# One-Step Diffusion Transformer for Controllable Real-World Image Super-Resolution

Supplementary Material

## 1. Overview

In this material, we provide the following contents:

- More implementation details of ODTSR, including hyperparameters, more explanation of the Prior Noise stream, model structure and loss visualization (referring to Sec. 5.1 in the main paper);
- More details of training and testing dataset (referring to Sec. 5.1 in the main paper);
- More ablation studies related to the Prior Noise and GAN training strategies (referring to Sec. 5.3 in the main paper);
- User Study details (referring to Sec. 5.4 in the main paper);
- More qualitative comparison results of Real-ISR and controllable Real-ISR (referring to Sec 5.2 in the main paper);
- Limitation and future works (referring to Sec. 6 in the main paper)

## 2. Implementation Details

In this section, we present various details of the model to help readers gain a deeper understanding of its design.

### 2.1. Hyperparameter Overview

Table 1. A summary of the main hyperparameters used in ODTSR.

| Hyperparameter | Value |
|---|---|
| Generator's Learning Rate | constant 5e-5 |
| Discriminator's Learning Rate | constant 5e-6 |
| Batch Size | 32 (gradient accumulation = 4) |
| Generator's Optimizer | RMSprop (alpha=0.9, momentum=0.0) |
| Discriminator's Optimizer | RMSprop (alpha=0.9, momentum=0.0) |
| Lora Rank | 128 |
| Lora Alpha | 128 |
| Training Iters | 10,000 |
| Training Resources | 8*NVIDIA H20 GPUs, within 48 hours |
| $t_p$ in Eq.(5) of main paper | 0.43 |
| Fidelity Weight $f$ | uniformly in the interval [0, 1] |
| $\lambda_1$ in Eq.(9) of main paper | 1.0 |
| $\lambda_{\text{min}}$ in Eq.(12) of main paper | 0.02 |
| $\lambda_{\text{max}}$ in Eq.(12) of main paper | 0.1 |
| variance $\sigma$ in Eq.(14) of main paper | 0.005 |
| $\lambda_2$ in Eq.(15) of main paper | 5.0 |
| Mixed Precision Training | True (*float8_e4m3fn* and *bfloat16*) |
| Gradient Checkpointing | True |

### 2.2. Explanation of $t_p$ for Prior Noise stream

In Eq.(5) of the main paper, the hyperparameter $t_p$ is a crucial value, as it determines the range of the noise level for LQ interpolation when the *Fidelity Weight* changes. In T2I pretrained models based on Rectified Flow [10], a **scheduler** is typically used during pre-training. This scheduler establishes a relationship between *timestep* (ranging from 0 to 999, 1000 discrete timesteps in total) and $t$ in Eq.(1) of the main paper. By sampling the timestep, the corresponding $t$ can be obtained. The scheduler used in Qwen-Image [15] is illustrated in Fig. 1.

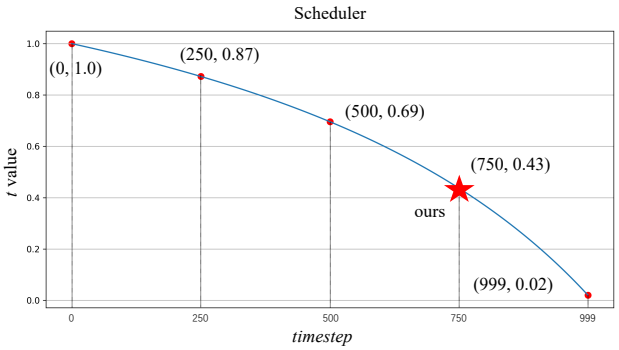


Figure 1. The **scheduler** used in Qwen-Image. The horizontal axis is the *timestep* ranging from 0 to 999, 1000 discrete timesteps in total, and the vertical axis represents the values of $t$. During pre-training, *timestep* is uniformly sampled to obtain $t$.

In our single-step training, we determine the value of $t_p$ by deciding the timestep, which is aligned with pre-training. This is why the value 0.43 looks a bit unusual, which actually corresponds to timestep 750.

