

Appendix to "InstructRestore: Region-Customized Image Restoration with Human Instructions"

In the appendix, we provide the following materials:

- **Details of Prompt Tuning with Qwen:** We provide the implementation details of using Qwen for prompt tuning as discussed in Section 3 of the main paper, including the model setup, prompt design.
- **Word Cloud Illustration of the Dataset:** We present a word cloud illustrating the semantic frequency of object categories in our dataset (referring to Section 3 in the main paper).
- **Ablation Study:** We present ablation studies on mask decoder and feature modulation mechanism (referring to Section 5 in the main paper).
- **Verification for the Robustness of Instruction Variation:** We present the test result with other instructions not following the templates (referring to Section 5 in the main paper).
- **Quantitative Results for Controllable Bokeh Tuning:** We present the quantitative results, including the estimation of blur degree, to further validate our controllable bokeh tuning (referring to Section 5 in the main paper).
- **More Visual Results for Instruction Following:** We showcase additional visual results demonstrating the localized restoration on images from the Instruct100set based on human instructions (referring to Section 5 in the main paper).
- **More Visual Comparisons on Instruct100Set:** We present more visual comparisons with other methods on the Instruct100Set (referring to Section 5 in the main paper).
- **More Visual Comparisons on Bokeh Dataset:** We present more visual comparisons on the Bokeh Dataset, illustrating the model’s performance for the bokeh effect preservation (referring to Section 5 in the main paper).
- **Broader Impacts:** We analyze the potential social impact of our work.

A.1 Details of Prompt Tuning with Qwen

In Section 3 of the main paper, we mentioned that the initial masks and region descriptions obtained from Semantic-SAM [2] and Osprey [8] cannot meet the requirements for our training data. To address this, we employed a Large Language Model (LLM), specifically Qwen [1], to reformat the region captions into a noun phrase structure and extract the subject from the descriptions, which facilitates the merging of masks and region captions with identical semantics. As illustrated in Figure 1, we designed a specific prompt for this purpose. In the first round of parsing, we utilized the Qwen-7B [1] model to ensure time efficiency.

However, due to the inherent randomness of LLM outputs, issues such as spelling errors or repetitive text (*i.e.*, “parroting”) can arise. We found that the frequency of the parsed subjects could serve as a useful indicator of their correctness. Given that the number of subjects parsed in real-world scenarios typically does not exceed 200, subjects with frequencies ranked beyond the top 200 are likely to be parsing errors. For these instances, we revisited the original data and implemented the parsing process using the larger Qwen-72B [1] model. This iterative procedure was repeated twice, resulting in a total of three iterations to refine and finalize our annotations.

A.2 Semantic Frequency Word Cloud of the Dataset

In Section 3 of the main paper, we have introduced the construction process of triplets, each consisting of a high-quality image, a region mask, and a corresponding region caption describing the content within the mask. To further demonstrate the semantic diversity and broad applicability of our dataset, we generated a word cloud based on the frequency of subjects extracted from the annotation descriptions. The size of each word in the cloud corresponds to its relative frequency, with larger words representing more prevalent subjects. As shown in Fig. 2, our dataset encompasses a wide range of common semantic categories, including animals, plants, natural landscapes, pedestrians, and other subjects frequently encountered in restoration.

those with GT masks across all metrics in both target and remaining areas. This minimal performance gap demonstrates that our mask decoder generates high-quality masks, enabling accurate region-customized restoration. Furthermore, both predicted and GT masks show that the target region changes much with different fidelity scales, while the remaining areas stay stable, confirming precise local control. Another observation is that at a lower fidelity scale (0.5), where more details are generated in the target region, the difference between predicted and GT masks is slightly larger. This suggests that mask quality is more critical when generating finer details. Nevertheless, the performance gap remains small, further confirming the robustness of our mask decoder.

Table 1: Comparison between predicted masks and ground truth masks to evaluate mask decoder effectiveness.

Fidelity Scale	Target Area						Other Area	
	PSNR↑	SSIM↑	LPIPS↓	CLIPQA↑	MUSIQ↑	MANIQA↑	PSNR↑	SSIM↑
0.5 (predict_mask)	29.71	0.7522	0.1610	0.6801	67.86	0.6108	31.27	0.8949
1.1 (predict_mask)	30.73	0.8649	0.1253	0.6659	66.92	0.5934	31.56	0.9108
0.5 (gt_mask)	29.49	0.7297	0.1663	0.6810	67.76	0.6118	31.32	0.9017
1.1 (gt_mask)	30.74	0.8661	0.1247	0.6663	66.88	0.5930	31.56	0.9107

Feature Modulation Mechanism. The fidelity modulation mechanism in our method controls the amount of generated details by adjusting the coefficient for integrating conditional features from ControlNet into the SD backbone. Intuitively, a smaller fidelity coefficient allows the SD backbone to dominate, resulting in richer generated details, while a larger coefficient makes the output closer to the degraded image, increasing fidelity but reducing generated details.

