ID-PreFeR: ID-Preserving Face Restoration with Mixed Data Quality

Chengxuan Zhu[1,2†], Yuchen Hong[1,2†], Qi Zhang[2], Bingtao Fu[3], Jinxiu Liang[4], Jinwei Chen[2], Huaqi Zhang[2], Chao Xu[1], Boxin Shi[1*], and Qingnan Fan[2*]

[1]Peking University [2]vivo BlueImage Lab
[3]Ant Group [4]National Institute of Informatics, Japan
{peterzhu,shiboxin}@pku.edu.cn, qingnanfan@vivo.com

Abstract. This paper introduces ID-PreFeR, a robust identity-preserving face restoration method that tackles the ill-posed face restoration problem by incorporating personalized identity information. Existing approaches often suffer from high training and storage costs and are sensitive to the quality of reference images. To address these issues, we propose a lightweight personalization injector that enables efficient personalization without the need for regularization data. We also introduce an identity–quality disentanglement training strategy to ensure robust identity learning, even when some reference images are of poor quality. Furthermore, we propose an identity-preserving sampling strategy to enhance identity fidelity during inference. Extensive experiments on both synthetic data and a newly collected real-world mobile-phone dataset verify the effectiveness and practicality of the proposed method.

Keywords: Diffusion · Face Restoration · Image Enhancement

1 Introduction

Low-quality (LQ) facial images are ubiquitous in practice, *e.g.*, small faces in mobile photos or unclear portraits collected online. Face restoration aims to reconstruct high-quality (HQ) facial images from degraded LQ inputs affected by blur, low resolution, noise, and compression artifacts. This problem is inherently ill-posed because multiple plausible HQ faces can correspond to the same LQ observation, and the challenge is amplified by real-world variations in illumination, pose, and expression. Because human perception is highly sensitive to facial details, even small deviations can noticeably alter perceived identity. As shown in Fig. 1, as degradation increases, a representative blind face restoration method [32] is more likely to drift from the input identity. These observations

* Corresponding authors. † Work done during their internship at vivo.
[1]Chengxuan Zhu and Chao Xu are with National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. Yuchen Hong and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

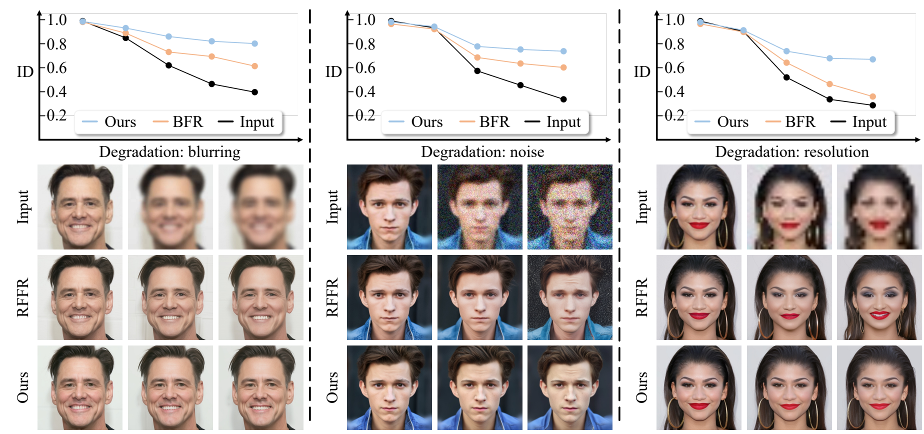

Fig. 1: The restoration results of the proposed method and a reference-free face restoration (RFFR) method [32], with different types of degradation. The more the degradation, the less identity information (measured by ID score [10]) exists in the input image, showing the effectiveness of the proposed method for identity preservation.

motivate a restoration model that jointly recovers realistic facial details and preserves identity.

Previous face restoration methods have substantially improved visual quality by leveraging generative priors from GANs [27, 41, 48, 57], codebooks [13, 15, 67], and diffusion models [8, 51, 55]. However, pretrained priors typically lack identity-specific information, which can lead to identity inconsistency and loss of crucial facial cues. To improve faithfulness, early personalized face restoration methods [12, 29–31] use multiple HQ references of the same identity to guide restoration. Yet differences across references (pose, illumination, makeup, and accessories) can cause misalignment and weaken feature transfer. Recent methods [5, 11, 43] instead fine-tune diffusion models [39] on identity-specific references to form personalized representations. Benefiting from diffusion priors, these methods generally improve visual quality and identity preservation.