We also ablate different timestep (i.e. 250 and 500) and use their corresponding $t_p$ values (i.e. 0.87 and 0.69) to train the model. See details in Sec. 4.1.

### 2.3. Model Structure Visualization

The base model we employ, Qwen-Image [15], adopts a double-stream design, consisting of a visual and a textual stream. We provide the detailed structure of a single transformer layer, as shown in Fig. 5.

### 2.4. Loss Curve Visualization

To help readers better understand the training dynamics, we visualize the loss curves, as shown in Fig. 2. Moreover, we summarize the common patterns observed in stable convergence as follows:

- $\mathcal{L}_{\textbf{MSE}}(I_{\textbf{pred}}, I_{\textbf{GT}})$: the loss decreases rapidly at the beginning and later fluctuates within a stable range.
- $\mathcal{L}_{\textbf{LPIPS}}(I_{\textbf{pred}}, I_{\textbf{GT}})$: shows a steady decrease (at least during the first 10k iterations).

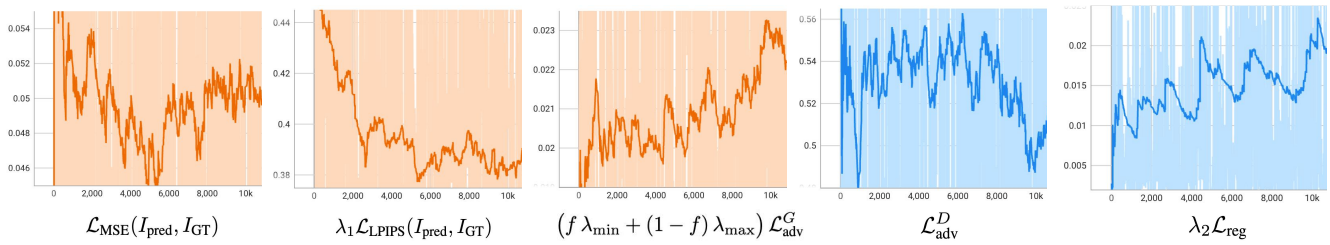


Figure 2. The visualization of the loss curves: the generator and discriminator losses are marked in different colors. Our training only needs 10,000 iterations to achieve good results.The meaning of the symbols corresponds to that in the main paper.

- $\mathcal{L}^G_{\textbf{adv}}$: shows a slight upward trend amid fluctuations.
- $\mathcal{L}^D_{\textbf{adv}}$: shows a slight downward trend amid fluctuations, and the value is less than $-\ln(\text{sigmoid}(0)) = -\ln(0.5) \approx 0.693$. This indicates that, on average, the discriminator has the ability to distinguish between real and fake samples.
- $\mathcal{L}_{\textbf{reg}}$: shows a slight upward trend amid fluctuations.

# 3. Dataset Details

## 3.1. Training Set

Like most previous methods, we use the LSDIR [9] and the first 10,000 images of FFHQ [8] as our training sets. The paired 512x512 data are constructed using the standard Real-ESRGAN [13] degradation pipeline. Specifically, the degradation hyperparameter configurations are exactly the same as those in PiSA-SR [12].

The difference is that our model requires textual descriptions during training. The discriminator always receives the description, while the generator uses the description with a probability of 0.75 and an empty description with a probability of 0.25 during each training iter. We design a set of six prompt templates to extract textual descriptions from GT images using Qwen2.5-VL-7B-Instruct [2]. These templates cover different levels of granularity, ranging from short tags to detailed captions and quality assessments:

- "Use a few word tags to summarize the main content of this image."
- "Describe the main content of this image in one sentence."
- "Describe the content of this image in a few sentences."
- "Describe possible detailed features in this image, such as text or faces; if no such targets exist, describe other details; if multiple targets exist, describe them separately."
- "Describe the quality of this image and evaluate it based on factors such as clarity, color, noise, lighting, and focal length/lens effects."
- "Describe the artistic style or form of this image (e.g., photography, illustration, painting, CG) and explain the overall tone and mood."

Since the base model we use supports both English and Chinese, we include six additional Chinese templates, each corresponding to one of the English templates. During each training iteration, we randomly select one prompt from the total of 12 templates.

