DNF-SR: Dual-Input and Negative-Aware Feature Fine-Tuning for Real-World Image Super-Resolution

Supplementary Material

In this Supplementary Material, we provide additional details, including the comparison with GAN-based methods in Section A, discussion on perception and restoration in Section B, more visual comparisons in Section C, and the algorithm in Section D. We conduct these additional comparisons and analyses to validate the effectiveness of DNF-SR.

A. Comparison with GAN-based Methods

We compare DNF-SR with three GAN-based Real-ISR methods: BSRGAN [22], RealESRGAN [13], and LDL [8]. Quantitative evaluations are conducted on the DIV2K [1], RealSR [4], DrealSR [15] and RealLQ [2] datasets, with results summarized in Tab. 1. The experimental results demonstrate that DNF-SR, leveraging a dual-input strategy and a novel post-training optimization method NF^{2}T, achieves significantly superior no-reference metrics compared to GAN-based methods.

Additionally, Fig 2 presents a visual comparison between DNF-SR and other GAN-based methods. The results show that DNF-SR reconstructs more photorealistic and natural outcomes. When compared to GAN-based methods, DNF-SR demonstrates distinct advantages in visual fidelity. Specifically, it achieves higher precision in restoring structured elements (e.g., text and architectural details) while rendering complex materials such as fabrics and natural textures with enhanced realism. This enables DNF-SR to more accurately reproduce the fine-grained detail hierarchy and authentic visual texture characteristic of high-resolution images, outperforming the GAN-based methods in both structural integrity and perceptual quality.

B. Discussion on Perception and Restoration

In experiments, we observe that existing multi-modal large language models (MLLMs) can effectively perceive the content in images. Even for challenging low-resolution (LR) images, they can infer reasonable content for blurred regions based on the overall image information. Meanwhile, current DiT-based generative models can adhere well to captions for image generation. However, in SR tasks, restoring strongly semantic structures such as text and logos is extremely challenging. As shown in the first row of Fig. 1, when no additional caption is used for the SR model, it is difficult to restore text with normal semantics. As shown in Fig 1, when captions generated by MLLMs are used as conditions, DiT-based SR models can effectively alleviate this

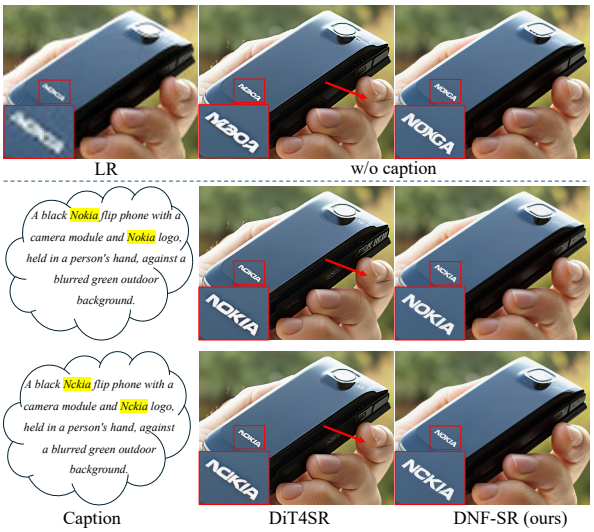

Figure 1. Visual comparisons between DNF-SR and DiT4SR when using different captions including no caption, a reasonable caption with "Nokia", and an unreasonable caption where "Nokia" is replaced with "Nckia". DiT4SR exhibits structural issues at the position indicated by the red arrow.

issue. Nevertheless, when we manually replace "Nokia" with "Nckia", DiT4SR exhibits poor prompt-following performance. This is because when applying text-to-image models to SR task, the model is caused to focus more on LR images, which impairs the inherent prompt-following capability of the original text-to-image model. In contrast, our method DNF-SR narrows the gap between LR and the original input of generative models through a dual-path input design. It also initializes with an image editing model to better perceive the information of LR used as conditions. This enables DNF-SR to better preserve the inherent generation capability and prompt-following ability of the original generative model when applied to SR tasks, thereby allowing it to better adhere to captions and restore more realistic and reasonable images.