Despite recent progress, two key challenges remain for personalized face restoration. **(1) Heavy personalization burden**. Many personalized methods [5, 11] fine-tune most or all diffusion-model parameters [39] and store a full model per identity. They also require substantial regularization data, often far larger than the reference set, to reduce catastrophic forgetting [40], which increases both training and storage cost. **(2) Strong HQ dependency**. Existing personalized methods [5, 11, 31, 37, 43] are typically learned from HQ references, so performance strongly depends on reference quality, yet in realistic scenarios one cannot guarantee that all acquired references are HQ. As illustrated in Fig. 2, the performance of prior methods degrades as the proportion of LQ references increases. A lightweight and robust personalized framework under mixed-quality references is therefore essential.

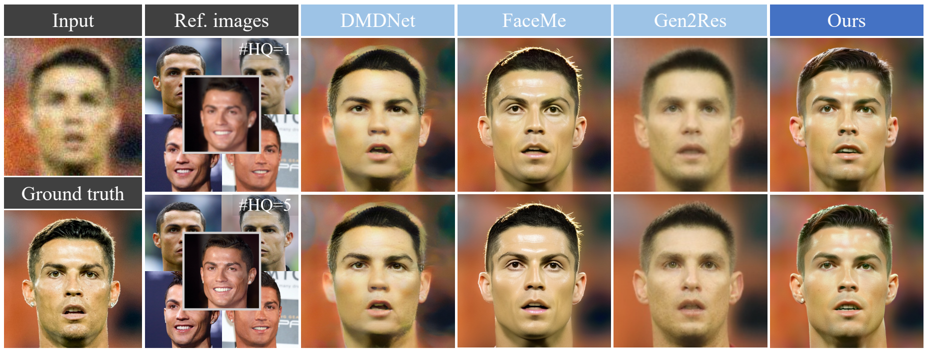

Fig. 2: The quality of results from previous personalized face restoration methods deteriorates as the quality of reference images decreases. In contrast, the proposed method maintains robust performance.

To address these challenges, we propose **ID-PreFeR**, a robust **ID-Pre**serving **F**ace **R**estoration framework. To reduce personalization cost, ID-PreFeR learns a lightweight *personalized injector* by fine-tuning only cross-attention weights with low-rank adaptation (LoRA) [20]. Trained *once per identity* from a small reference set with the diffusion backbone frozen, the injector forms an identity-specific adapter stored and loaded on demand, greatly reducing both training and storage cost versus full-model personalization. To mitigate the impact of LQ references, we introduce an *ID-quality disentanglement* strategy that uses a multi-modal large language model [60] to provide quality-aware prompt cues, encouraging the model to focus on identity-consistent characteristics instead of degradation patterns. We further propose an ID-preserving sampling strategy that uses references to guide reverse diffusion and improve identity fidelity. Adding only about 0.25% trainable parameters to the SDXL backbone [38], ID-PreFeR achieves efficient personalization and remains robust even when only one reference image is HQ, as shown in Figs. 1 and 2. Moreover, because many prior personalized benchmarks [31,43] rely on celebrity images collected online [31,35], we collect a mobile-phone dataset to evaluate performance in realistic scenarios.

Our contributions are:

- a regularization-free personalized injector that eliminates the need for external regularization data, substantially reducing the optimization burden.
- an ID–quality disentanglement strategy that improves robustness to low-quality references, mitigating the reliance on high-quality reference images.
- an ID-preserving sampling strategy that enhances identity consistency at inference time without modifying the base restoration model.
- a real-world mobile-captured dataset to evaluate personalized face restoration, featuring diverse degradations and practical capture conditions.