## 3.2. Testing Set

**Controllable Real-ISR: RealSR** We utilize the RealSR [3] dataset to evaluate the performance on general controllable image restoration. The difference from a standard Real-ISR task is that we provide the model with a text description. The textual descriptions for RealSR are extracted from the ground truth images using Qwen2.5-VL-7B-Instruct [2], and all methods use the same descriptions. The prompt used to extract the description is: "Describe the content of this image in a few sentences."

**Controllable Real-ISR: RealCE** We utilize the RealCE-val [11] dataset to evaluate the performance on scene text image restoration, especially with text annotations. We center-crop all images to 512×512 and manually remove pairs where LQ and GT are not aligned. Finally, we get 260 LQ-GT pairs. For text annotations, we employ *PP-OCRv5* of Paddle-OCR [4] to extract text from GT and use it directly as the corresponding prompt.

# 4. More Ablation Studies

## 4.1. Selection of timestep in the Prior Noise stream

In this section, we investigate the impact of different *timestep* selections of Prior Noise. We train another two models with different timestep (i.e. 250 and 500) for the same training steps and compare their performance. As shown in Tab. 2, as timestep decreases from 750 to 250. which means $t_p$ and the level of the Prior Noise increases, the reconstruction fidelity is severely compromised. This is because our Fidelity-Aware Adversarial Training strategy samples GAN weight based on $t_p$, and higher $t_p$ value requires more training for fidelity (i.e. $f$ approaches 1 in Eq.(6) and Eq.(12) of the main paper). We thus choose timestep 750 empirically.

Ours (Relativistic GAN) No GAN Vanilla GAN

Figure 3. Visualization of GAN ablation. Trained with no GAN, model generates grid-like and blur artifacts. Trained with Vanilla GAN rather than Relativistic GAN, the restored image lacks detail and has been smoothed out.

Table 2. **Ablation of timestep of Prior Noise.** We train two models with different timestep (i.e. 250 and 500) of Prior Noise for the same training steps and compare the performance on *RealSR* dataset. *Fidelity Weight* is set to 1.0. The best is highlighted in **bold**.

| Timestep | $t_p$ | LPIPS↓ | MANIQA↑ | FID↓ | CLIP-T↑ |
|---|---|---|---|---|---|
| 250 | 0.87 | 0.3318 | 0.6610 | 144.79 | 29.76 |
| 500 | 0.69 | 0.3057 | 0.6387 | 118.08 | 31.00 |
| 750 (Ours) | 0.43 | **0.2398** | **0.6622** | **101.49** | **32.37** |

### 4.2. GAN training strategies

In this section, we investigate two key questions in GAN training: (1) why GAN is necessary and (2) why relativistic GAN [7] is chosen. We train two models following these two questions and evaluate their performance quantitatively and qualitatively. As shown in Tab. 3 and Fig. 3, trained without GAN ("No GAN") is comparable in terms of fidelity, but fails in image quality, which is also demonstrated in visualization, showing grid-like artifacts and blur; trained with vanilla GAN ("Vanilla GAN") solves grid-like artifacts but produces over-smooth restoration results and also falls behind in the metric evaluation. Overall, relativistic GAN demonstrates the best performance. This also demonstrates that PSNR cannot accurately reflect image quality.

Table 3. **Ablation of GAN Strategies.**

| GAN | LPIPS↓ | MANIQA↑ | FID↓ | PSNR↑ |
|---|---|---|---|---|
| No GAN | 0.2600 | 0.6180 | 119.28 | **26.07** |
| Vanilla GAN | 0.2931 | 0.6451 | 142.93 | 25.60 |
| Ours (Relativistic GAN) | **0.2398** | **0.6622** | **101.49** | 25.07 |

## 5. User Study Details

We invite 20 volunteers to evaluate ODTSR and other three latest SOTA methods (TSD-SR [5], PiSA-SR [12] and DiT4SR [6]) in generic Real-ISR setting. 100 LQ images are randomly chosen from four datasets (RealSR [3], DRealSR [14], Div2k-val [1], RealCE-val [11]). *Fidelity Weight* is set to 1.0 and input prompt is empty in ODTSR. In the user study, volunteers are presented with six images: LQ, GT and restoration results from four methods randomly ordered. Volunteers are asked to choose the best result according to (1) fidelity with LQ and GT (2) overall image quality. We build a simple webpage for the user study, with the interface shown in Fig. 4.

