8. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
9. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing (TIP) (2018)
10. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-In-One image restoration for unknown corruption. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
11. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
12. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2023)
14. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Quan, R., Yu, X., Liang, Y., Yang, Y.: Removing raindrops and rain streaks in one go. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning (ICML) (2021)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

18. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
19. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of European Conferences on Computer Vision (ECCV) Workshops (2018)
20. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. Proceedings of The British Machine Vision Conference (BMVC) (2018)
21. Xu, J., Li, H., Liang, Z., Zhang, D., Zhang, L.: Real-world noisy image denoising: A new benchmark. arXiv preprint arXiv:1804.02603 (2018)
22. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
23. Zhou, S., Li, C., Change Loy, C.: Lednet: Joint low-light enhancement and deblurring in the dark. In: Proceedings of European Conferences on Computer Vision (ECCV) (2022)
24. Zhou, Y., Ren, D., Emerton, N., Lim, S., Large, T.: Image restoration for under-display camera. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)