

Beyond the Boundary: RL-Driven Solution Space Exploration for Blind Face Restoration

Bin Wu^{1,2}, Wei Wang^{1,2}[✉], Yahui Liu³, Chi Zhang^{1,2}, Yao Zhao^{1,2}

¹Institute of Information Science, Beijing Jiaotong University, Beijing, China

²Visual Intelligence +X International Cooperation Joint Laboratory of MOE, Beijing, China

³Xiaohongshu Inc., China

{bin.wu, wei.wang}@bjtu.edu.cn

Abstract. Blind Face Restoration (BFR) encounters inherent challenges in exploring its large solution space, leading to missing details and identity ambiguity in deterministic outputs. While diffusion models offer powerful generative priors, their potential is often constrained by deterministic sampling processes that lack exploration in the vast solution space. To tackle this, we propose **Likelihood-Regularized Policy Optimization (LRPO)**, the first online reinforcement learning (RL) framework that uniquely leverages the inherent stochasticity of diffusion models for the BFR task. Instead of relying on a single deterministic mapping, LRPO systematically explores multiple diverse restoration candidates per input. By evaluating these trajectories, our framework steers the denoising policy towards optimal solutions, significantly increasing the likelihood of high-quality outputs. To effectively facilitate this exploration mechanism, LRPO incorporates three key innovations: 1) a composite reward function balancing human preference and fidelity, 2) ground-truth-guided likelihood regularization, and 3) a noise-level advantage assignment. Extensive experiments demonstrate that LRPO significantly improves face restoration quality over baselines and achieves *state-of-the-art* performance. Source code is available at the GitHub repository [BFR_RL](https://github.com/BFR-RL).

Keywords: Blind Face Restoration · Reinforcement Learning · Diffusion Models

1 Introduction

Blind Face Restoration (BFR), which aims to reconstruct high-quality (HQ) faces from low-quality (LQ) inputs with unknown degradations, has made rapid progress in recent years. Modern BFR methods typically exploit various types of priors to establish direct mappings from LQ to HQ. These priors can be categorized into several types: (1) *geometric priors* (e.g., facial landmarks [7, 18], parsing maps [5], and component heatmaps [45]) that provide structural guidance; (2) *generative priors* [4, 35, 43] derived from pre-trained models such as StyleGAN [15, 16],

[✉] Corresponding author.

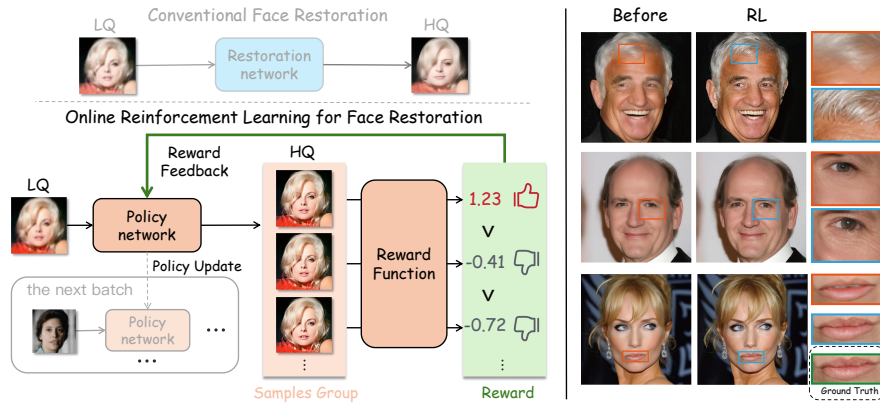


Fig. 1: (Left) Our proposed online RL-based face restoration framework: an LQ face is input to the policy network π_θ , which generates a group of HQ face candidates. The reward function evaluates each candidate and converts the scores into within-group relative advantages that guide policy optimization for the next iteration. The comparisons (Right) demonstrate the quality improvement achieved through RL optimization over the base model.

which enable realistic detail reconstruction; (3) *discrete codebook priors* [9, 51] improve restoration fidelity; and (4) *diffusion priors* [6, 20, 36, 40, 47] are built upon the denoising and noise-adding diffusion model. Diffusion models offer distinct advantages including robust generative capability, stable optimization, and superior control over output diversity, making them particularly effective for producing high-quality, visually pleasing face restorations.

However, despite the advantages of diffusion priors, BFR remains fundamentally challenging. The task is an ill-posed inverse problem where a single LQ input can correspond to multiple plausible HQ solutions, making it difficult to determine the optimal restoration. Current methods are constrained by their deterministic nature. They learn a fixed one-to-one mapping that produces a single output without considering alternative solutions. This lack of exploration within the vast solution space prevents these methods from discovering potentially superior restorations, leading to suboptimal results [50].

To address these exploration limitations, we formulate the BFR denoising process as a reinforcement learning (RL) optimization problem. RL has demonstrated remarkable success across various domains, particularly in language models [28, 44] and vision models [8, 21, 34, 46], by enabling diverse exploration strategies rather than deterministic outputs. Building on recent advances that successfully integrate RL with diffusion models [21, 41]—where the denoising process is formulated as a Markov decision process—we leverage RL’s exploration capabilities alongside diffusion models’ inherent randomness to systematically search BFR’s solution space for optimal restorations that enhance both fidelity and perceptual quality.

To this end, we construct a BFR-native online RL framework: Likelihood-Regularized Policy Optimization (LRPO). As the first to apply online RL to this task, our framework utilizes a policy network to generate multiple diverse HQ candidate faces from each LQ input, effectively exploring the solution space rather than following a single deterministic trajectory. As shown in Figure 1 (left), this multi-candidate sampling strategy allows systematic exploration of potential solutions. A reward function evaluates each candidate, and relative advantages computed among candidates from the same input guide the policy network optimization. However, naively transferring standard RL algorithms to BFR inevitably triggers reward hacking and disrupts the strict perception-fidelity balance. To overcome these fundamental limitations, we construct LRPO with three core innovations specifically tailored for diffusion dynamics: (1) We design a composite reward function that evaluates restoration quality by incorporating human preferences, perceptual quality, and fidelity metrics. (2) To prevent reward hacking [1, 29]—where the policy exploits reward signals while deviating from authentic facial distributions—we implement ground-truth (GT)-guided likelihood regularization during the training phase that anchors the policy to the true data manifold. (3) We propose a noise-level advantage assignment mechanism that weights the advantages according to the importance of each denoising step, ensuring more effective policy updates. It is worth noting that our LRPO is designed as a highly data-efficient post-training paradigm rather than training from scratch. Specifically, we apply this RL-based exploration directly to a fully-trained face restoration model (*e.g.*, DiffBIR) using only a small subset of the training data, ensuring minimal training cost.