## 6. More Qualitative Results

### 6.1. Real-ISR

We present more Real-ISR results in Fig. 6 and Fig. 7, comparing our ODTSR with three state-of-the-art methods PiSA-SR [12] (one-step, based on Diffusion U-Net), TSD-SR [5] (one-step, based on Diffusion Transformer) and DiT4SR [6] (multi-step, based on Diffusion Transformer), which is identical to the settings in the user study. In generic Real-ISR setting, ODTSR achieves remarkable performance in fidelity and detail quality. For example, Row 2 of Fig. 6 shows restored sculpture image with refined and realistic facial details and Row 5 of Fig. 7 shows restored correct texture of the blanket with no prompt guidance.

### 6.2. Controllable Real-ISR

**Text scene** We present more results on RealCE-val, a scene text image super-resolution dataset. We show results of ODTSR both with and without prompt. As shown in Fig. 8, ODTSR achieves higher restoration quality than other methods without prompt. When text annotation is incorporated as a prompt, some distorted characters were corrected while also demonstrating prompt controllability.

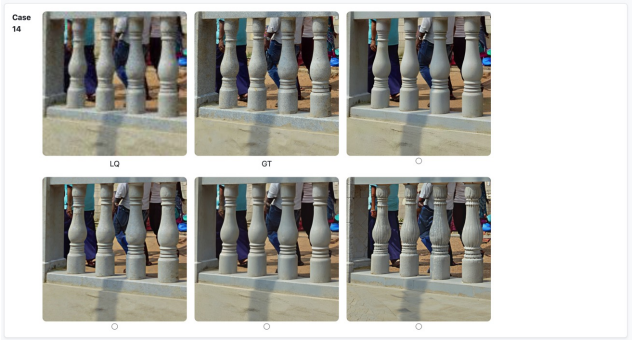


Figure 4. User Study Interface.

**General scene** We present more results on RealSR with prompt extracted from GT for easier visual comparison. Fig. 9 shows that ODTSR, compared with multi-step methods SUPIR [16] and DiT4SR [6], restores realistic and high-quality details under prompt control. For example, in Row 4 of Fig. 9, ODTSR restores correct details of seats and wall.

**Adjustable Fidelity Weight** In Fig. 10, we present results with adjustable *Fidelity Weight* (denoted as $f$) in ODTSR to demonstrate the scope and effectiveness of controllability. In general, as $f$ decreases from 1 to 0, detail generation and prompt adherence gradually strengthen.

## 7. Limitation and Future Works

Although ODTSR achieves high fidelity and prompt controllability, its large number of parameters leads to the high computational cost. We plan to apply various model acceleration techniques to mitigate computational costs. Additionally, the *Fidelity Weight* in ODTSR is currently adjusted for the whole image, leaving room for more fine-grained control.

## References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 3

[2] 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. 2

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3086–3095, 2019. 2, 3

[4] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. 2

[5] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23174–23184, 2025. 3

[6] Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. *arXiv preprint arXiv:2503.23580*, 2025. 3, 4

[7] A Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 3

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[9] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1775–1787, 2023. 2

[10] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1

[11] Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for chinese-english scene text image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19452–19461, 2023. 2, 3

[12] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2333–2343, 2025. 2, 3

[13] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 2

[14] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, pages 101–117. Springer, 2020. 3

