– Supplementary Material –
Detail-Preserving Latent Diffusion for Stable Shadow Removal

Jiamin Xu[1], Yuxin Zheng[1], Zelong Li[1], Chi Wang[1], Renshu Gu[1], Weiwei Xu[2], Gang Xu[1*]
[1] Hangzhou Dianzi University [2] Zhejiang University
superxjm@yeah.net, yuxin6@hdu.edu.cn, jokerli@hdu.edu.cn, wangchi1995@zju.edu.cn,
renshugu@hdu.edu.cn, xww@cad.zju.edu.cn, gxu@hdu.edu.cn

In the supplementary material we present the following:
- Additional implementation details.
- Additional experiments and ablations.

A. Additional Implementation Details

In our approach, we utilize Stable Diffusion v2 [6] in Diffusers [8], setting the text prompt to an empty string for both training and testing phases. During the LDM fine-tuning stage (stage one), we adapt the U-Net by increasing its input channels from 4 to 8. These 8 channels include 4 for noise and 4 for the latent representation of the conditioning image. The additional parameters in the first layer are initialized by duplicating the original parameters and downscaling them by a factor of 2.

To adapt our method for large-size inputs, such as those in the WSRD+ dataset (1920×1440), we make slight modifications to our two-stage approach. In the first stage, the input images are downscaled to $W/k \times H/k$, with $k = 3$ for the WSRD+ dataset. We use smaller images in this stage to ensure high-quality shadow removal, accepting some loss of detail while prioritizing the capture of global contextual information. As a result, we resize the images rather than cropping them into local patches.

In the second stage, the input consists of the latent generated from the downscaled image in the first stage. The original image of shape $W \times H \times C$ is first reshaped to $\frac{W}{k} \times \frac{H}{k} \times Ck^2$, where the features of each $k \times k$ region are flattened into a vector. Using the VAE decoder with our Detail Injection model, the output is a feature map of shape $\frac{W}{k} \times \frac{H}{k} \times Ck^2$, which is then reshaped to the original dimensions, resulting in a large-size shadow-free image. Here, the input and output channels of the VAE are expanded from 3 to $3k^2$. The additional parameters in the first and last layers are initialized by duplicating and downscaling the original parameters. With this design, the features from the VAE encoder capture details from the large-size image, which are then injected into the decoder through the RRDB network to enhance its detail recovery. During training on the WRSD+ dataset [7], we use 400 epochs for the first stage and 100 epochs for the second stage, with a batch size of 16. The results, presented in Fig. A, demonstrate that our method effectively removes both hard and soft shadows from the high-resolution images.

Figure A. **Our results on large-size inputs from the WSRD+ dataset** (1920×1440)**.**

B. Additional Experiments and Ablations

B.1. The MAE for shadow and non-shadow regions.

In Table A, we report the MAE results on the ISTD+, SRD, and INS datasets, with 'S' for shadow regions and 'NS' for non-shadow regions. For the INS dataset, we generate the binary mask using a 0.2 probability threshold. Our method

*Corresponding author.

achieves the lowest MAE on the INS dataset. On the ISTD+ dataset, our method performs the best among mask-free methods, only slightly behind ShadowFormer, ShadowDiffusion, and HomoFormer, which use shadow masks. We also observe that our MAE in the shadow regions is significantly lower than that of other methods, which suggests that our approach is more effective at completely removing shadows. For the SRD dataset, among mask-free methods, our method's MAE is just slightly higher than DeS3 (+0.01). Again, it performs strongly in the shadow regions.