In summary, our main contributions are as follows:

- We are the first to formulate the BFR learning process as an RL-driven multi-candidate exploration. Rather than simply applying existing algorithms, we construct the LRPO framework to overcome the limitations of single deterministic trajectory generation by exploring multiple restoration candidates through RL training.
- We introduce three critical components for our proposed LRPO framework: a multi-faceted reward function that captures diverse restoration quality aspects, GT-guided likelihood regularization that maintains authentic facial distributions while preventing reward exploitation, and adaptive advantage weighting that optimizes learning across different denoising stages.
- Our LRPO framework delivers substantial improvements in face restoration quality and establishes new *state-of-the-art* performance on standard evaluation metrics.

2 Related Work

Diffusion-based Blind Face Restoration. Blind Face Restoration (BFR) aims to recover high-quality face images from these degraded inputs while preserving identity consistency and perceptual quality. Recently, diffusion models [12, 30] have gained popularity due to their generative diversity and training stability.

DR2 [36] uses a diffusion model for degradation removal, followed by refinement through an enhancement module. DiffFace [47] constructs a posterior distribution to map LQ images to HQ counterparts, utilizing the error-shrinkage property of pre-trained diffusion models for robust restoration. To accelerate training, LDM [26] recommends training diffusion in the latent space. DiffBIR [20], built on LDM, employs ControlNet to guide restoration using low-quality faces as control signals. Some variations of diffusion models have been used for face restoration. InterLCM [19] leverages the latent consistency model (LCM) to improve semantic consistency and restore images efficiently. FlowIE [52] uses conditional rectified flow for faster inference with comparable restoration quality. Despite their generative power, existing diffusion-based BFR methods typically rely on deterministic sampling trajectories, which restrict their ability to explore the vast solution space. Consequently, they often settle for suboptimal restorations with compromised identity or missing details. In contrast, our LRPO framework breaks this deterministic constraint. By harnessing the inherent stochasticity of diffusion models through online reinforcement learning, we formulate the restoration as a multi-candidate exploration process, effectively navigating the solution space to discover the optimal, high-fidelity face.

RL in Vision Generation. Reinforcement learning has recently achieved remarkable success in improving large language model reasoning, particularly through policy gradient approaches such as PPO [27] and GRPO [28]. In the text-to-image (T2I) generation field, many methods have explored incorporating policy gradient approaches (*e.g.*, PPO) to align with human preferences. These methods explicitly cast diffusion denoising as a multi-step decision process and update the policy accordingly. DDPO [2] improves alignment and aesthetics by optimizing rewards tied to human feedback. DPOK [8] studies online RL with Kullback–Leibler (KL) regularization on SD [26]. As a complementary approach to RL, Diffusion-DPO [32] successfully adapts direct preference optimization to diffusion likelihoods, achieving significant improvements in human-preference alignment on SDXL [24]. Beyond standard diffusion models, Flow-GRPO [21] is the first to bring GRPO into flow-matching model. Building on this, TempFlow-GRPO [10] further improves efficiency and stability by introducing trajectory branching and enabling process-level rewards without an intermediate reward model. Motivated by these successes, we present the first integration of policy gradient-based online RL into the BFR domain.

3 Preliminary

BFR Problem Modeling. BFR is an ill-posed inverse problem. From a mathematical perspective, given a low-quality observation $\mathbf{c}_{\text{LQ}}$, the posterior distribution of its corresponding high-quality $\mathbf{x}_0$, denoted as $p(\mathbf{x}_0|\mathbf{c}_{\text{LQ}})$, has multiple feasible solutions [23]. This posterior can be modeled as a mixture distribution:

$$p(\mathbf{x}_0|\mathbf{c}_{\text{LQ}}) = \sum_{k=1}^K w_k \cdot p_k(\mathbf{x}_0|\mathbf{c}_{\text{LQ}}). \quad (1)$$

Here, K represents the number of possible high-quality solutions. Each distribution $p_k(\mathbf{x}_0|\mathbf{c}_{LQ})$ represents a real face distribution compatible with $\mathbf{c}_{LQ}$, having a specific identity, expression, or detail, with a peak at $\boldsymbol{\mu}_k$. The w_k denotes the probability of the k -th solution, where $\sum w_k = 1$. This multi-solution nature poses a challenge for existing methods. Without dedicated exploration mechanisms, restoration models tend to converge toward average solutions, resulting in blurry faces with weakened textural detail [22].

Diffusion Process as MDP. Diffusion models generate data through forward noise addition and reverse denoising processes. In the forward diffusion process, the data is progressively perturbed by Gaussian noise, defined as $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I})$, where $\alpha_t \in (0, 1]$ controls the noise intensity, and $\mathbf{x}_t$ represents the noisy data at time step t .

The denoising process recovers data via the conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$, where $\mathbf{c}$ is a condition. In the DDIM [31] framework, the network predicts the noise $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$. The one-step denoising formula (from t to $t - 1$) is given by:

$$\mu_\theta(\mathbf{x}_t, t, \mathbf{c}) = \sqrt{\alpha_{t-1}} \cdot \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \cdot \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})}{\sqrt{\alpha_t}} + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}), \quad (2)$$

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t, \mathbf{c}) + \sigma_t \cdot \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

where $\sigma_t = \eta \sqrt{\frac{1 - \alpha_{t-1}}{1 - \alpha_t}} \cdot \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)$, and $\eta \in [0, 1]$ controls the randomness. When $\eta = 0$, the sampling is deterministic, producing a fixed generation path. When $\eta > 0$, random noise ϵ is introduced, bringing randomness.