[15] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25669–25680, 2024. 4

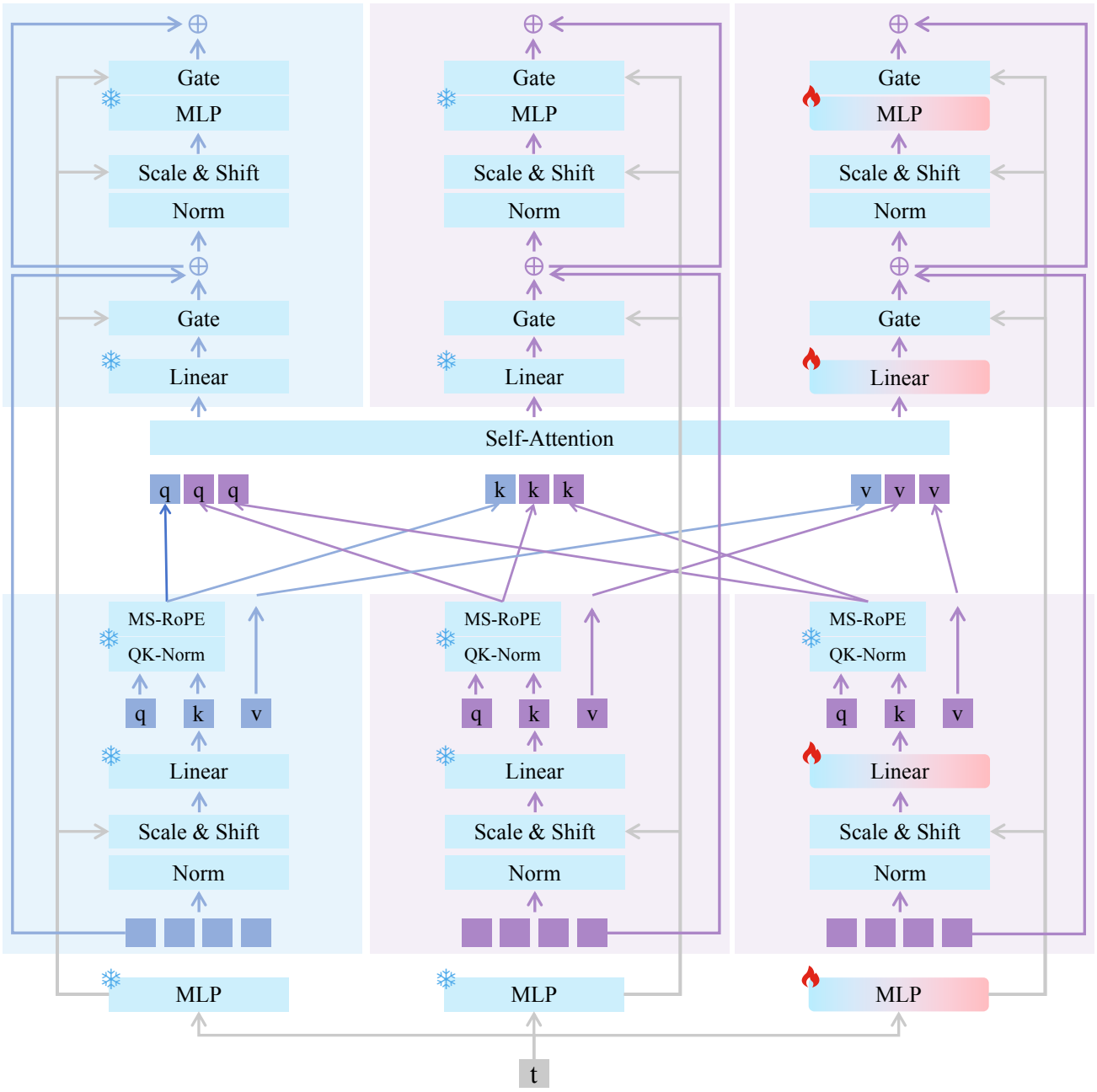


Figure 5. The model contains 60 transformer layers in total. The details of a single transformer layer are shown here: the left branch corresponds to the Text stream, the middle to the Prior Noise stream, and the right to the Control Noise stream. Among them, only the linear layers in the Control Noise stream are trained with LoRA, while all other parameters remain frozen.


Figure 6. More qualitative results on **Real-ISR**. *Fidelity Weight* is set to 1.0 and input prompt is empty in ODTSR. ODTSR achieves remarkable performance in fidelity and detail quality.