However, current SR methods still obtain captions by leveraging MLLMs to perceive image content before applying them to restoration tasks. Only using text caption may sometimes fail to fully convey image information, and the separate execution of perception and restoration steps also introduces redundancy. Thus, it is highly meaningful to develop an integrated perception-restoration model that can accurately perceive semantic information in LR images

Table 1. A comprehensive evaluation against state-of-the-art GAN-based methods across synthetic and real-world datasets. The top-performing results under each metric are marked in **red**.

Datasets	Methods	PSNR ↑	LPIPS [3] ↓	CLIPIQA [12] ↑	MUSIQ [7] ↑	MAINIQA [20] ↑	QALIGN [16] ↑	VQ-R1 [18] ↑
DIV2k	BSRGAN	**24.583**	0.3351	0.5246	61.193	0.5040	3.1703	3.3063
	RealESRGAN	24.293	**0.3112**	0.5276	61.049	0.5484	3.2764	3.2623
	LDL	23.828	0.3256	0.5179	60.040	0.5328	3.1798	3.1018
	DNF-SR	23.631	0.3234	**0.7723**	**71.546**	**0.6703**	**4.1563**	**4.3630**
DrealSR	BSRGAN	**28.702**	0.2858	0.5092	57.159	0.4844	2.9572	3.0559
	RealESRGAN	28.618	0.2818	0.4517	54.275	0.4902	2.8638	2.7683
	LDL	28.196	**0.2790**	0.4473	53.948	0.4894	2.8576	2.6129
	DNF-SR	28.141	0.3531	**0.7559**	**68.732**	**0.6515**	**3.7997**	**3.9152**
RealSR	BSRGAN	**26.379**	**0.2656**	0.5114	63.283	0.5419	3.1829	3.4907
	RealESRGAN	25.687	0.2710	0.4489	60.364	0.5504	3.1081	3.1342
	LDL	25.280	0.2750	0.4556	60.930	0.5495	3.0898	2.9897
	DNF-SR	24.970	0.3239	**0.7257**	**70.040**	**0.6930**	**4.0718**	**4.2646**
RealLQ250	BSRGAN	-	-	0.5940	66.289	0.5963	3.4794	3.8124
	RealESRGAN	-	-	0.6253	66.990	0.6148	3.6471	3.8088
	LDL	-	-	0.6183	67.027	0.6147	3.6357	3.6940
	DNF-SR	-	-	**0.7997**	**73.700**	**0.7029**	**4.4752**	**4.6090**

and perform restoration within a single framework.

C. More Visual Comparisons

In Fig. 3 and 4, we provide more visual comparisons with other diffusion-based Real-ISR methods. As shown in Fig. 3, DNF-SR can better adhere to the content of the caption and restore more accurate text information. And in close-up scenarios, DNF-SR can better restore the texture and details of the image. Meanwhile, as shown in Fig. 4, DNF-SR can restore more realistic images under severe degradation. These examples all demonstrate the performance and robustness of DNF-SR for Real-ISR.

D. Algorithm Details

The training of DNF-SR consists of two stages: supervised fine-tuning (SFT) and post-training. In the SFT stage, we use multiple losses to perform supervised fine-tuning on the paired (x_L, x_H, c) dataset. In the post-training stage, we adopt a Negative-aware Feature Fine-Tuning method for reinforcement learning. Specifically, we sample K noises to generate K restored images, then compute rewards using multiple reward functions, which are normalized and aggregated into a single r. Subsequently, we define positive and negative optimization directions to improve model performance. Here, $K = 8$. Details are in Algorithm 1.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 1

[2] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. In *NeurIPS*, 2025. 1

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 2

[4] 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 *ICCV*, 2019. 1

[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 *CVPR*, 2025. 4, 5

[6] Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *ICCV*, 2025. 4, 5

[7] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 2

[8] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *CVPR*, 2022. 1, 3

[9] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024. 4, 5

[10] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ArXiv preprint*, 2025.