[8] proposes that DDIM can be formulated as a Markov Decision Process (MDP) defined by the tuple $(S, A, \rho_0, \pi, P, R)$. At time step t , the state is $\mathbf{s}_t \triangleq (\mathbf{c}, t, \mathbf{x}_t)$. The action is the denoised sample predicted by the model, $\mathbf{a}_t \triangleq \mathbf{x}_{t-1}$, with the policy defined as $\pi(\mathbf{a}_t|\mathbf{s}_t) \triangleq \pi_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$. The state transition is deterministic, given by $P(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \triangleq (\delta_{\mathbf{c}}, \delta_{\mathbf{a}_t})$, where $\delta_{\mathbf{c}}$ denotes the Dirac distribution at $\mathbf{c}$. The initial state distribution is $P_0(\mathbf{s}_0) \triangleq (p(\mathbf{c}), \mathcal{N}(\mathbf{0}, \mathbf{I}))$. The reward is provided only at the final step: $R(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{x}_0, \mathbf{c})$. When $\eta > 0$, the DDIM MDP can achieve reinforcement learning training.

4 Methodology

LRPO is an approach that constructs a BFR-native optimization algorithm using online RL, as shown in Figure 2. First, we initialize its policy network with an *off-the-shelf* diffusion-based face restoration model. Specifically, we employ DiffBIR [20] (our base model), a ControlNet-based [48] approach that uses the LQ input as a control signal to guide restoration. Rather than naively applying standard RL algorithms, we propose three core innovations tailored for the BFR task: (1) We design a composite reward function that provides comprehensive assessments for the diffusion denoising process. (2) We propose

a ground-truth-guided likelihood regularization term to structurally anchor the policy to authentic facial manifolds. (3) We develop a noise-aware advantage assignment mechanism to appropriately weight advantages based on the physical significance of each denoising step. LRPO is designed to post-train off-the-shelf face restoration models via RL. Crucially, since the fully-trained base model already possesses strong face restoration capabilities, our RL alignment is highly data-efficient, requiring only a limited subset of standard training data to extend the performance frontier.

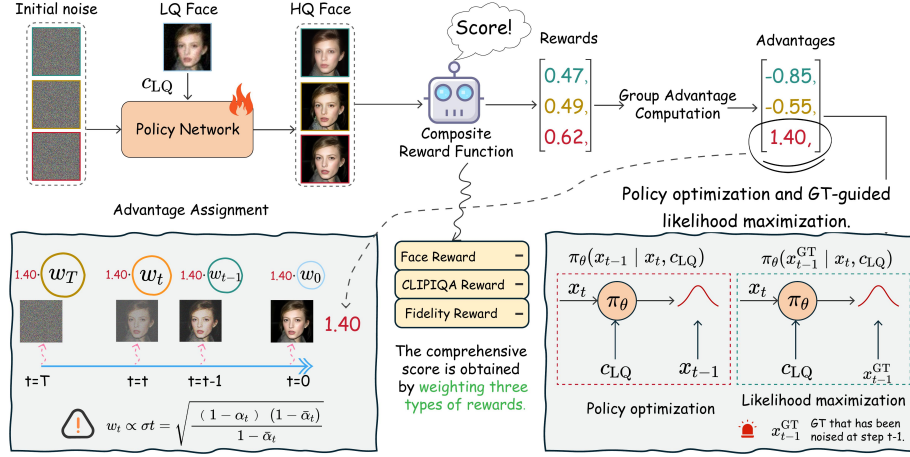


Fig. 2: The overview of our proposed LRPO framework. The policy network produces multiple HQ restoration candidates from a single LQ input, which are then assessed by the reward function and transformed into advantage scores. The framework assigns weighted advantage scores to individual denoising steps according to their contribution to restoration quality, and integrates ground-truth-guided likelihood regularization into the RL optimization objective to maintain fidelity.

4.1 Multi-Candidate Exploration via Relative Advantage

Instead of forcing the diffusion model to learn a deterministic trajectory, we formulate the optimization as a multi-candidate exploration process. Inspired by recent advances in group-based RL [28], we sample a group of G trajectories $\{\{\mathbf{x}_T^{(i)}, \mathbf{x}_{T-1}^{(i)}, \dots, \mathbf{x}_0^{(i)}\}\}_{i=1}^G$ from the policy $\pi_{\theta_{old}}$, obtaining G candidate reconstructions $\hat{\mathbf{x}}_0^{(i)}$, which are decoded from $\mathbf{x}_0^{(i)}$ in latent diffusion. We then compute their rewards $r^{(i)} = r(\hat{\mathbf{x}}_0^{(i)}, \mathbf{x}_{GT})$, where $\mathbf{x}_{GT}$ denotes the ground truth corresponding to $\hat{\mathbf{x}}_0^{(i)}$. Then, the advantage of the i -th sampled trajectory at time

t is calculated by normalizing the group-level rewards:

$$\hat{A}_t^{(i)} = \frac{r^{(i)} - \text{mean}(\{r^{(i)}\}_{i=1}^G)}{\text{std}(\{r^{(i)}\}_{i=1}^G)}. \quad (4)$$

By utilizing this relative advantage, the policy model is updated by maximizing the following objective:

$$\begin{aligned} \mathcal{J}_{\text{RA}}(\theta) = & \mathbb{E}_{\mathbf{c}_{\text{LQ}} \sim p(\mathbf{c}_{\text{LQ}}), \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | \mathbf{c}_{\text{LQ}})} \\ & \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^{T-1} \left[\min \left(r_t^{(i)}(\theta) \hat{A}_t^{(i)}, \text{clip} \left(r_t^{(i)}(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t^{(i)} \right) - \beta D_{KL}(\pi_{\theta} \| \pi_{\text{ref}}) \right], \end{aligned} \quad (5)$$

where $r_t^{(i)}(\theta) = \frac{\pi_{\theta}(\mathbf{x}_{t-1}^{(i)} | \mathbf{x}_t^{(i)}, \mathbf{c}_{\text{LQ}})}{\pi_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^{(i)} | \mathbf{x}_t^{(i)}, \mathbf{c}_{\text{LQ}})}$, with T denoting the total timesteps and π_{ref} denotes the initialized pretrained model.