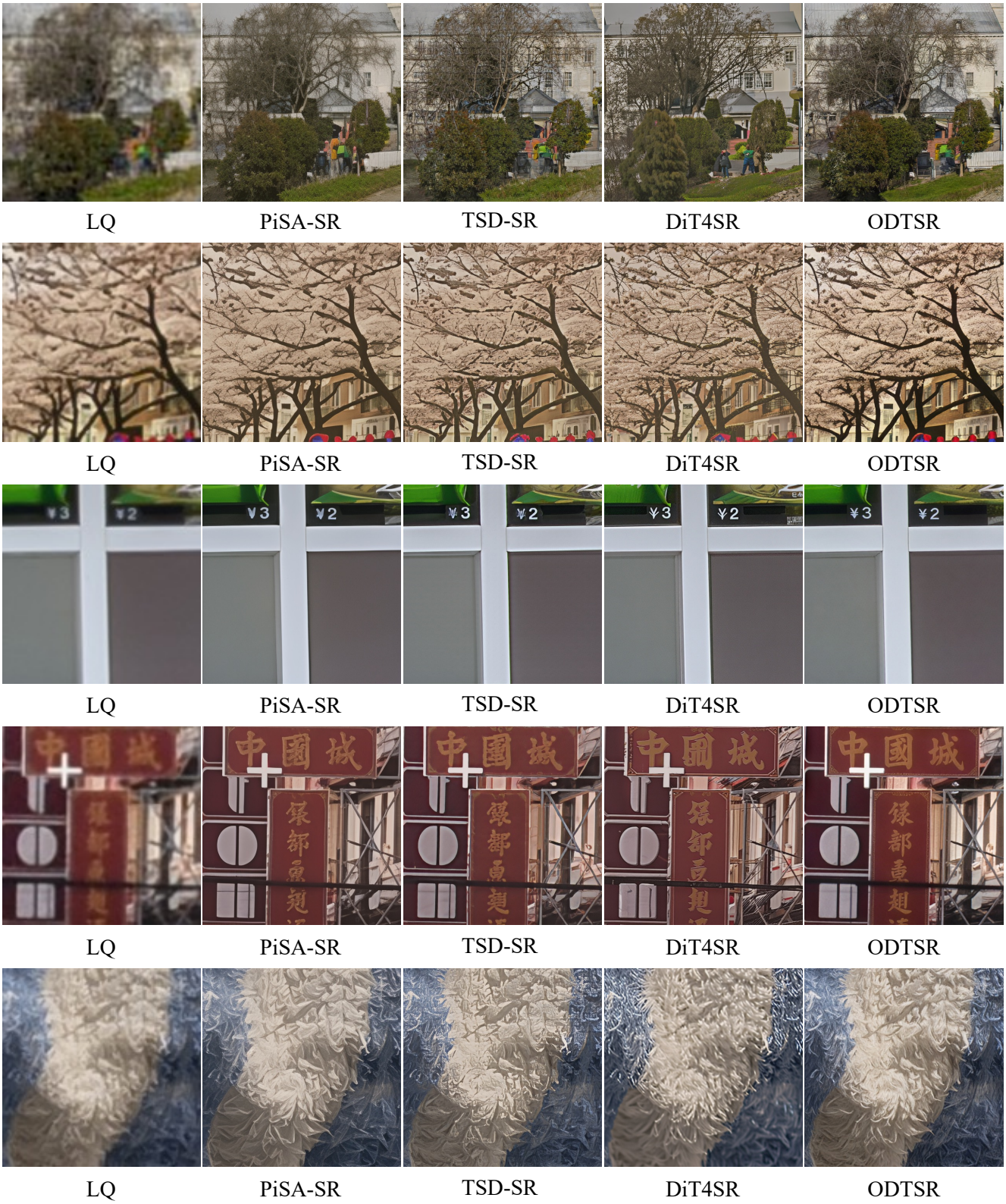


Figure 7. More qualitative results of **Real-ISR**. *Fidelity Weight* is set to 1.0 and input prompt is empty in ODTSR. ODTSR achieves remarkable performance in fidelity and detail quality.