To analyze the effect of the fidelity modulation mechanism, experiments were conducted within our multi-step inference framework (20 steps in total), with the fidelity scale set to 0.6 for the target region and 1 for the remaining area. We tested the impact of applying the fidelity modulation starting from different inference steps. In our default setting, we apply the fidelity modulation from the very first step (step 0). The results are summarized in Table 2. We observe that the later the modulation is applied (i.e., the larger the starting step t), the higher the fidelity metrics and the lower the no-reference quality metrics. This indicates that delaying the application of fidelity modulation will suppress the model’s ability to generate details, and narrow the adjustable range of effects when different fidelity scales are set according to user instructions. Therefore, to ensure sufficient controllability, we apply fidelity modulation from the very beginning of the inference process.

Table 2: Impact of different time steps implementing feature modulation mechanism.

t steps	PSNR↑	SSIM↑	LPIPS↓	CLIPQA↑	MUSIQ↑	MANIQA↑
15	30.52	0.8454	0.1333	0.6821	67.65	0.6055
10	30.41	0.8325	0.1371	0.6824	67.86	0.6083
5	30.28	0.8137	0.1431	0.6866	67.95	0.6116
0 (Ours)	30.09	0.7922	0.1511	0.6911	68.12	0.6152

A.4 Verification for the Robustness of Instruction Variation

In this section, we evaluate the robustness of our method to more natural or varied user expressions. We designed three instruction variants that were not seen during training. Specifically, we tested the following variants:

v1: “make the entire image sharper while make {region expression} clear”;

v2: “enhance the clarity of {region expression}”;

v3: “I would like {region expression} to be clear”.

For each variant, we set the foreground fidelity scale to either 0.5 or 1.1, the background to 1, and conducted experiments on the Instruct100 set. The test result is shown in Table 3. Notably, v1, which is most similar to the original template, achieves nearly identical results to the original instruction. For

v2 and v3, which differ more significantly in phrasing, our method still achieves comparable results, suggesting a certain degree of robustness to instruction variations. All the three instruction variants result in significant changes in the specified region and minimal changes elsewhere, with metrics similar to those under the standard instruction. This suggests that, although the phrasing differs, these instructions can still localize the intended region to some extent. However, we acknowledge that the generated masks are not identical across different instructions, which may cause some result variations. Incorporating diverse instruction formats during training construction would likely further enhance robustness.

Table 3: The quantitative test for instruction variations.

Fidelity Scale	Target Area						Other Area	
	PSNR↑	SSIM↑	LPIPS↓	CLIPQA↑	MUSIQ↑	MANIQA↑	PSNR↑	SSIM↑
0.5 (our)	29.71	0.7522	0.1610	0.6801	67.86	0.6108	31.27	0.8949
1.1 (our)	30.73	0.8649	0.1253	0.6659	66.92	0.5934	31.56	0.9108
0.5 (v1)	29.82	0.7636	0.1575	0.6772	67.89	0.6084	31.24	0.8929
1.1 (v1)	30.71	0.8641	0.1239	0.6702	67.19	0.5949	31.55	0.9124
0.5 (v2)	29.92	0.7813	0.1483	0.6580	66.73	0.5914	31.52	0.9050
1.1 (v2)	30.90	0.8814	0.1136	0.6229	64.24	0.5551	31.81	0.9238
0.5 (v3)	29.90	0.7751	0.1506	0.6622	67.09	0.5965	31.40	0.8994
1.1 (v3)	30.84	0.8781	0.1154	0.6361	65.16	0.5654	31.73	0.9210

A.5 Quantitative Result for Controllable Bokeh tuning

We provide visual results for controllable bokeh tuning in Figure 6 of Section 5.3 of the main paper. In this section, we provide the quantitative result for it. We fixed the foreground fidelity at 0.8 while varying background bokeh fidelity (0.4, 0.7, 1), where lower values indicate stronger blur. We measured foreground PSNR, background PSNR, and background blur strength using the Brenner metric (lower values means stronger blur).