Due to the inherent randomness of the diffusion model, given the same LQ face $\mathbf{c}_{\text{LQ}}$, a set of diverse candidate faces $\{\hat{\mathbf{x}}_0^{(1)}, \hat{\mathbf{x}}_0^{(2)}, \dots, \hat{\mathbf{x}}_0^{(G)}\}$ can be generated from policy π_{θ} . Ideally, when π_{θ} learns a posterior distribution capturing different potential solutions, the generated samples cluster around distinct peaks $\{\boldsymbol{\mu}_k\}$ in the solution space. Such diverse candidates enable the optimization process to make meaningful comparisons and drive policy improvements.

Once obtaining the candidate set, we evaluate each generated sample $\hat{\mathbf{x}}_0^{(i)}$ using the composite reward function $r(\cdot)$. The transformation of absolute rewards $r^{(i)}$ into within-group relative advantages $\hat{A}_t^{(i)}$ (See Eq. 4) improves the ability to distinguish between high and low-quality solutions while reducing sensitivity to reward scaling. The core insight of this transformation is redirecting optimization from absolute quality assessment to relative comparison. Consequently, the policy gradient update direction is proportional to:

$$\sum_{i=1}^G \sum_{t=0}^{T-1} \hat{A}_t^{(i)} \nabla_{\theta} \log \pi_{\theta}(\mathbf{a}_t^{(i)} | \mathbf{s}_t^{(i)}). \quad (6)$$

Thus, when a sample's reward exceeds the group average, its advantage $\hat{A}_t^{(i)}$ is positive, increasing the policy's selection probability; conversely, below-average rewards yield negative advantages $\hat{A}_t^{(j)}$, reducing selection likelihood for suboptimal solutions. This group sampling approach enables parallel exploration of the policy's learned solution space. To fully exploit this mechanism and adapt it to the unique physical dynamics of BFR, we introduce three key innovations that structurally reformulate the optimization objective.

4.2 Likelihood-Regularized Policy Optimization

Composite Reward Function. To effectively guide the optimization of the policy network π_{θ} and enable it to find a better balance in the complex Perception-Distortion Tradeoff [3], we design a multi-objective composite reward function,

$R(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{GT}})$. This composite function measures the quality of generated face images $\hat{\mathbf{x}}_0$ from three complementary perspectives:

- *Human Preference Reward* (r_{pref}): We employ a Face Reward Model [39], pre-trained on a human preference dataset, to score the overall realism and naturalness of facial details, ensuring alignment with human aesthetic preferences.
- *Perceptual Quality* (r_{aq}): We incorporate CLIPQA [33], a no-reference image quality assessment metric, to objectively measure perceptual quality using knowledge from pre-trained CLIP [25] models.
- *Fidelity Reward* (r_{fid}): We formulate a fidelity reward based on feature similarity and wavelet low-frequency constraints to enforce identity consistency, resulting in substantial fidelity improvements (see the supplementary material for implementation details).

The final total reward $R(\cdot)$ is defined as the weighted sum of these three components:

$$R(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{GT}}) = \lambda_1 r_{\text{pref}}(\hat{\mathbf{x}}_0) + \lambda_2 r_{\text{aq}}(\hat{\mathbf{x}}_0) + \lambda_3 r_{\text{fid}}(\hat{\mathbf{x}}_0, \mathbf{x}_{\text{GT}}). \quad (7)$$

where $\lambda_{(\cdot)}$ are the weight coefficients for each term.

Ground-truth Guided Likelihood Regularization. Optimization based solely on reward maximization is susceptible to reward hacking—the policy may exploit reward function biases to produce aesthetically pleasing but unfaithful outputs. We address this by replacing the standard KL divergence penalty with a ground-truth (GT) guided likelihood regularization $\mathcal{R}_{\text{likelihood}}$, activated only in the final S timesteps.

The regularization leverages GT supervision by using ideal denoising trajectories derived from $\hat{\mathbf{x}}^{\text{GT}}$. For each GT image, we pre-compute the ideal latent trajectory $\{\mathbf{x}^{\text{GT}}\}_{t=0}^T$ through forward noising. During training, when the policy π_θ samples a state $\mathbf{x}_t$, we apply regularization by maximizing the log-likelihood of the model producing the ideal subsequent state $\mathbf{x}_{t-1}^{\text{GT}}$. Thus, the regularization item is defined as:

$$\mathcal{R}_{\text{likelihood}} = -\log \pi_\theta(\mathbf{x}_{t-1}^{\text{GT}} | \mathbf{x}_t, \mathbf{c}_{\text{LQ}}) \propto \frac{\|\mathbf{x}_{t-1}^{\text{GT}} - \mu_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{LQ}})\|_2^2}{2\sigma_t^2}. \quad (8)$$

where μ_θ and σ_t refer to the mean and standard deviation of DDIM’s one-step denoising, with the detailed form provided in Sec. 3.

In standard RL frameworks, KL divergence prevents policy collapse by anchoring the new policy to the pre-trained reference model π_{ref} . However, in the BFR task, π_{ref} often yields inferior, blurry priors when conditioned on severely degraded inputs. Forcing a KL constraint would structurally lock the policy into these suboptimal priors. By completely discarding the KL divergence and substituting it with Eq. 8, we grant the model the freedom to break away from inferior priors while safely steering the exploration towards the authentic facial manifold.