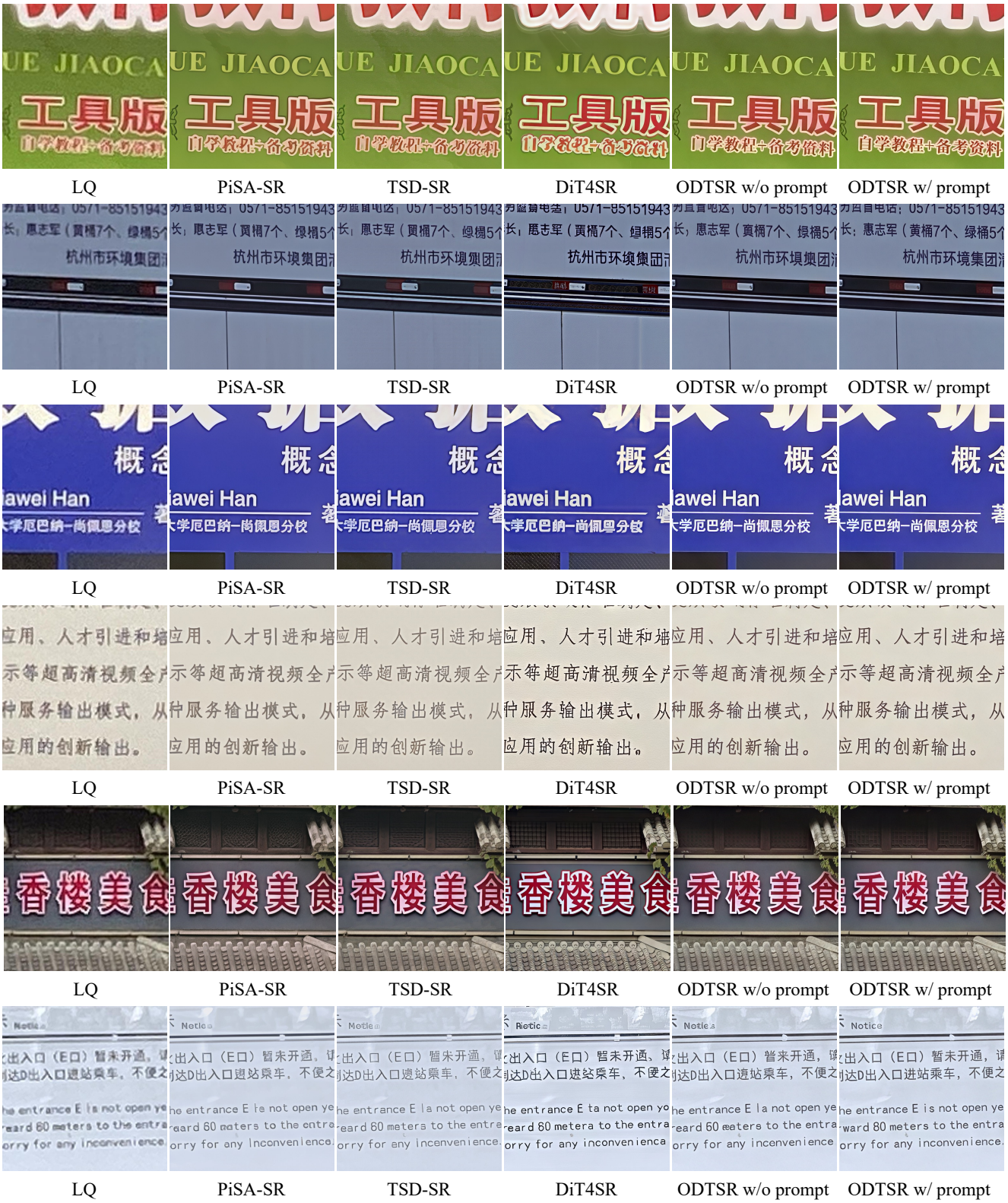


Figure 8. More qualitative results of **Controllable Real-ISR** on *RealCE-val* dataset. *Fidelity Weight* is set to 1.0 and we show results of ODTSR both with and without prompt (text annotation). ODTSR achieves higher restoration quality than other methods.