2 Related Works

2.1 Face Restoration

Face restoration requires strong facial priors to maintain realism and fidelity. Earlier methods exploit geometric priors such as segmentation [6,42], facial landmarks [3, 25, 36, 45], and 3D face shapes [21, 22, 68], but these are reliable only for high-quality inputs and become fragile under severe degradation. Generative priors are another major direction, leveraging GANs [4, 6, 23, 47, 68], vector-quantization codebooks [15,52,67], and diffusion models [32,51,55,61] to model realistic face distributions. Owing to their strong visual quality, diffusion models are now a central focus: DR2 [51] uses a low-pass filter to guide denoising, DiffBIR [33] first restores structure and then synthesizes details using the LQ input as control [63], and AuthFace [32] adds an adversarial loss over ControlNet [63] to enhance authenticity in sensitive regions. However, these methods are conditioned only on the LQ input and may drift from the source identity.

Given reference images of the same subject, personalized methods recover more identity-specific details: GWAINet [12] warps reference image to the input, ASFFNet [29] uses landmarks to select the most similar reference, DMDNet [31] aligns input and reference features via a learned dictionary, and PFStorer [43] fine-tunes personalization blocks over ControlNet [63]. Recent identity-preserving diffusion methods (InstantRestore [62], RestorerID [59], ReF-LDM [19], FaceMe [34]) further improve fidelity but are learned from HQ references. These methods all struggle when the references are themselves low-quality; we instead learn a neural identity representation with ID-quality disentanglement, reducing dependence on HQ references.

2.2 Text-to-Image Personalization

Text-to-image personalization adapts a generative model to a specific identity given reference images, by fine-tuning model components [1, 26, 40, 53, 58], optimizing concept-specific embeddings [1,7,14], or adding conditioning and prompt-control strategies [2, 17, 44, 63]. These often require many regularization images to avoid language drift [40] or incur high optimization cost [1,14]. In contrast, we fine-tune only LoRA [20] adapters on cross-attention modules and avoid explicit regularization-image sets by anchoring to a frozen backbone.

3 Methodology

Given several portrait references of an individual, ID-preserving face restoration aims to restore low-quality portraits while preserving the subject's identity. As illustrated in Fig. 3, ID-PreFeR separates a short *per-identity* personalization stage that learns a compact LoRA injector from a few references, and a restoration stage that reuses the frozen backbone with the stored injector. We thus do *not* train a single universal LoRA across identities: each identity is personalized once into a cached adapter, after which any number of inputs can

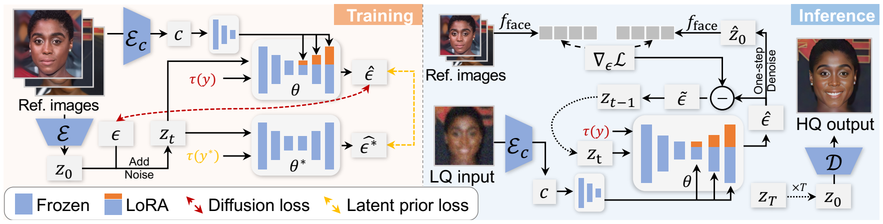

Fig. 3: Overview of ID-PreFeR. **Left: per-identity personalization (training).** Given reference images of a target identity, we optimize the LoRA weights of the cross-attention layers of the U-Net decoder (orange) and the learnable prompt tokens. The diffusion loss uses an identity prompt y (containing the identifier token `[V]` and quality tokens), while a latent prior loss computed with a class prompt y^* (without `[V]`) regularizes the personalized network against the frozen backbone θ^* to prevent language drift. **Right: restoration (inference).** For a low-quality input image, we run conditional latent diffusion with the learned injector. An ID-preserving sampling is applied, by taking a small number of gradient steps on the predicted noise at denoising step, to further increase the portrait similarity between the intermediate decoded result and the reference set.

be restored, with ID-preserving sampling applied optionally at inference. After briefly reviewing diffusion preliminaries (Sec. 3.1), we present three complementary components: a lightweight *personalized injector* (Sec. 3.2) that modifies only a fraction of the text-conditioning parameters to learn *who* the subject is; an *ID-quality disentanglement* strategy (Sec. 3.3) that prevents reference quality from being encoded as identity and exploits mixed-quality references; and an *ID-preserving sampling* strategy (Sec. 3.4) that corrects residual identity drift during generation, even when few HQ references are available.