[11] 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 *CVPR*, 2025. 4, 5

[12] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2

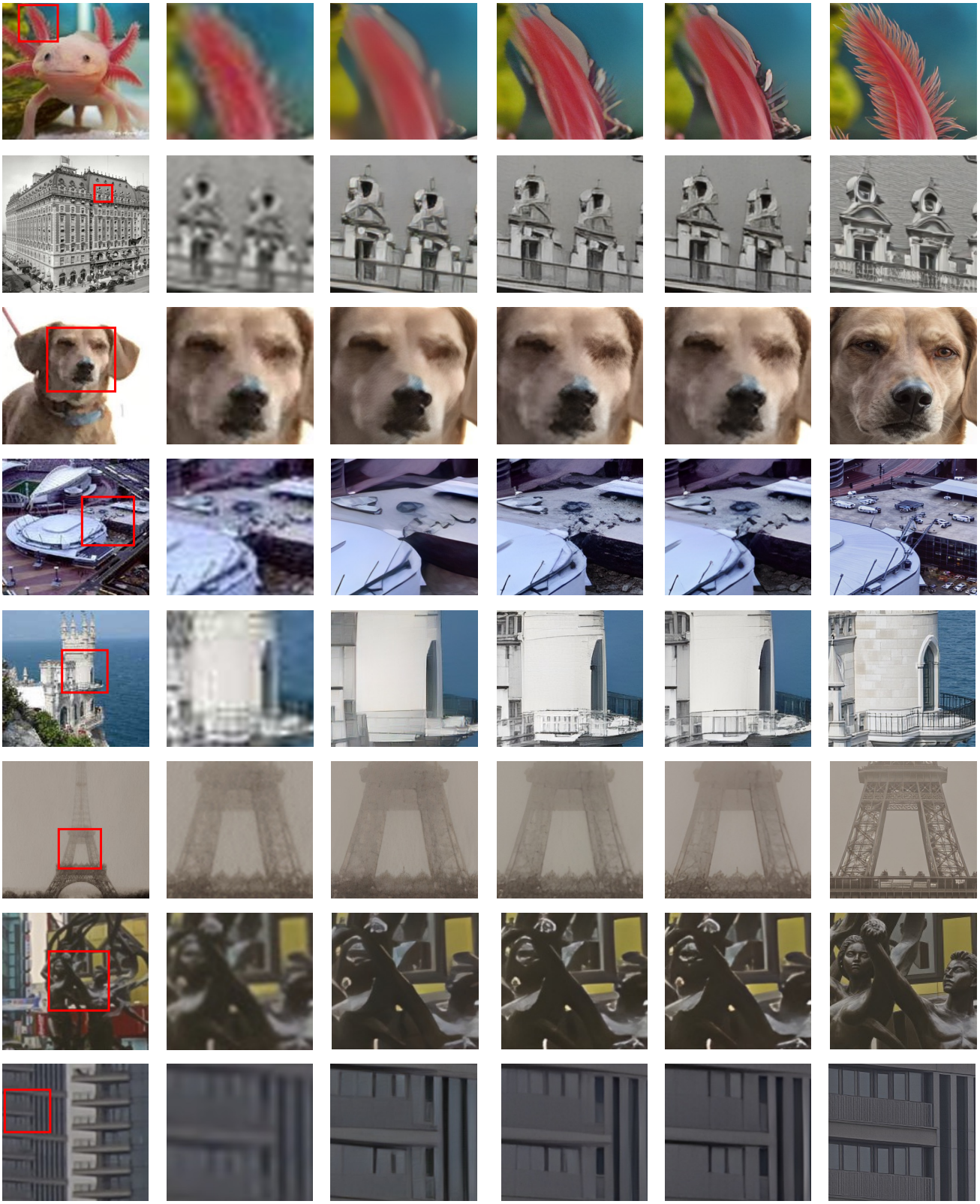

Figure 2. Vision comparisons between DNF-SR and GAN-based Real-ISR methods [8, 13, 22]. Zoom in for a better view.