As shown in Table 4, our tuning led to little variation in foreground PSNR (0.5 dB) but significant changes in background PSNR (2.52 dB). Moreover, as the bokeh fidelity scale decreased, the Brenner measure values consistently reduced, confirming that the blur intensity indeed increased as intended.

Table 4: Quantitative results for bokeh scale tuning

Scale	Foreground PSNR (dB)	Background PSNR (dB)	Background Brenner Measure
0.4	29.42	28.80	137.23
0.7	29.87	30.42	150.27
1.0	29.90	31.32	156.44

A.6 More Examples of Localized Enhancement with User-Instruction

In Figure 4 of Section 5.2 of the main paper, we used flowers as an example to demonstrate the effectiveness of our method in achieving region-specific restoration based on user instructions. In this section, we provide more examples of various semantic scenes to further validate the comprehensive semantic coverage of our dataset and the effectiveness of our method in following user instructions. In Fig. 3, as the fidelity scale decreases, the texture of the pagoda becomes increasingly detailed, with clearly recognizable windows emerging. Meanwhile, the restoration results of the surrounding trees remain unchanged. In Fig. 4, the text on the sign closely resembles the input when the fidelity scale is set to 1.1. As the fidelity scale decreases, the restored patterns begin to deviate from the input. In contrast, the restoration results of the surrounding plants remain unchanged. In Fig. 5, when the fidelity scale is set to 1.2, the cat’s fur appears clustered. As the fidelity scale decreases, the fur becomes more refined. However, the facial features of the cat, especially the eyes, increasingly deviate from the input, demonstrating that the fidelity indeed decreases as instructed. These examples highlight the capability of our dataset to support the processing of common semantic categories, as

well as the effectiveness of our method in understanding instructions, localizing target regions, and adaptively adjusting the restoration results.

Instruction: "make a large pagoda among trees clear with { fidelity scale s_1 } and keep other parts clear with 1"

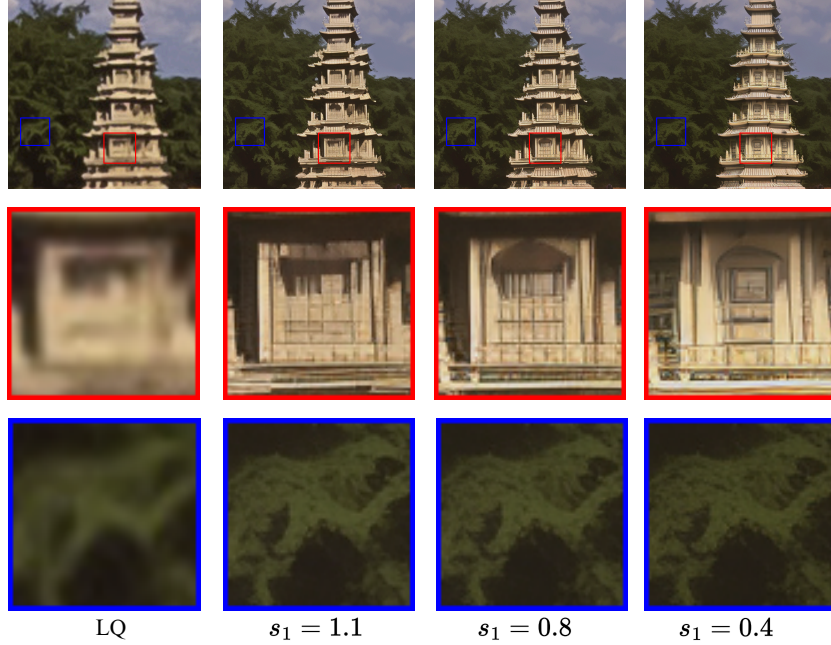


Figure 3: The varied localized enhancement for pagoda.

A.7 More Visual Comparisons on Instruct100Set

In this section, we provide additional examples comparing our method with existing restoration approaches on the Instruct100Set. As shown in Fig. 6, Real-ESRGAN [5] fails to restore fine details. Except for OSEDiff [6], other methods generate unnecessary artifacts in the windows of the building. However, OSEDiff restores the indoor lights as prominent white marks. In addition, the foliage restored by these methods exhibits a smeared appearance. In contrast, our method ensures the fidelity of the building while enhancing the details of plants, resulting in a more realistic overall appearance. In Fig. 7, our method restores more intricate details in the tree branches while maintaining background fidelity. In contrast, other methods, except for DiffBIR [3], lack realistic details in the restoration of branches. However, similar to StableSR [4] and SUPIR [7], DiffBIR exhibits incomplete denoising in the background regions. In Fig. 8, none of the other methods successfully restores the fleshy leaves of the succulent plant, which can be seen in the input. Additionally, in the results of StableSR and DiffBIR, the texture details of the adjacent clay pot are overly cluttered.