3.1 Preliminaries

Diffusion models. In summary, diffusion models [18,49] aim to learn a mapping from the noise distribution to a desired image distribution. With a fixed forward process that adds noise to an image, diffusion models learn a denoising process that removes noise to approximate the original image. The denoising process is performed with a U-Net or Transformer parameterized by θ, and the predicted noise is denoted as ϵ_θ. It is usually conditioned on an intermediate result x_t in the process of denoising, and the current timestep t. To control the generated results towards any desired distribution, the input prompts y are turned into text embeddings $\tau(y)$ with a pretrained text encoder τ, and serve as a condition for the denoising process as well.

For efficiency, diffusion models are adapted to the latent domain [39], where an encoder $\mathcal{E}$ first maps the input image x into a latent representation $z = \mathcal{E}(x)$ with a lower resolution but richer information, and a decoder $\mathcal{D}$ maps z back into pixel domain as $x' = \mathcal{D}(z)$. The latent diffusion model (LDM) operates on the latent z and is trained with the following diffusion loss:

$$\mathcal{L}_{\mathrm{LDM}} = \left\| \epsilon - \epsilon_\theta \left(z_t, t, \tau(y) \right) \right\|_2^2, \tag{1}$$

and the diffusion process and denoising process is given by

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \hat{z} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, \tau(y))}{\sqrt{\bar{\alpha}_t}}, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0,1)$ is the added noise to z_t, and z_t is the noised latent representation at timestep t. $\hat{z}$ is the estimated denoised latent representation, while α_t is the scaling factors that control the noise level at each timestep. $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative scaling factor.

ControlNet [63] extends LDM with extra conditioning signals such as edge maps or segmentation masks. It freezes the noise prediction network and trains a copy of the encoder, initialized with the original weights and connected to the decoder through a zero-convolution, thereby leveraging the pretrained Stable Diffusion knowledge while guided by another condition c. The loss function for ControlNet is:

$$\mathcal{L}_{\mathrm{Ctrl}} = \left\| \epsilon - \epsilon_\theta \left(z_t, t, \tau(y), c \right) \right\|_2^2. \tag{3}$$

A naive solution for general image restoration is to condition the denoising process of ControlNet on the LQ image, as done by previous methods [32, 33]. However, such formulations remain severely under-determined and tend to drift from the actual identity.

3.2 Personalized Injector

The diversity of plausible face distribution calls for a design that injects the identity prior into the denoising process. Inspired by previous works on personalization [26,40], we finetune the cross attention layers of the decoder part in the noise prediction network on the provided reference images with the LoRA [20] technique, and optimize with the diffusion loss defined in Eq. (1). However, the finetuning inevitably hinders the generative capability by causing the language drift problem [28, 40], though the change is limited to a low rank.

To preserve the generative prior, previous works [40] constrain the training with regularization images, by supervising on the images from the class where the subject belongs. More recent works have alleviated the need for regularization image, and turned to the original diffusion model weights θ^* for supervision. Based on the intuition that using the class prompts without the personalized token should yield a similar response, researchers have supervised on intermediate representations, such as the text embedding [50], the attention response [66], the noise prediction results [54]. Due to the implicit nature of attention response and text embedding, we preserve the generative prior by regularizing on the predicted noise with a latent prior loss:

$$\mathcal{L}_{\mathrm{reg}} = \| \epsilon_\theta(z_t, t, \tau(y^*)) - \epsilon_{\theta^*}(z_t, t, \tau(y^*)) \|_2^2 \tag{4}$$

where y is the prompts with the identifier token, and y^* is the prompts for the class. For example, if y is "a photo of a [V] face", y^* should be "a photo of a face". The regularization is applied in the latent domain for finer control, since latent distance aligns better with pixel error than attention response does. It encourages the personalized branch to change only what is needed for identity-specific restoration, while keeping general facial priors and language semantics anchored to the pretrained model.

3.3 ID-Quality Disentanglement

Learning the identity precisely from several photos requires that each reference contributes meaningfully. However, the quality of these reference images can vary significantly, ranging from high-quality (HQ) portraits with clear details to low-quality (LQ) images degraded by noise, blur, or compression artifacts. If directly optimized on LQ images, the model may learn the artifacts as a property of the identity to personalize. To address this challenge, we propose an ID-quality disentanglement strategy that separates identity information from quality degradation in the reference images.