Noise-Level Advantage Assignment. Conventional GRPO-based approaches [8, 21] treat all timesteps equally when assigning advantage weights, ignoring the inherently non-uniform importance of different steps in the diffusion process. To

address this, inspired by previous work [10], we introduce a noise-level-aware advantage assignment approach that correlates advantage weights with the exploration magnitude achieved at each denoising step. In DDIM sampling, the single-step exploration radius from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ is determined by the standard deviation σ_t of the added noise. Therefore, we set the timestep weight w_t proportional to σ_t :

$$w_t \propto \sigma_t, \quad \text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} w_t = 1. \quad (9)$$

The specific form of σ_t is detailed in Sec. 3. After normalizing the weights $\{w_t\}_{t=1}^T$, we apply them to weight the original advantage $\hat{A}_t^{(i)}$, yielding the final advantage $\tilde{A}_t^{(i)}$ for policy update:

$$\tilde{A}_t^{(i)} = w_t \cdot \hat{A}_t^{(i)}. \quad (10)$$

Since σ_t decreases from high initial values to nearly zero, early denoising steps possess greater exploration capability. By weighting these steps more heavily in advantage computation, we facilitate enhanced exploration that yields more diverse high-quality restoration outcomes. Simultaneously, it reduces interference during the high-frequency detail refinement that occurs in later denoising phases. **LRPO Optimization Objective.** By unifying these diffusion-native strategies, we construct the final LRPO optimization objective. It fundamentally reformulates Eq. 5 by embedding the noise-aware advantages $\tilde{A}_t^{(i)}$ and substituting the restrictive KL penalty with our likelihood regularization $\mathcal{R}_{\text{likelihood}}^{(i)}$:

$$\begin{aligned} \mathcal{J}_{\text{LRPO}}(\theta) = & \mathbb{E}_{\mathbf{c}_{\text{LQ}} \sim p(\mathbf{c}_{\text{LQ}}), \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | \mathbf{c}_{\text{LQ}})} \\ & \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^{T-1} \left[\min \left(r_t^{(i)}(\theta) \tilde{A}_t^{(i)}, \text{clip} \left(r_t^{(i)}(\theta), 1 - \epsilon, 1 + \epsilon \right) \tilde{A}_t^{(i)} \right) - \alpha \mathcal{R}_{\text{likelihood}}^{(i)} \right], \end{aligned} \quad (11)$$

5 Experiments

5.1 Experimental Settings

Experimental hyperparameters and composite reward function configurations are provided in the supplementary material.

Training and Testing Data. For our RL post-training, we utilize only a small subset of face images randomly sampled from the FFHQ [13] dataset, highlighting the high data efficiency of our LRPO framework. The degradation strategy from HQ to LQ is based on the following degradation function:

$$\mathbf{I}_{\text{LQ}} = \left\{ \left[(\mathbf{I}_{\text{HQ}} \otimes \mathbf{k}_{\sigma})_{\downarrow_r} + \mathbf{n}_{\delta} \right]_{\text{JPEG}_q} \right\}_{\uparrow_r} \quad (12)$$

where the HQ images are first convolved with a Gaussian kernel $\mathbf{k}_{\sigma}$, followed by a downsampling with a factor of r , and then corrupted with Gaussian noise $\mathbf{n}_{\delta}$.

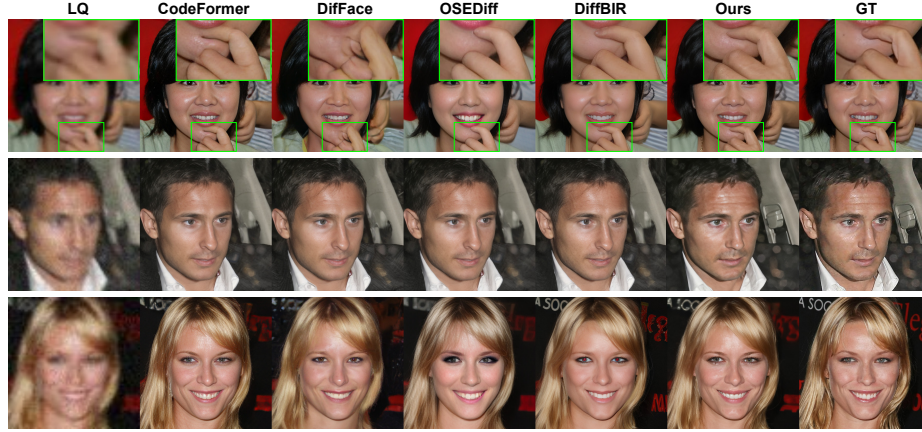


Fig. 3: Qualitative results on the CelebA-Test dataset. (Zoom in for details)

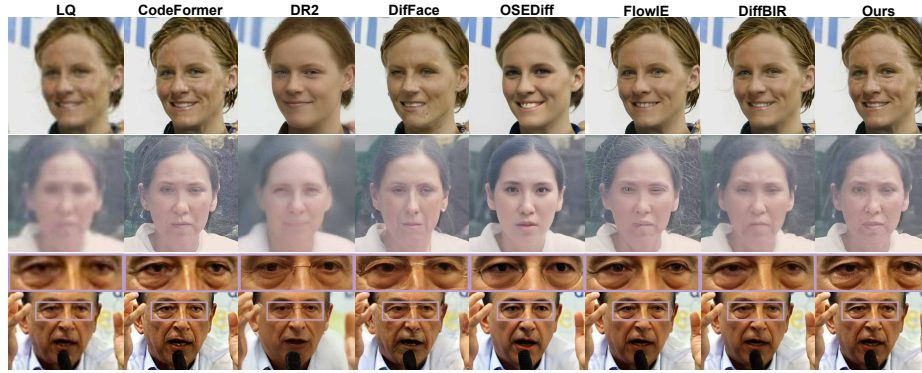


Fig. 4: Qualitative results on real-world datasets. (Zoom in for details)

Subsequently, the images undergo JPEG compression with a quality factor of q . Finally, the LQ image is resized back to the original 512×512 . Here, σ , r , δ , and q are randomly sampled from the intervals $[0.1, 12]$, $[1, 12]$, $[0, 15]$, and $[30, 100]$, respectively. Following previous work [9, 35], we employ the synthetic dataset CelebA-Test [14] and two real-world datasets [35] (*i.e.*, LFW-Test and WebPhoto-Test) to validate our method.