Figure 3. Vision comparisons between DNF-SR and different diffusion-based Real-ISR methods [5, 6, 9–11, 14, 17, 19, 21].

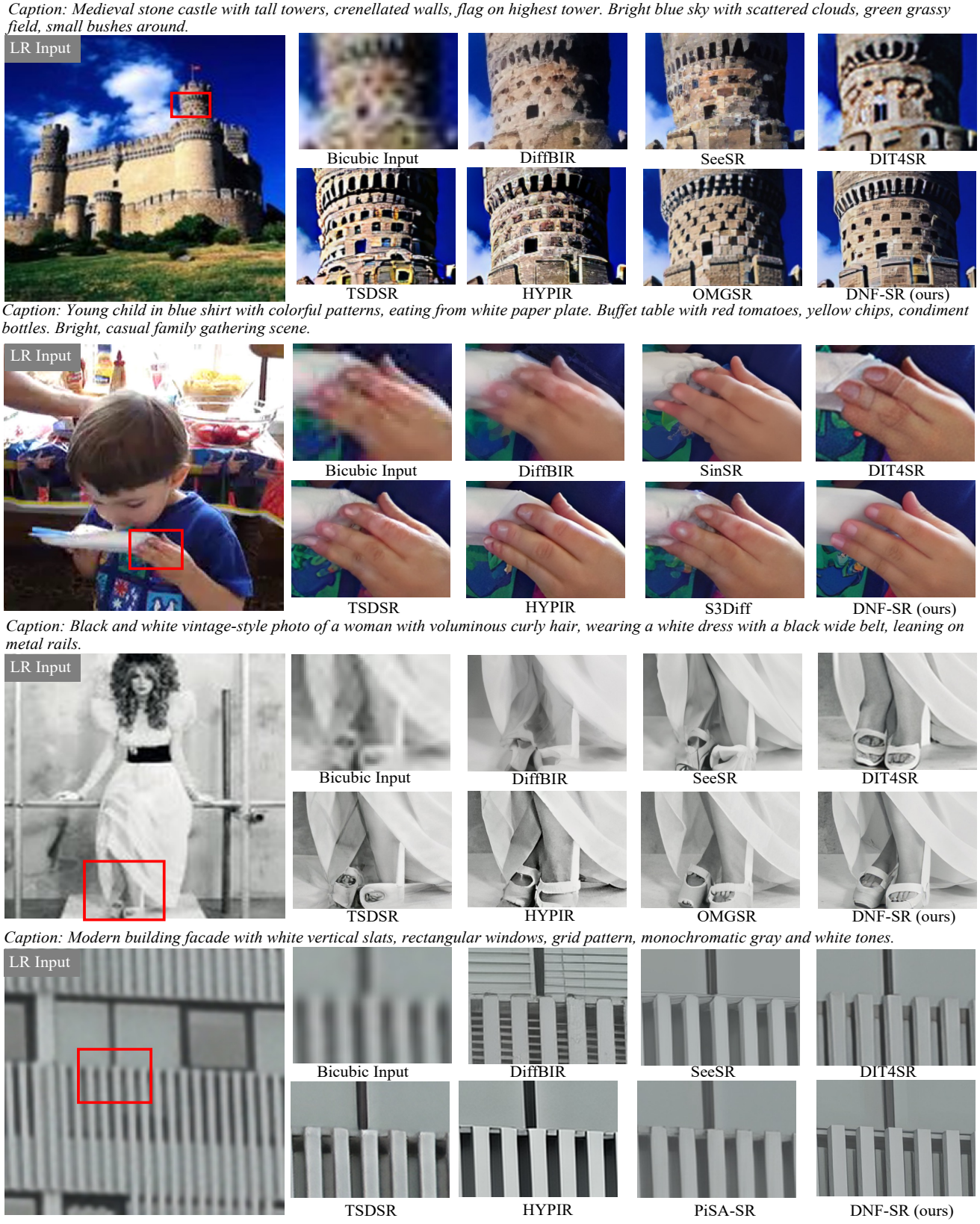

Figure 4. Vision comparisons between DNF-SR and different diffusion-based Real-ISR methods [5, 6, 9–11, 14, 17, 19, 21].

