FinPercep-RM: A Fine-grained Reward Model and Co-evolutionary Curriculum for RL-based Real-world Super-Resolution

Supplementary Material

1. Detailed Formulations of Alignment Algorithms

This section details the mathematical formulations of the three primary Reinforcement Learning from Human Feedback strategies discussed in the main paper: Reward Feedback Learning [8], Direct Preference Optimization (DPO) for diffusion [5], and Group Relative Policy Optimization [3]. We demonstrate how our FinPercep-RM can be integrated into each framework.

1.1. Reward Feedback Learning (ReFL)

ReFL [8] is a direct optimization method designed to fine-tune diffusion models using feedback from a reward model without the need for a complex PPO pipeline. It leverages the observation that image quality becomes identifiable at the later stages of denoising.

Formulation. Let $\epsilon_\theta(x_t, t, c)$ be the noise prediction network with parameters θ, where x_t is the latent at timestep t, and c is the text prompt. ReFL randomly samples a timestep $t \in [T_{start}, T_{end}]$ (typically late steps, e.g., $[30, 40]$ for a 40-step scheduler). It then predicts the clean latent x_0 directly from x_t:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t, c)}{\sqrt{\alpha_t}}, \tag{1}$$

where α_t denotes the noise schedule parameters. The predicted image $I_{pred} = \mathcal{D}(\hat{x}_0)$ is then scored by our FinPercep-RM $R(\cdot)$. The ReFL objective is:

$$\begin{aligned}\mathcal{L}_{\text{ReFL}}(\theta) = \lambda\mathbb{E}_{c\sim\mathcal{Y}}[\phi(R(I_{pred}))] + \\ \mathbb{E}_{(c,x)\sim\mathcal{D}_{pre}}[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2],\end{aligned} \tag{2}$$

where $\phi(\cdot)$ is a weighting function (e.g., ReLU) and the second term is a pre-training regularization.

1.2. Diffusion Direct Preference Optimization (Diffusion-DPO)

Diffusion-DPO [5] reformulates the RLHF objective to directly optimize the policy on preference pairs (x^w, x^l) using the Evidence Lower Bound (ELBO) as a proxy for log-likelihood.

Derivation. The objective encourages the model to lower the denoising error for the preferred image x^w relative to a frozen reference model θ_{ref}, while allowing the error for the rejected image x^l to increase:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x^w,x^l,c),t,\epsilon} \\ \left[\log\sigma\left(-\frac{\beta}{2}\left(\delta_{\text{MSE}}(x^w) - \delta_{\text{MSE}}(x^l)\right)\right)\right],\end{aligned} \tag{3}$$

where $\delta_{\text{MSE}}(x) = \|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 - \|\epsilon - \epsilon_{\text{ref}}(x_t, t, c)\|_2^2$. In our context, the pairs (x^w, x^l) are labelled using the score $S_{fgc-global}$ from our FinPercep-RM.

1.3. Group Relative Policy Optimization (Flow-GRPO)

Flow-GRPO [3] introduces Group Relative Policy Optimization to Flow Matching models. It computes advantage without a value network by using group statistics.

ODE-to-SDE Conversion. To enable exploration, the deterministic ODE $dx_t = v_t dt$ is converted into an SDE:

$$dx_t = \left(v_t(x_t) - \frac{\sigma_t^2}{2}\nabla\log p_t(x_t)\right)dt + \sigma_t dw. \tag{4}$$

Objective. For a prompt c, a group of G outputs $\{x_0^i\}_{i=1}^G$ is generated. The advantage A_i is normalized within the group based on rewards R_i from FinPercep-RM:

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\}) + \epsilon}. \tag{5}$$

The GRPO objective is:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Big(\min\big(\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i\big) \\ - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big)\Bigg].\end{aligned} \tag{6}$$

2. Extended Ablation and Comparison Studies

2.1. Ablation on Loss Components

To rigorously evaluate the contribution of each loss term in our training objective, we performed a step-wise ablation study. We start with the baseline ranking loss and progressively incorporate the map loss and alignment loss. The results, presented in Table 1, confirm that each component provides a cumulative benefit to the final perceptual quality.

Model	$\mathcal{L}_{rank}$	$\mathcal{L}_{map}$	$\mathcal{L}_{align}$	MUSIQ ↑	MANIQA ↑
A	✔	✗	✗	71.052	0.628
B	✔	✔	✗	72.804	0.651
Full	✔	✔	✔	**73.456**	**0.658**

Table 1. Ablation study on the training objectives of FinPercep-RM. We progressively add the ranking loss, fine-grained map loss, and anchor alignment loss. Model A represents a standard preference model baseline; Model B incorporates fine-grained spatial supervision; the Full model adds score alignment for stability across training stages.

2.2. Performance Comparison with Other Reward Models

To demonstrate the superiority of our proposed **FinPercep-RM** in the context of Real-ISR, we conducted a comparative experiment against several state-of-the-art reward models widely used in the Text-to-Image (T2I) domain. Specifically, we compared our method with **ImageReward** [8], **HPSv2** [7], **PickScore** [2], **CLIP Score** [1], and **Aesthetic Score** [4].