Evaluation Metrics. On the CelebA-Test dataset, we utilized six common reference-based metrics: SSIM [37], PSNR, LPIPS [49], CLIP Score [11], Deg. (ArcFace embedding angle to GT) [35], and LMD (mean L2 distance to GT over 98 landmarks) [9], along with four non-reference metrics: MUSIQ [17], MANIQA [42], CLIPQA [33], and Aesthetic.

Comparison Methods. We compare with not only the base models but also the latest state-of-the-art methods, including GFPGAN [4], CodeFormer [51], VQFR [9], DR2+SPAR [36], RestoreFormer [38], DiffFace [47], OSedDiff [40],

DiffBIR [20], FlowIE [52] and InterLCM [19]. Among them, FlowIE and InterLCM use ODE solvers. For fairness, all baselines use face-specific checkpoints; LRPO and DiffBIR both start from the BFR checkpoint. DiffBIR serves as a public ControlNet-based backbone for reproducible post-training: LRPO freezes SwinIR, updates only the second-stage ControlNet, and can extend beyond BFR by replacing task-specific checkpoints and rewards.

5.2 Main Results

Table 1: Performance comparisons on CelebA-Test. The highest score for each metric is highlighted in **red**, the second-highest in **blue**. Metrics with $\uparrow$ indicate higher is better, $\downarrow$ means lower is better. Values in parentheses represent our method’s improvements over the base model.

Methods	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	CLIP Score $\uparrow$	Deg. $\downarrow$	LMD $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	Aesthetic $\uparrow$	CLIPQA $\uparrow$
Input	0.6994	25.33	0.4866	0.7894	47.94	3.7560	17.00	0.3957	4.0484	0.2957
GFP-GAN	0.6772	24.65	0.3646	0.8410	34.58	2.4110	73.90	0.6522	5.6992	0.6781
CodeFormer	0.6925	25.85	0.3335	0.8931	31.08	1.9963	74.23	0.6520	5.8103	0.6493
VQFR	0.6654	23.76	0.3557	0.8562	42.48	2.9444	73.84	0.6544	5.7844	0.6750
DR2+SPAR	0.6512	22.89	0.4146	0.7437	57.24	4.5449	70.19	0.6374	5.6602	0.5960
DiffFace	0.6762	24.80	0.3994	0.8380	45.81	2.9766	68.96	0.6204	5.4708	0.5711
OSDiff	0.6864	23.96	0.3478	0.7962	46.20	2.8871	73.41	0.6560	5.7720	0.6120
FlowIE	0.6769	24.85	0.3442	0.8961	33.44	2.1995	74.08	0.6720	5.6782	0.6866
InterLCM	0.6819	24.88	0.3349	0.8905	33.58	2.1519	75.16	0.6781	5.7735	0.6748
DiffBIR	0.6775	25.44	0.3811	0.8877	35.16	2.2661	74.46	0.6752	5.7943	0.7200
LRPO (ours)	0.7021	26.15	0.3635	0.9100	31.19	1.9533	75.24	0.6808	5.8126	0.8061
	(+0.0246)	(+0.71)	(+0.0176)	(+0.0223)	(+3.97)	(+0.3128)	(+0.78)	(+0.0056)	(+0.0183)	(+0.0861)

Results on Synthetic Data. As shown in Table 1, LRPO achieves improvements on all metrics compared with DiffBIR on the synthetic CelebA-Test dataset. These results indicate that our RL framework simultaneously improves perceptual quality and identity preservation in restored faces. Furthermore, LRPO achieves superior performance compared to state-of-the-art approaches across the majority of evaluation metrics, including SSIM, LMD, and MUSIQ, confirming that it enhances identity consistency while maintaining perceptual quality.

Figure 3 demonstrates LRPO’s superior performance over methods that fail to restore faces. LRPO delivers more realistic textures than the baseline, better identity alignment than OSDiff, and more natural results without the over-smoothing seen in OSDiff and DiffBIR. The supplementary material contains the complete qualitative results.

Results on Real-world Data. Table 2 shows the quantitative performance evaluation on real-world datasets LFW-Test and WebPhoto-Test. LRPO demonstrates significant performance gains compared to the base DiffBIR and outperforms other state-of-the-art approaches on MUSIQ and CLIPQA metrics. Qualitative results are illustrated in Figure 4. Due to severe degradation in real-world inputs, many approaches fail to restore texture details. In contrast, our method recovers more details while introducing fewer artifacts.

Table 2: Performance comparisons on wild datasets. The highest score is highlighted in red, and the second-highest in blue. Metrics with $\uparrow$ indicate higher is better, $\downarrow$ means lower is better. Values in parentheses represent our method’s improvements over the base model.

Dataset Methods	LFW-Test		WebPhoto-Test	
	MUSIQ $\uparrow$	CLIPQA $\uparrow$	MUSIQ $\uparrow$	CLIPQA $\uparrow$
Input	26.87	0.2834	18.63	0.4128
GFP-GAN	73.57	0.6983	72.09	0.6888
CodeFormer	70.69	0.6335	71.16	0.6573
VQFR	74.39	0.7100	70.91	0.6767
DR2+SPAR	72.22	0.6427	63.65	0.5586
DiffFace	69.85	0.6110	65.21	0.5821
OSDiff	73.40	0.6327	72.60	0.6454
FlowIE	64.29	0.5974	71.45	0.6838
InterLCM	74.18	0.6588	73.91	0.6658
DiffBIR	73.71	0.7296	67.45	0.6630
LRPO (ours)	74.60 (+0.89)	80.73 (+0.0777)	72.71 (+5.26)	70.40 (+0.0410)

Human Preference Evaluation. A user study was conducted with 30 participants of varying backgrounds to evaluate 100 face images from the CelebA-Test dataset. Participants evaluated our method against the base model (DiffBIR) on two criteria: *fidelity* (identity preservation) and *realism* (naturalness with minimal artifacts). As shown in Table 3, our method outperforms the base model in both fidelity and realism according to human preferences. This confirms that the base model enhanced by LRPO aligns with human preferences, as validated by human subjective assessment.