Algorithm 1: Training Procedure of Negativate-aware Feature Fine-tuning in DNF-SR

Input: Training datasets $\{x_L, x_H, c\}$, fine-tuning one-step Diffusion-based SR Model including VAE encoder E_{ref} and velocity prediction network v_{ref}, pre-trained VAE decoder D_φ, number of samples K per training step, N raw reward functions $r^{raw}(\cdot) \in \mathbb{R}$, one fixed mid-timestep t_{mid}.

Output: Post-trained one-step velocity prediction network v_θ for SR.

Initialize data collection policy for velocity prediction $v_{old} \leftarrow v_{ref}$.
Initialize training policy for velocity prediction $v_\theta \leftarrow v_{ref}$.
Initialize data buffer $\mathcal{D} \leftarrow \emptyset$
while *train* **do**
  /* Rollout Step, Data Collection */
  for *each sampled data*$(x_L, x_H, c) \sim \mathcal{D}$ **do**
    Sample K standard Gaussian noises $\epsilon^{1:K}$ and collect K restored images $\hat{x}_H^{1:K}$ using v_{old}.
    Compute rewards $\{r_{1:N}^{raw}\}^{1:K}$ using N raw reward functions, respectively.
    Standardize raw rewards in group: $r_i^{std} := (r_i^{raw} - mean(\{r_i^{raw}\}^{1:K}))/std(\{r_i^{raw}\}^{1:K})$.
    Normalize rewards using the standard Gaussian cumulative distribution function: $r_i = \Phi(X < r^{std})$.
    Average the N rewards: $r = avg(r_{1:N})$
    $D \leftarrow \{c, x_L, \hat{x}_H^{1:K}, r^{1:K}\}$
  end
  /* Gradient Step, Policy Optimization */
  for *each mini batch* $\{c, x_L, \hat{x}_H, r\}$ **do**
    Encode the LR image: $z_L = E_{ref}(x_L)$.
    Forward diffusion process: $z_t = t_{mid}\epsilon + (1 - t_{mid})z_L$.
    /* Calculate Positive Optimization Direction */
    Implicit positive velocity: $v_\theta^+ := (1-\beta)v^{old}(z_t, c, t_{mid}) + \beta v_\theta(z_t, c, t_{mid})$.
    Implicit positive image: $\hat{x}_\theta^+ := D_\varphi(z_L - t_{mid}v_\theta^+)$.
    Positive optimization direction: $\mathcal{L}_\theta^+ = r\mathcal{L}_{rec}(\hat{x}_\theta^+, \hat{x}_H)$.
    /* Calculate Positive Optimization Direction */
    Implicit negative velocity: $v_\theta^- := (1+\beta)v^{old}(z_t, c, t_{mid}) - \beta v_\theta(z_t, c, t_{mid})$.
    Implicit negative image: $\hat{x}_\theta^- := D_\varphi(z_L - t_{mid}v_\theta^-)$.
    Negative optimization direction: $\mathcal{L}_\theta^- = (1-r)\mathcal{L}_{rec}(\hat{x}_\theta^-, \hat{x}_H)$.
    /* Update Model Parameters */
    $\theta \leftarrow \theta - \lambda\nabla_\theta\left[\mathcal{L}_\theta^+ + \mathcal{L}_\theta^-\right]$
  end
  /* Online Update */
  Update data collection policy $v_{old} \leftarrow v_\theta$, and clear buffer $\mathcal{D} \leftarrow \emptyset$.
end

[13] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 1, 3

[14] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 4, 5

[15] 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 *ECCV*, 2020. 1

[16] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024. 2

[17] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 4, 5

[18] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *ArXiv preprint*, 2025. 2

[19] Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. *ArXiv preprint*, 2025. 4, 5

[20] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022. 2

[21] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *ArXiv preprint*, 2024. 4, 5

[22] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 1, 3