A.8 More Visual Comparisons on Bokeh Testset

In Section 5.3 of the main paper, we highlighted that our method outperforms existing approaches in preserving the natural background bokeh effect of scenes. In this section, we provide more visual comparisons to demonstrate this advantage. For a fair comparison, we set the fidelity scale for background bokeh blur to 1 in the instruction, which corresponds to the minimal blurring effect in our method. In Fig. 9, DiffBIR restores the blurred light spot in the lower-left corner into a flower, while PASD recovers partial vegetation in the upper-right corner. Other methods restore sharp details to varying degrees. In Fig. 10, although StableSR, PASD, and RealESRGAN preserve the background bokeh effect, the texture of the leaves is less clear compared to our results. In addition, the books beneath the leaves are restored to resemble wooden planks, indicating that their foreground fidelity is not as well maintained as ours. In Fig. 11, only our method keeps the background bokeh blur. Overall,

Instruction: "make a sign next to bushes clear with { fidelity scale s_1 } and keep other parts clear with 1"

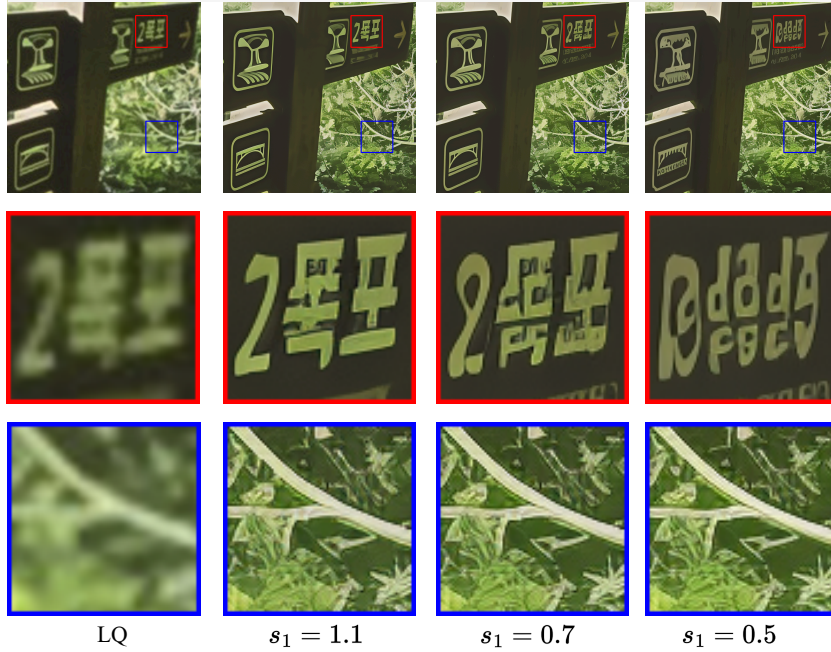


Figure 4: The varied localized enhancement for sign.

these visual comparisons further demonstrate the effectiveness of our method in preserving the bokeh effect.

A.9 Broader Impacts

This paper investigates customizing region-specific restoration effects based on user instructions to meet diverse user needs. Users can adjust the degree of background bokeh to align with aesthetic preferences or privacy protection requirements. In addition, they can optimize the trade-off between data fidelity and perceptual quality for different semantic regions to enhance visual effects. The key academic contribution lies in proposing the novel task of instruction-based regional restoration, addressing the limitation of existing global uniform processing methods in academia and advancing interactive restoration. In terms of broader social impact, this approach improves the flexibility of image restoration and supports the creative and media industries. Users can precisely tune image enhancements for specific regions, improving post-processing efficiency—particularly beneficial for social media content generation, where visual appeal is crucial. However, high-precision regional customization can increase computational costs, necessitating a balance between efficiency and performance, especially for mobile deployment.