5.3 Ablation Study

We conduct the ablation study on the synthetic dataset. As shown in Table 4, we analyze the effects of four key components: Reinforcement Learning (RL), Kullback-Leibler divergence (KL), GT-guided likelihood regularization (Reg.), and noise-level advantage assignment (AdA). The variants are designed to isolate each component under comparable training settings. Variant 2 corresponds to the standard GRPO objective with the KL divergence term, while Variant 1 removes this KL term from the optimization objective. Variants 3 and 4 both follow the GRPO training framework but isolate the two LRPO-specific designs: Variant 3 adds GT-guided likelihood regularization while assigning the advantage uniformly across timesteps, whereas Variant 4 uses the proposed noise-level advantage assignment without the likelihood regularization term. LRPO combines both Reg. and AdA to jointly improve fidelity preservation and timestep-aware exploration. Variant 1 demonstrates improvements across all metrics after incorporating RL, confirming that RL directly enhances BFR performance. However, adding KL divergence in Variant 2 degrades all metrics without improving visual quality (See Figure 5(a)). We therefore remove KL divergence from the RL objective in Eq. 11, reducing computational cost while maintaining visual quality. Without AdA

Table 3: User study. Participants selected the winner between DiffBIR and LRPO restored images in terms of fidelity and realism.

Comparison	Fidelity %	Realism %
DiffBIR vs LRPO	39.4% vs 60.6%	29.5% vs 70.5%

Table 4: Ablation Study of LRPO

Struct	RL	KL	Reg	AdA	SSIM↑	LMD↓	CLIPQA↑
Base					0.6775	2.2661	0.7200
Reg only			✓		0.7150	2.1864	0.3654
Variant 1	✓				0.6849	2.0503	0.7809
Variant 2	✓	✓			0.6750	2.1602	0.7816
Variant 3	✓		✓		0.6867	2.0078	0.7852
Variant 4	✓			✓	0.6806	1.9551	0.7778
LRPO	✓	✓	✓	✓	0.7021	1.9533	0.8061

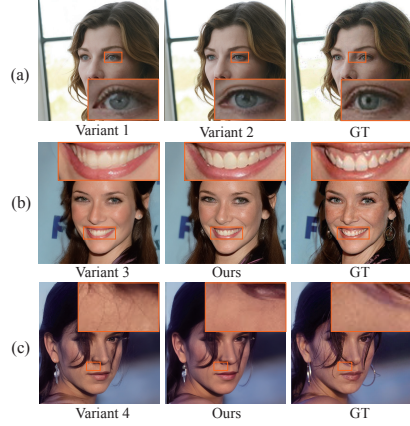


Fig. 5: Ablation study visualizations.

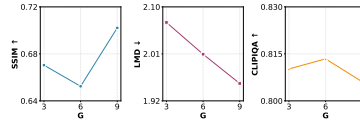
(Variant 3), the model suffers from detail blurring caused by over-optimization during late denoising stages (Figure 5(b)). Removing Reg. (Variant 4) leads to decreased SSIM scores and poor fidelity, with the model producing unrealistic, fantasy-like textures as shown in Figure 5(c). Furthermore, applying GT-guided regularization alone with RL disabled (“Reg only” in Table 4) preserves high SSIM but collapses perceptual quality, confirming that the perceptual gains stem from RL-based exploration rather than from GT supervision alone.

Figure 7 illustrates the training dynamics of the reward score: the noise-level advantage assignment consistently outperforms uniform weighting, achieving superior reward scores and faster discovery of optimal restoration solutions. Here, one training step denotes one complete denoising trajectory on which the policy is updated.

Table 5: Ablation of Reward Functions.

Reward Composition	Deg. ↓	CLIPQA ↑	MUSIQ ↑
r_{fid}	33.06	0.6823	65.09
r_{pref}	36.93	0.7375	74.19
$r_{\text{pref}} + r_{\text{aq}}$	43.07	0.8170	76.93
$r_{\text{pref}} + r_{\text{fid}}$	32.52	0.7355	73.51
$r_{\text{pref}} + r_{\text{fid}} + r_{\text{aq}}$	31.19	0.8061	75.24

Fig. 6: Quantitative metrics under different group sizes G .



We conduct an ablation on the composite reward to justify our reward setting in Table 5. While r_{fid} alone ensures fidelity and identity consistency, it lacks perceptual details. Using r_{pref} alone improves perceptual quality but weakens identity consistency, and combining r_{pref} with r_{aq} further increases CLIPQA and MUSIQ at the cost of substantially worse Deg. The full composition is therefore crucial: it balances perceptual quality with fidelity and identity preservation.

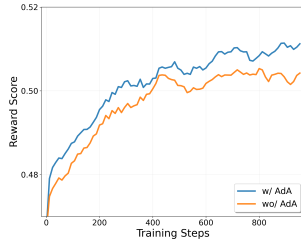


Fig. 7: Reward scores during training with and without noise-level advantage assignment.



Fig. 8: Reward curves of different S values within the likelihood regularization term.



Fig. 9: On the CelebA-Test dataset, we qualitatively compare InterLCM with InterLCM (+ours).

In addition, Figure 6 shows the experiments conducted to select the number of trajectories G within each group. Because $G = 9$ exhibits superior performance in terms of both structural similarity and identity consistency, we set $G = 9$ for the training stage. The proposed likelihood regularizer is applied only during the last S denoising steps. Throughout training, varying S leads to different reward curves, as shown in Figure 8. Choosing an excessively large S introduces overly strong penalties that conflict with the RL learning trajectory, ultimately causing training collapse. $S = 5$ achieves convergence to a higher reward score without causing training collapse; therefore, we adopt $S = 5$.