Quality-token construction. We model quality with two learnable tokens, one for HQ references and one for LQ references. Each reference image is assigned a binary quality label (HQ/LQ) by a lightweight MLLM-based classifier, and the corresponding token is inserted into the training prompt. Concretely, we use prompts of the form "a [Q] [V] face", where $[V]$ is the identity token and $[Q] \in \{[Q_{\mathrm{HQ}}], [Q_{\mathrm{LQ}}]\}$ is the quality token. In our implementation, $[Q_{\mathrm{HQ}}]$ and $[Q_{\mathrm{LQ}}]$ are initialized from the text embeddings of "sharp" and "blurry", respectively, and are jointly optimized with $[V]$. At inference, quality tokens are removed and only $[V]$ is retained, so that restoration is driven by identity rather than quality-specific cues: $[V]$ captures identity-consistent cues shared across references, while $[Q]$ explains quality-dependent variation.

We view the disentanglement from another perspective: each LQ reference can be seen as a shared identity-consistent latent z_{HQ} plus a per-reference residual, defined by $z_{\mathrm{LQ}}^{(i)} = z_{\mathrm{HQ}} + \epsilon_{\mathrm{deg}}^{(i)}$, where z_{HQ} is the high-quality component shared across all references of the identity, and $\epsilon_{\mathrm{deg}}^{(i)}$ is the unknown per-sample degradation residual. Substituting the LQ reference latent $z = z_{\mathrm{LQ}}$ into the forward process of Eq. (2) yields the noise combination

$$z_t = \sqrt{\bar{\alpha_t}} z_{\mathrm{HQ}} + \sqrt{1-\bar{\alpha_t}}(\epsilon + \omega_t \epsilon_{\mathrm{deg}}), \tag{5}$$

where $\omega_t = (\frac{\bar{\alpha_t}}{1-\bar{\alpha_t}})^{\frac{1}{2}}$. Since the diffusion models are trained to fully denoise the image, the output of the pretrained backbone should be close to a slightly altered noise $\epsilon + \omega_t \epsilon_{\mathrm{deg}}$, instead of ϵ.

To separate degradation from identity, the jointly optimized quality tokens absorb ϵ_{deg} while $[V]$ captures the identity-consistent cues shared across references. We further employ the Min-SNR-γ weighting during training, and Eq. (1) is modified to

$$\mathcal{L}_{\mathrm{LDM}} = \min\left(\gamma \cdot \omega_t^{-2}, 1\right) \cdot \|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2\,, \tag{6}$$

where $\gamma = 5$ is a fixed hyperparameter (selected from a small sweep over $\{1, 2, 5, 10\}$). Note that $\omega_t^2 = \bar{\alpha_t}/(1 - \bar{\alpha_t})$ is the signal-to-noise ratio (SNR). At a high timestep the SNR is low and $\omega_t \to 0$, so the degradation term $\omega_t \epsilon_{\text{deg}}$ is dominated by ϵ and vanishes on its own; the harmful regime is a low timestep, where the SNR is high and $\omega_t \epsilon_{\text{deg}}$ would otherwise push the injector to reproduce the degradation. The Min-SNR-γ weighting caps this contribution precisely when the SNR is high, offering more robust training [16], reducing overfitting to high-frequency artifacts from LQ references, and emphasizing stable identity structure. A detailed derivation is provided in the supplementary material.

3.4 ID-Preserving Sampling

Despite the ID-quality disentanglement strategy, an insufficient number of HQ reference images can still degrade performance, since the degradation in LQ images renders the target data distribution less well-defined. To mitigate this, we propose an ID-preserving sampling strategy that explicitly injects identity information into the reverse diffusion trajectory.

Let $\hat{\epsilon}_t = \epsilon_\theta(z_t, t, \tau(y), c)$ denote the predicted noise at timestep t from the personalized denoiser, where c is the conditioning signal from the LQ input (*e.g.*, ControlNet features). We employ a frozen face recognition network f_{face} [10] that maps an image to a unit-normalized identity embedding. Given the reference set, we precompute $f_{\text{face}}(x_{\text{ref}})$ and select x_{ref} as the closest reference to the input in the face-embedding space (or, when available, the best-quality HQ reference).