Method	MAE↓(ISTD+)	MAE↓(SRD)	MAE↓(INS)
	S/NS/ALL	S/NS/ALL	S/NS/ALL
AutoExposure	6.5/3.8/4.2	8.55/5.74/6.50	9.56/2.41/3.45
Zhu et al.	—/—/—	7.44/3.74/4.79	—/—/—
BMNet	5.6/2.5/3.0	6.61/3.61/4.46	9.51/2.34/3.38
ShadowFormer	5.2/2.3/2.8	5.90/3.44/4.04	8.36/2.27/3.16
DMTN	6.1/2.6/3.2	5.92/3.03/3.82	8.31/2.35/3.24
ShadowDiffusion	4.9/2.3/2.7	4.98/3.44/3.63	8.08/2.81/3.59
HomoFormer	5.0/2.2/2.6	4.33/2.76/3.29	8.00/2.24/3.09
InstanceShadow	5.1/2.5/2.9	6.53/3.76/4.33	8.75/3.86/4.63
TBRNet	6.6/3.3/3.8	7.68/4.89/5.57	—/—/—
Refusion	6.2/3.3/3.8	6.89/4.36/4.87	8.73/2.69/3.59
DeS3	6.5/3.3/3.9	5.88/2.83/3.72	8.86/3.29/4.08
OmniSR	6.6/2.4/3.1	6.77/3.70/4.33	6.96/2.11/2.82
Ours	4.4/2.6/2.9	5.33/3.32/3.73	6.50/2.18/2.82

Table A. The MAE results on ISTD+, SRD and INS datasets.

B.2. Performance comparisons

In Table B, we present the inference time, number of parameters, and MACs (for a 640×480 image), along with comparisons to several other methods. Our method achieves an inference time of 781.6ms using FP16 by default on an RTX 4090 GPU and 2182.4ms using FP32.

	Ours (FP16)	OmniSR	Shadow-Former	Shadow-Diffusion	Instance-Shadow	DeS3
Time (ms)	781.6	120.1	43.7	506.9	3070.8	1233.6
#para (MB)	1329.8	128.7	11.4	55.5	262.6	108.4
GMACs	4561.4	628.9	343.3	855.3	2826.0	1364.0

Table B. The performance comparisons.

B.3. More results on the ISTD+ [3], SRD [5], and INS dataset [10]

In this section, we present the comparison results on the ISTD+ [3], SRD [5], and INS [10] datasets in Fig. B. Our shadow removal method outperforms others, including those that rely on detected shadow masks, such as HomoFormer [9] and ShadowDiffusion [1], as well as methods that do not require shadow masks, such as Refusion [4], Des3 [2], and OmniSR [10]. As shown in Fig. B, in the outdoor dataset, our method successfully removes hard shadows. In the INS [10] indoor dataset, our method effectively removes complex shadows, such as those under the table and the thin shadow near the sofa.

B.4. More results from our first and second stages

We present additional results from the first and second stages, along with the colored RRDB features of different detail injection models, spanning from decoder layer one to four. As shown in Fig. C, whether in outdoor or indoor settings, our detail injection model successfully injects shadow-free details into the decoder outputs, leading to shadow removal results that preserve intricate details. Additionally, by visualizing the RRDB features using PCA, we observe that shadow regions exhibit distinct colors compared to the surrounding areas. This suggests that the detail injection model in the second stage is able to learn to locate and remove shadows with the aid of the latent features from the first stage.

B.5. More results from the cross-dataset evaluation

As described in the main paper, we evaluate the generalizability of our method through cross-dataset testing. This includes training on ISTD+ and testing on SRD, training on SRD and testing on ISTD+, and training on INS (a synthetic dataset) while testing on WRSD+ (real-world captures). Notably, in both the ISTD+ and SRD datasets, the objects casting shadows in the training and testing splits are limited and often quite similar—such as the umbrellas shown in the third and fourth rows of Fig. D. Therefore, cross-dataset evaluation provides a more robust demonstration of each method.

As shown in Fig. D, when tested on a different dataset, methods like Refusion [4], ShadowDiffusion [1], DeS3 [2], and OmniSR [10] tend to leave residual shadows in their results. In contrast, our method achieves superior performance under this evaluation setting. Leveraging the generative priors of Stable Diffusion [6], our approach demonstrates strong generalizability to unseen shadow-casting objects and diverse background types, which is a critical requirement for practical real-world shadow-removal applications.

B.6. Evaluation on fixed and unfixed VAE decoder

In our second stage, we fixed the decoder parameters and modulated each layer's output features by incorporating features from the corresponding VAE encoder layer. In this section, we evaluate our method using an unfixed VAE decoder and compared it with the fixed version. The results revealed no significant differences between the two configurations (Table C). Therefore, we chose the fixed VAE decoder version for its advantage in reducing memory usage.