5.4 Scalability to Few-Step Diffusion

To directly address the applicability of our RL framework to rapidly developing few-step generation paradigms, we validate LRPO’s generality by applying it to InterLCM, a state-of-the-art few-step diffusion face restoration model. Fig. 9 presents a qualitative comparison between the baseline InterLCM and its counterpart fine-tuned with LRPO (denoted as +ours). The results demonstrate that InterLCM (+ours) yields more realistic and richer facial details while maintaining high inference speed. This validates that LRPO successfully generalizes across diverse diffusion solvers and remains highly effective for ultra-fast inference scenarios.

5.5 SFT vs. RL Exploration



Fig. 10: SFT vs. RL, zoom in for details on two examples.

To isolate RL exploration, we compare LRPO with a same-reward SFT variant that removes group sampling, relative advantage, and PPO clipping (Table 6, Figure 10). SFT directly backpropagates the reward gradient; it is consistently worse and shows stylization and missing details, confirming that the gains come from RL exploration under GT-guided regularization.

Table 6: SFT vs. RL on CelebA-Test.

	Deg.↓	MANIQA↑	MUSIQ↑	CLIP Score↑
SFT	36.35	0.665	74.94	0.868
LRPO	31.19	0.681	75.24	0.910

6 Limitations

LRPO can still struggle with highly individualized attributes, such as complex jewelry and accessories (Figure 11). These rare objects are underrepresented in the base model’s training data, leading to blurred or distorted high-frequency details. Inference speed matches the base model, but RL post-training adds online sampling and reward computation; future work will improve data diversity and training efficiency.

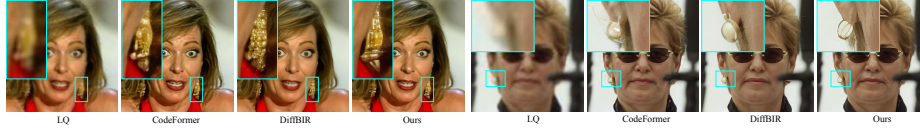


Fig. 11: Failure cases on rare individualized objects such as jewelry.

7 Conclusion

We propose LRPO, the first online reinforcement learning framework for BFR. By combining RL exploration, composite image rewards, GT-guided likelihood regularization, and noise-level advantage assignment, LRPO moves beyond deterministic restoration while preserving identity. Experiments show consistent gains over existing approaches in both perceptual quality and identity consistency.

Acknowledgements

This work was supported by the Natural Science Foundation of Beijing, China (Grant No. L252025), and the National Natural Science Foundation of China (Grant Nos. 62372033 and 62120106009).

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023), <https://arxiv.org/abs/2305.13301>
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
4. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
5. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
6. Chen, X., Tan, J., Wang, T., Zhang, K., Luo, W., Cao, X.: Towards real-world blind face restoration with generative diffusion prior. IEEE Transactions on Circuits and Systems for Video Technology (2024)
7. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
8. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)
9. Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., Cheng, M.M.: Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In: European Conference on Computer Vision (ECCV) (2022)
10. He, X., Fu, S., Zhao, Y., Li, W., Yang, J., Yin, D., Rao, F., Zhang, B.: Tempflow-grpo: When timing matters for grpo in flow models. arXiv preprint arXiv:2508.04324 (2025)
11. Hessel, J., Holtzman, A., Forbes, M., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) (2020)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis Machine Intelligence **43**(12), 4217–4228 (dec 2021). <https://doi.org/10.1109/TPAMI.2020.2970919>
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

18. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. arXiv preprint arXiv:1908.08239 (2019)
19. Li, S., Wang, K., van de Weijer, J., Khan, F.S., Guo, C.L., Yang, S., Wang, Y., Yang, J., Cheng, M.M.: Interlcm: Low-quality images as intermediate states of latent consistency models for effective blind face restoration. arXiv preprint arXiv:2502.02215 (2025)
20. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European Conference on Computer Vision (ECCV) (2024)
21. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
22. Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R.: Srflow: Learning the super-resolution space with normalizing flow. In: European conference on computer vision. pp. 715–732. Springer (2020)
23. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
27. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
28. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
29. Skalse, J., Howe, N., Krashennnikov, D., Krueger, D.: Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems* **35**, 9460–9471 (2022)
30. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML) (2015)
31. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
32. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
33. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2023)

34. Wang, J., Tian, Z., Wang, X., Zhang, X., Huang, W., Wu, Z., Jiang, Y.G.: Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. arXiv preprint arXiv:2504.11455 (2025)
35. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
36. Wang, Z., Zhang, Z., Zhang, X., Zheng, H., Zhou, M., Zhang, Y., Wang, Y.: Dr2: Diffusion-based robust degradation remover for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)* **13**(4), 600–612 (2004)
38. Wang, Z., Zhang, J., Chen, R., Wang, W., Luo, P.: Restoreformer: High-quality blind face restoration from undegraded key-value pairs (2022)
39. Wu, B., Wang, W., Liu, Y., Li, Z., Zhao, Y.: Diffusionreward: Enhancing blind face restoration through reward feedback learning. arXiv preprint arXiv:2505.17910 (2025)
40. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems (NeurIPS)* (2024)
41. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
42. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
43. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR) (2021)
44. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
45. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: European conference on computer vision (ECCV) (2018)
46. Yuan, S., Liu, Y., Yue, Y., Zhang, J., Zuo, W., Wang, Q., Zhang, F., Zhou, G.: Ar-grpo: Training autoregressive image generation models via reinforcement learning. arXiv preprint arXiv:2508.06924 (2025)
47. Yue, Z., Loy, C.C.: Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2024)
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
50. Zhou, H., Huang, C., Gao, S., Zhuang, X.: Vspsr: Explorable super-resolution via variational sparse representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

51. Zhou, S., Chan, K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems (NeurIPS)* (2022)
52. Zhu, Y., Zhao, W., Li, A., Tang, Y., Zhou, J., Lu, J.: Flowie: Efficient image enhancement via rectified flow. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024)