At each denoising step, we refine $\hat{\epsilon}_t$ by performing a small number of gradient-ascent steps on the cosine similarity between the decoded intermediate prediction and the selected reference embedding:

$$\epsilon^{(k+1)} = \epsilon^{(k)} + \delta \, \nabla_{\epsilon^{(k)}} \big\langle f_{\text{face}}\big(\mathcal{D}(\hat{z}(\epsilon^{(k)}))\big), \, f_{\text{face}}(x_{\text{ref}}) \big\rangle, \tag{7}$$

where $\hat{z}(\epsilon) = (z_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon)/\sqrt{\bar{\alpha}_t}$ is the single-step latent estimate of Eq. (2) evaluated at the current noise iterate ϵ. We initialize $\epsilon^{(0)} = \hat{\epsilon}_t$ and use $\epsilon^{(K_{\text{ID}})}$ as the refined noise estimate for the subsequent diffusion update. In practice K_{ID} is small (we use K_{ID}=1 by default) with step size δ=0.05, and the refinement is applied only to the last few denoising steps, acting as a lightweight late-stage identity correction while synthesis remains governed by the diffusion prior. The full per-step procedure is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Details

Baselines. For personalized face restoration, we select a CNN-based model (DMDNet [31]) and two diffusion-based methods (FaceMe [34] and Gen2Res [11]). We focus on baselines with public implementations and compatible protocols; several contemporary personalization frameworks either require full-model tuning with large regularization sets or are not released in a form suitable for controlled comparison under the same reference setting. In addition, we include

ALGORITHM 1: ID-Preserving Sampling (one denoising step)

Input: noisy latent z_t, timestep t, prompt embedding $\tau(y)$, conditioning c, reference image x_{ref}, step size δ, iterations K_{ID}

$\epsilon^{(0)} \leftarrow \epsilon_\theta(z_t, t, \tau(y), c)$;

for $k = 0$ **to** $K_{ID} - 1$ **do**

  $\hat{z} \leftarrow \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon^{(k)}}{\sqrt{\bar{\alpha}_t}}$;

  $\hat{x} \leftarrow \mathcal{D}(\hat{z})$;

  $\epsilon^{(k+1)} \leftarrow \epsilon^{(k)} + \delta \nabla_{\epsilon^{(k)}} \langle f_{\text{face}}(\hat{x}), f_{\text{face}}(x_{\text{ref}}) \rangle$;

end

Output: refined noise $\epsilon^{(K_{\text{ID}})}$

representative blind face restoration baselines with different generative priors: GFPGAN [47], CodeFormer [67], DiffBIR [33], and AuthFace [32].

Datasets. The base model is trained on FFHQ [24] and the Unsplash dataset [9]. Following previous works [31, 43], quantitative evaluation is conducted on CelebRef [31] with the same degradation pipeline for all compared methods. For each identity, we randomly select K=5 images as references and use the rest for evaluation; K=5 is a default rather than an intrinsic optimum, and ID-PreFeR remains effective with fewer (including mixed-quality) references. To analyze robustness to the number of HQ references (#HQ), we replace part of the HQ references with degraded versions. We additionally evaluate on real-world images captured by mobile phones under challenging conditions (low light, long distance, motion blur, and compression).

Metrics. We evaluate pixel-wise, structural, and perceptual similarity to ground truth using PSNR, SSIM, and LPIPS [64], identity similarity ("ID") [10] as cosine similarity between face embeddings, and the no-reference quality metrics CLIPIQA [46] and MANIQA [56].

Training setup. We personalize on SDXL [38] with a rank-16 (α=16) LoRA on the decoder cross-attention layers, optimized by AdamW for 500 steps at batch size 4 with a constant schedule in FP16; the LoRA weights use a $\sqrt{\cdot}$-scaled learning rate of 10^{-3}, while the jointly trained quality tokens and $[V]$ use 5×10^{-4}.

Inference setup. At test time, restoration uses the frozen backbone with the stored identity-specific injector, requiring no parameter update except the optional ID-preserving sampling (K_{ID}=1, δ=0.05) applied to late denoising steps.