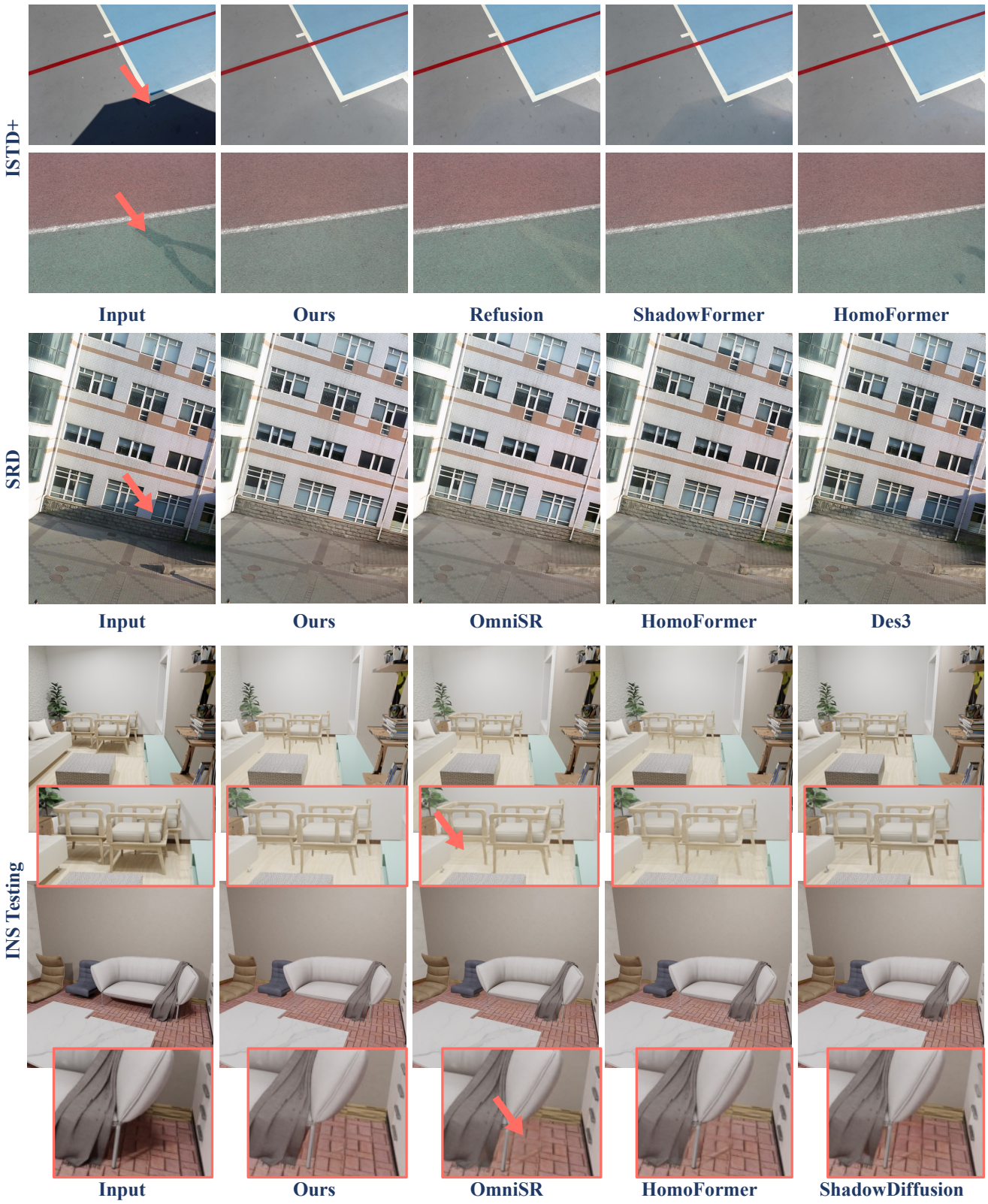

Figure B. **Additional comparison results on the ISTD+ [3], SRD [5], and INS dataset [10].**