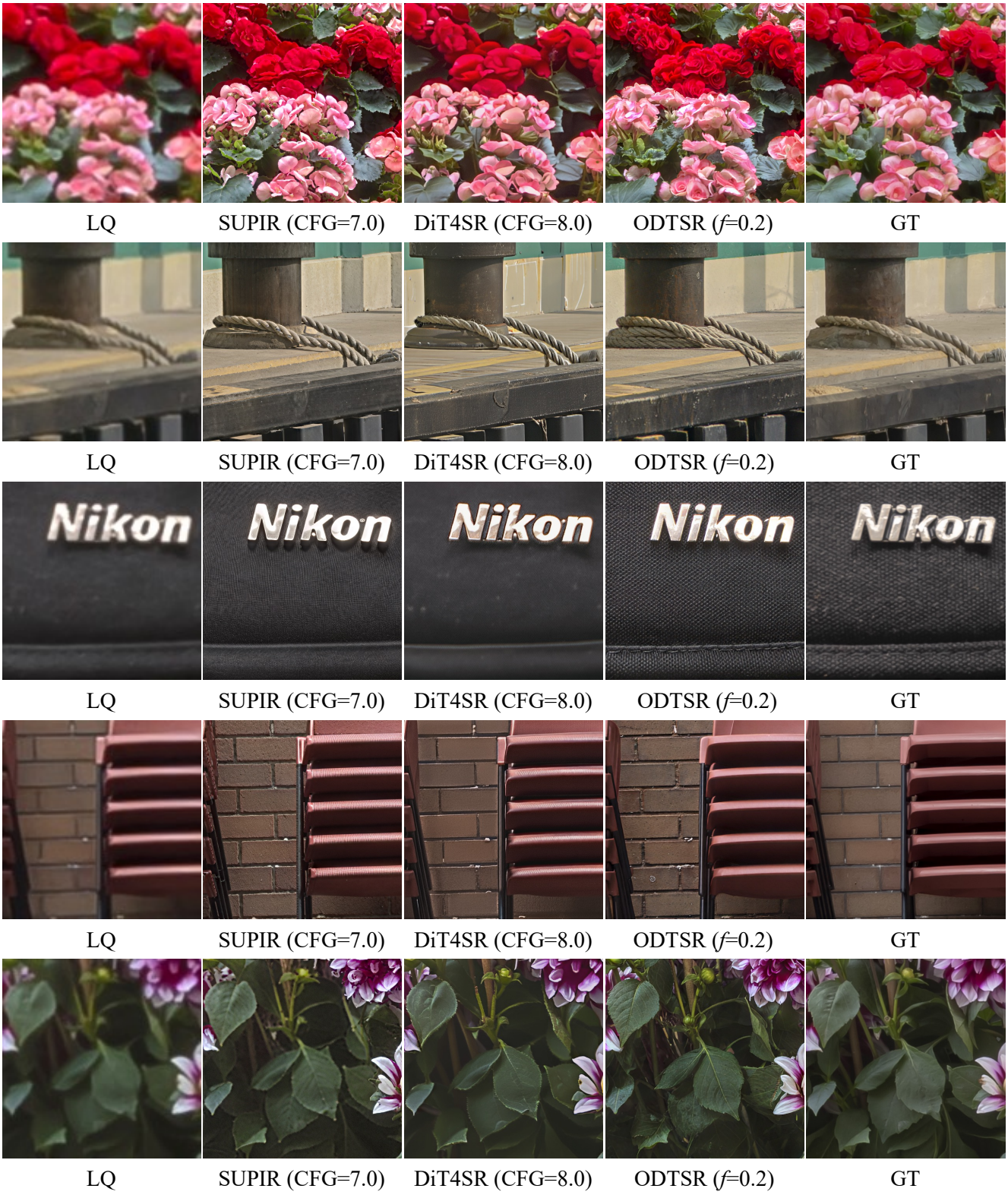


Figure 9. More qualitative results of **Controllable Real-ISR** on *RealSR* dataset. *Fidelity Weight* is denoted as $f$ and the input prompt is extracted from GT same across three methods. ODTSR restores more realistic and higher-quality details under prompt control.


Figure 10. More qualitative results of **Controllable Real-ISR** with prompt and adjustable *Fidelity Weight* (denoted as *f*) on *Div2k-val* dataset. As *f* decreases from 1 to 0, detail generation and prompt adherence gradually strengthen, demonstrating controllability of ODTSR. The prompt corresponding to the enhanced details is marked in red.