Reward Model	**MUSIQ ↑**	**MANIQA ↑**
Aesthetic Score [4]	71.12	0.589
CLIP Score [1]	70.56	0.591
PickScore [2]	71.65	0.605
HPSv2 [7]	72.03	0.618
ImageReward [8]	71.56	0.628
FinPercep-RM (Ours)	**73.45**	**0.658**

Table 2. Quantitative comparison with state-of-the-art T2I reward models on the RealLQ250 benchmark. Our FinPercep-RM significantly outperforms general-purpose reward models in the ISR task, indicating the necessity of fine-grained defect awareness.

3. Additional Qualitative Results

We provide additional visual comparisons on real-world datasets to demonstrate the robustness of our method. As shown in Fig. 1, our method generates more plausible details while effectively suppressing artifacts, yielding the best visual results.

4. FinPercep-RM Architecture Details

The FinPercep-RM adopts an Encoder-Decoder architecture designed to simultaneously predict a global quality score and a local degradation map.

4.1. Encoder

We utilize the CLIP-IQA model [6] as the backbone encoder. It processes the input image into a sequence of patch embeddings. We extract multi-scale feature maps $\{f_i\}_{i=1}^{4}$ from intermediate transformer layers to capture hierarchical visual information. Specifically, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder outputs features from layers $\{6, 12, 18, 24\}$ of the backbone, ensuring that both low-level texture information and high-level semantic information are preserved for the subsequent decoding task.

4.2. Decoder

The decoder consists of a Feature Pyramid Network (FPN)-like structure.

- **Lateral Connections:** Each selected feature map f_i from the encoder is projected to a common channel dimension via 1×1 convolutions.
- **Upsampling Path:** Lower-resolution features are upsampled and added to higher-resolution features.
- **Prediction Head:** The final fused feature map is passed through two 3×3 convolution layers followed by a Sigmoid activation to produce the Perceptual Degradation Map (M_{fg-pdm}).

4.3. Global Score Head

The global score is computed by spatially weighting the encoder's final semantic feature f_{final} with the predicted degradation map:

$$f_{weighted} = f_{final} \odot \text{Downsample}(M_{fg-pdm}). \quad (7)$$

This weighted feature vector is then fed into a 3-layer MLP to regress the scalar quality score $S_{fgc-global}$. This mechanism explicitly forces the global score to be aware of local defects identified by the decoder.

5. Implementation and Training Details

5.1. FinPercep-RM Training

- **Loss Weights:** In Eq. (8) of the main paper, we set $\lambda_{map} = 1$, $\lambda_{rank} = 0.8$, and $\lambda_{align} = 0.6$. The high weight on the map loss enforces accurate localization.
- **Optimizer:** AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$, weight decay $1e-4$.
- **Learning Rate:** $1e-5$ for the encoder backbone (fine-tuning) and $1e-4$ for the newly added decoder and MLP head. Cosine annealing scheduler.
- **Batch Size:** 64. Trained for 50 epochs on FGR-30k.

5.2. Co-evolutionary Curriculum Learning (CCL)

The CCL training process is structured into $N = 3$ stages to balance stability and performance. The detailed configuration is as follows:

- **Training Schedule:** Each stage consists of **10,000 steps**, totaling 30,000 steps for the complete training process.
- **Learning Rate:** The initial learning rate is set to 1×10^{-5} and decays according to a **cosine annealing** schedule.