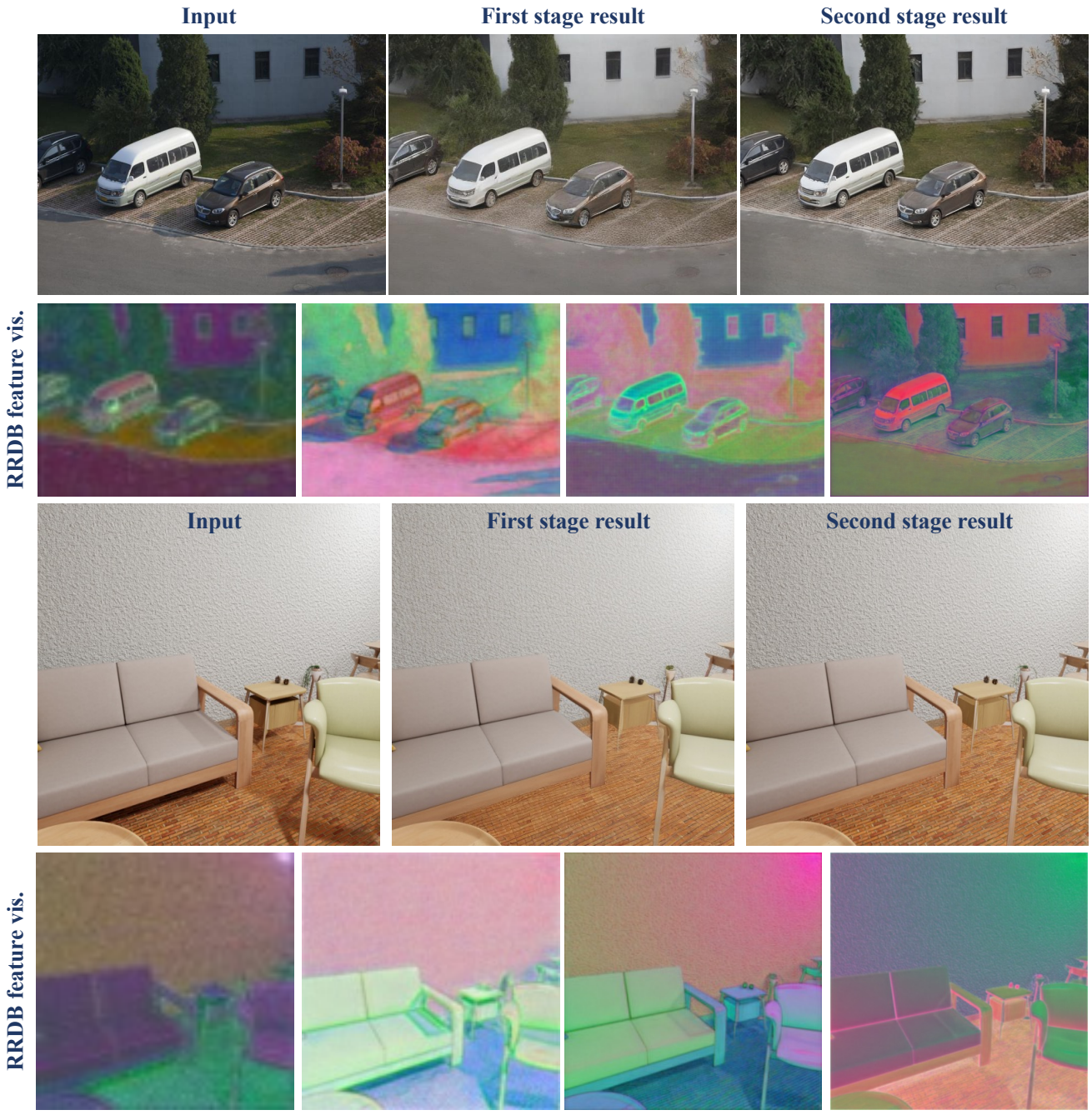

Figure C. **Additional results from our first and second stages, along with the visualization of RRDB features added to each decoder layer.**

B.7. Qualitative results of ablation studies

We present the qualitative results of the ablation studies. As shown in Fig. E, our full model achieves the best shadow removal performance compared to the other ablation configurations. In the "Wo/ DINO" ablation, we observe that DINO features help reduce some shadows, indicating that it can help filter out shadow-free information and inject it into the VAE decoder features during our Detail Injection stage.