Instruction: "make the cat with red scarf clear with { fidelity scale s_1 } and keep other parts clear with 1"

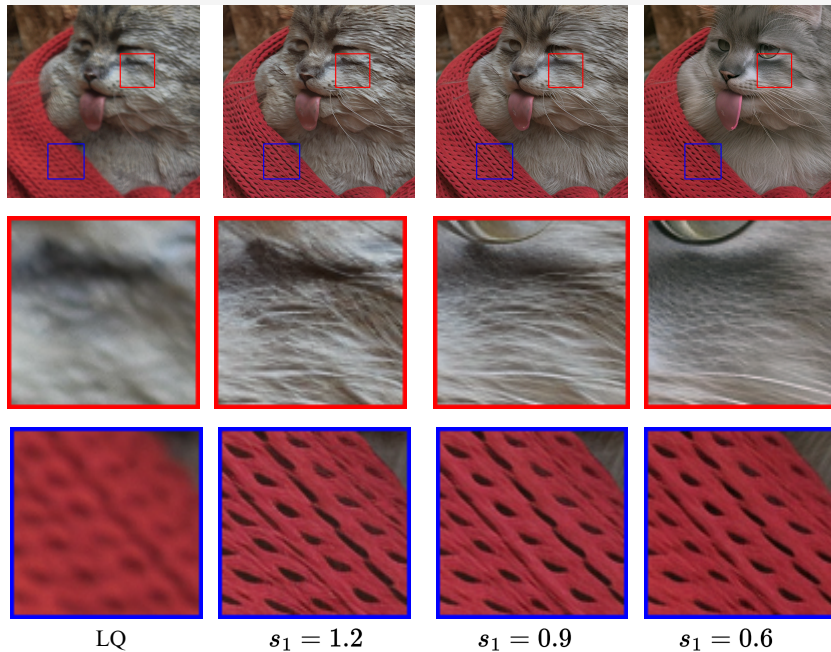


Figure 5: The varied localized enhancement for cat.

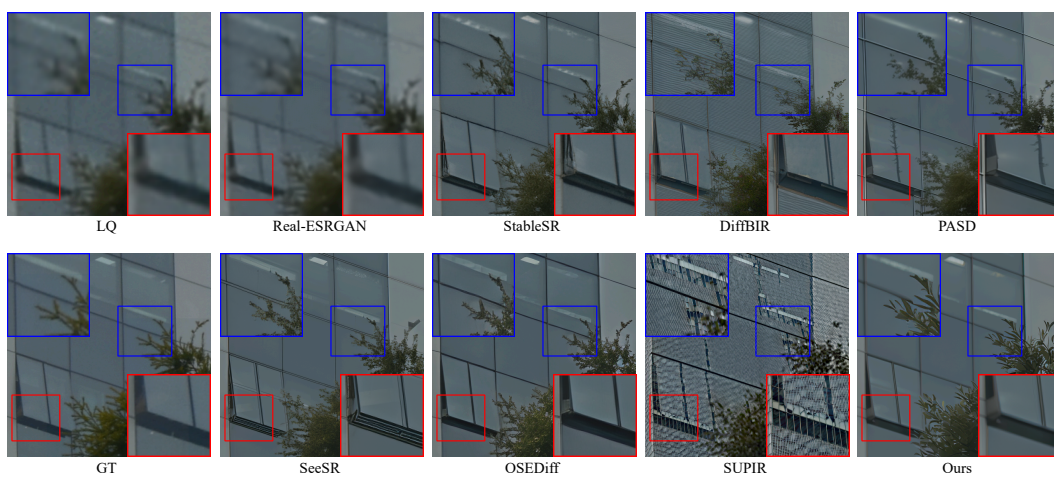


Figure 6: Example of visual comparison on Instruct100Set.

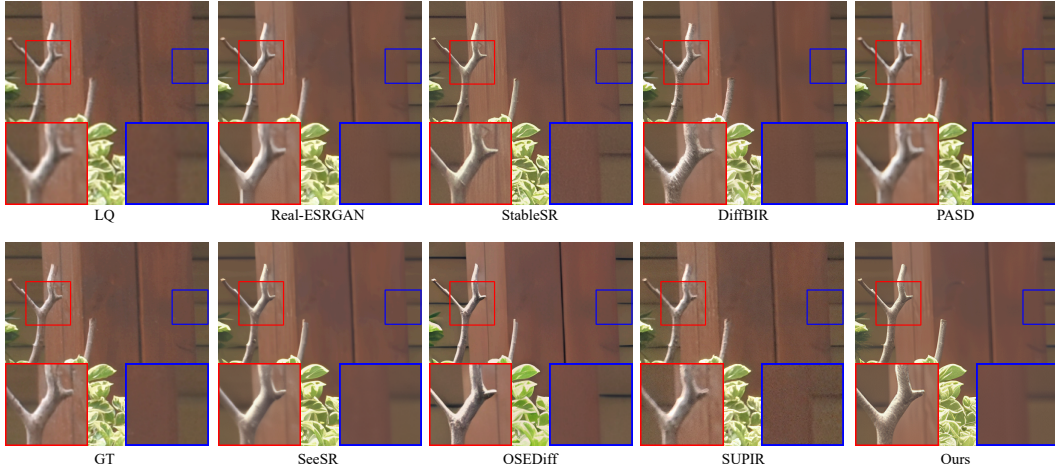


Figure 7: Example of visual comparison on Instruct100Set.