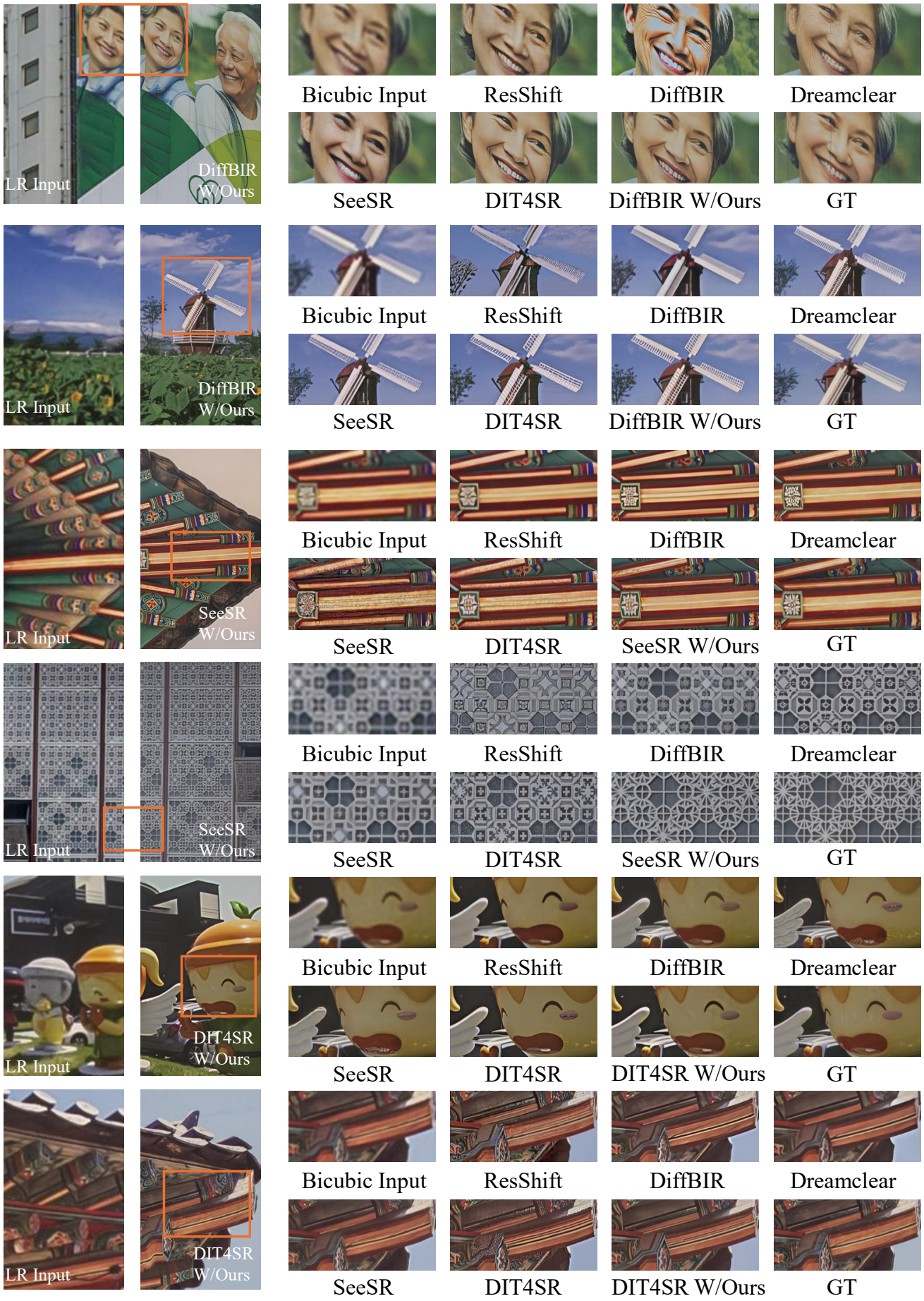

Figure 1. Qualitative comparisons with state-of-the-art Real-ISR methods on on RealSR based on RLHF method of REFL [8].

- **Stage 0 (Warm-up):** The Generator is trained using the frozen CLIP-IQA (RM_0) reward. This stage ensures initial stability by optimizing for global perceptual quality without the complexity of fine-grained local feedback.
- **Stage 1 (Partial Expansion):** The FinPercep-RM introduces a lightweight version of the decoder with fewer parameters. This allows the Generator to adapt to coarse-grained local defect penalties without being overwhelmed by high-frequency signals.
- **Stage 2 (Full Expansion):** The full parameters of the FinPercep-RM decoder and adapters are activated. The Generator fine-tunes against this fully capable reward model to meticulously refine local textures and eliminate subtle artifacts.

6. User Study Configuration

To rigorously evaluate perceptual quality, we conducted a user study with 20 human raters.

Interface. Raters were presented with a web interface showing the Low-Resolution input (for reference) and two super-resolved images side-by-side (Ours vs. Baseline, randomized order).

Criteria. Raters were asked to select the better image based on two distinct criteria:

1. **Fidelity:** Which image better preserves the identity and structure of the original content?
2. **Realism:** Which image looks more like a natural photograph, free from AI-generated artifacts or over-smoothing?

Dataset. We randomly sampled 50 images from each of the four test datasets (Total = 200 comparisons per rater). The results reported in the main paper (Table 2) are the averaged preference rates.

7. Broader Impacts

Our work on fine-grained perceptual reward models for Real-ISR has potential positive impacts but also carries certain risks that must be acknowledged.

Positive Impacts.

- **Restoration of Historical Archives:** Our method can significantly improve the quality of restoring old photographs and films, helping to preserve cultural heritage with higher fidelity.
- **Medical Imaging:** In medical scenarios where high-resolution details are crucial (e.g., MRI or CT scans), our fine-grained perception approach could potentially aid in clearer visualization, though rigorous validation is needed before clinical use.
- **Content Creation:** The ability to generate high-fidelity textures benefits the creative industry, including game development and VFX, by reducing the manual effort required for texture upscaling.

Potential Risks and Mitigation.

- **Hallucination and Misinformation:** Like all generative SR models, our method may hallucinate plausible but non-existent details. In contexts like surveillance or forensics, this could lead to misidentification. We strongly advise against using generative SR for legal evidence without expert verification.
- **Bias Amplification:** If the training data (e.g., FGR-30k or the pre-trained T2I priors) contains biases regarding race, gender, or age, the model might hallucinate biased features during restoration (e.g., changing facial characteristics). We plan to audit our FGR-30k dataset for demographic balance in future work.
- **Deepfakes:** The enhanced realism could be misused to create convincing deepfakes. We support the development of robust detection methods and digital watermarking to trace AI-generated content.

References

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[2] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 2

[3] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 1

[4] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 2

[5] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 1

[6] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 2

[7] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2

[8] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023. 1, 2, 3