References

[1] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for

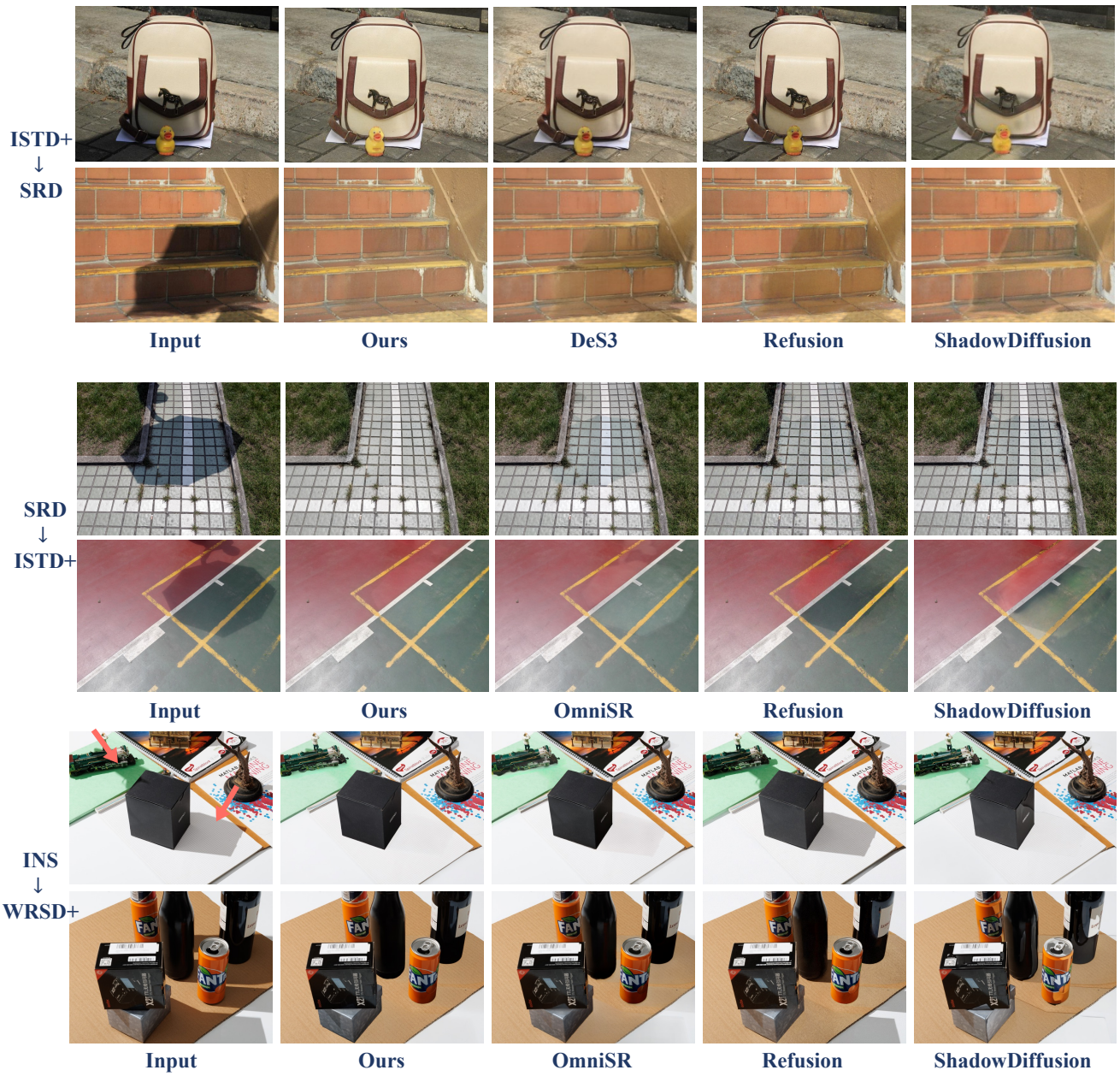

Figure D. **Additional results from our cross-dataset evaluation.**

Dataset	ISTD+ PSNR/SSIM	SRD PSNR/SSIM
Fixed VAE decoder	35.19/0.974	33.63/0.968
Unfixed VAE decoder	35.18/0.974	33.69/0.969

Table C. **Ablation study on fixed vs. unfixed VAE decoder.**

shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14049–14058, 2023. 2

[2] Yeying Jin, Wei Ye, Wenhan Yang, Yuan Yuan, and Robby T Tan. Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2634–2642, 2024. 2