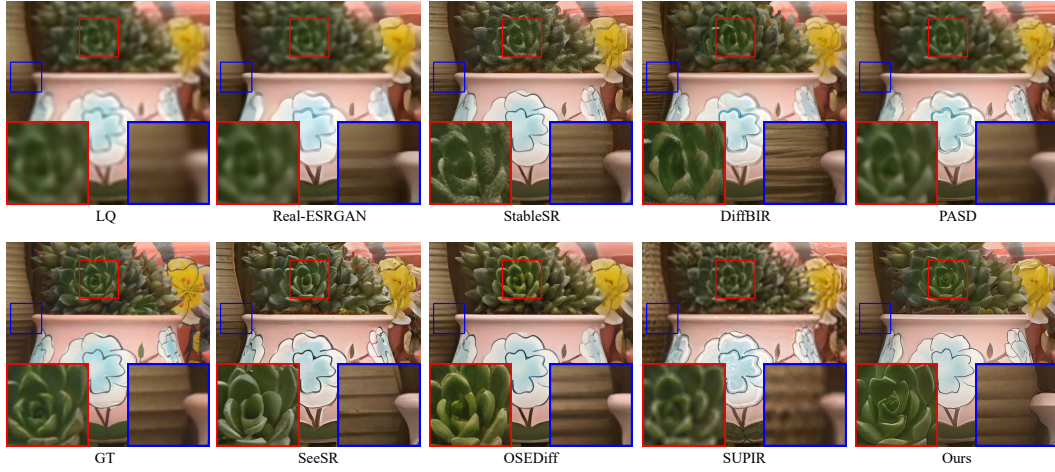


Figure 8: Example of visual comparison on Instruct100Set.



Figure 9: Example of visual comparison on Bokeh Testset.



Figure 10: Example of visual comparison on Bokeh Testset.

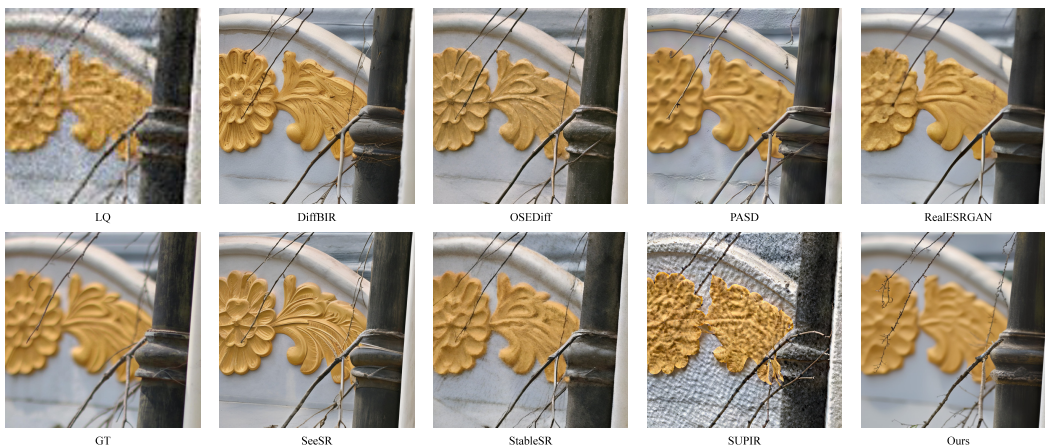


Figure 11: Example of visual comparison on Bokeh Testset.

References

- [1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [2] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *Eur. Conf. Comput. Vis.*, pages 467–484, 2024.
- [3] Xinqi Lin, Jingwen He, Ziyang Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *Eur. Conf. Comput. Vis.*, pages 430–448, 2024.
- [4] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *Int. J. Comput. Vis.*, pages 1–21, 2024.
- [5] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, pages 1905–1914, 2021.
- [6] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Adv. Neural Inform. Process. Syst.*, 37:92529–92553, 2025.
- [7] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25669–25680, 2024.
- [8] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 28202–28211, 2024.