[3] 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

[4] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realis-

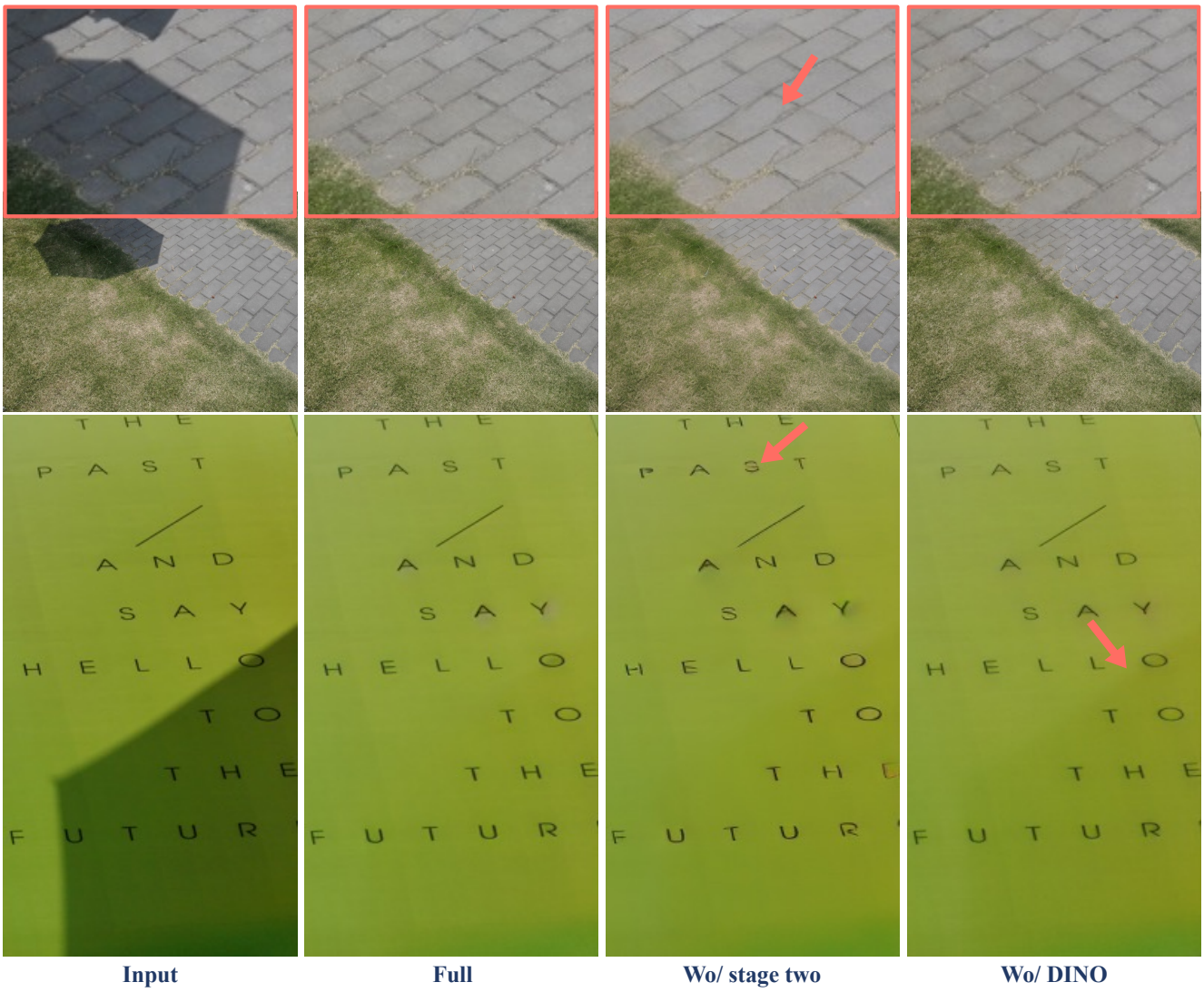

Figure E. **Qualitative results of ablation studies.**

tic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023. 2

[5] 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. 2, 3

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[7] 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

[8] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. 1

[9] Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. Homoformer: Homogenized transformer for image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25617–25626, 2024. 2

[10] Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. Omnisr: Shadow removal under direct and indirect lighting. *arXiv preprint arXiv:2410.01719*, 2024. 2, 3