4.2 Quantitative Results

Tab. 1 reports quantitative comparisons across baselines and our method, with metrics of the input LQ images and ground truth for reference. The inputs already show relatively high PSNR/SSIM, in some cases comparable to the baselines, while other metrics reveal poor perceptual quality. ID-PreFeR attains the lowest LPIPS [64], the best MANIQA [56], a competitive CLIPIQA [46] (second only to RestorerID [59], which trades fidelity and identity for no-reference

Table 1: Quantitative comparisons on CelebRef [31] of the proposed method with 4 blind face restoration methods (*i.e.*, GFPGAN [47], Authface [32], CodeFormer [67], and DiffBIR [33]) and 6 personalized/reference-based face restoration methods (*i.e.*, DMDNet [31], FaceMe [34], Gen2Res [11], InstantRestore [62], RestorerID [59], and ReF-LDM [19]). The specifications of reference images (Ref. spec.) are listed. ↑ (↓) denotes higher (lower) values indicate better performance. Numbers colored in red indicate the best performance, while blue for the second best. The metrics of the input and ground truth are for reference.

Method	Ref. spec.	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	ID↑
GFPGAN	×	25.50	0.748	0.301	0.503	0.449	0.613
CodeFormer	×	25.07	0.726	0.388	0.598	0.429	0.586
DiffBIR	×	25.29	0.690	0.401	0.615	0.504	0.562
AuthFace	×	25.32	0.683	0.294	0.699	0.643	0.649
DMDNet	HQ	25.04	0.712	0.316	0.613	0.491	0.635
FaceMe	LQ & HQ	25.82	0.724	0.265	0.646	0.504	0.704
Gen2Res	LQ & HQ	24.30	0.748	0.424	0.491	0.331	0.446
InstantRestore	HQ	24.65	0.722	0.297	0.561	0.488	0.769
RestorerID	HQ	22.72	0.636	0.434	0.717	0.574	0.522
ReF-LDM	HQ	23.58	0.676	0.244	0.669	0.562	0.695
Ours	LQ & HQ	25.51	0.729	0.242	0.701	0.649	0.767
Input	N/A	24.58	0.714	0.672	0.309	0.144	0.613
Ground truth	N/A	$+\infty$	1.000	0.000	0.650	0.546	1.000

quality), and a highly competitive ID score. Crucially, the only baseline with a marginally higher ID score, InstantRestore [62], and the other strong identity-preserving baselines (ReF-LDM [19], RestorerID [59]) all rely on HQ-only references, whereas ID-PreFeR reaches a comparable ID score with mixed LQ&HQ references, *i.e.*, the realistic setting this work targets. It is the only method ranking first or second on every metric, while baselines that lead one column (*e.g.*, FaceMe [34] on PSNR or RestorerID on CLIPIQA) drop sharply on others, showing that our gains do not trade one axis against another.

4.3 Evaluation on Additional Benchmarks

Larger reference benchmark. To verify that our conclusions are not specific to CelebRef, we additionally evaluate on the larger FFHQ-Ref-Moderate benchmark (857 subjects) [19], where ID-PreFeR improves identity similarity (IDS) and perceptual similarity (LPIPS/FID) by a clear margin over DMDNet [31] and ReF-LDM [19]. Full results are provided in the supplementary material.
Real-world mobile dataset. For our mobile-phone dataset, pixel-aligned ground truth is unavailable, so reference-based metrics that require pixel correspondence are not applicable. We therefore report no-reference quality metrics (LIQE [65], CLIPIQA [46], and MANIQA [56]) together with two complementary identity-similarity measures: the ArcFace [10] feature similarity against a held-out HQ

Table 2: Quantitative evaluation on the real-world mobile dataset. As paired ground truth is unavailable, we report no-reference quality metrics and two identity-similarity measures (ArcFace [10] feature similarity and a VLM-as-a-judge score, both higher is better). Red and blue denote the best and second-best.

Method	Reference-free quality↑			Identity similarity↑	
	LIQE	CLIPIQA	MANIQA	ArcFace	VLM
Gen2Res [11]	3.8491	0.6203	0.6375	0.5013	7.10
FaceMe [34]	4.5924	0.5503	0.5173	0.5750	7.62
InstantRestore [62]	4.0722	0.5072	0.4307	0.6093	6.84
RestorerID [59]	4.3060	0.7404	0.5594	0.5621	5.74
ReF-LDM [19]	4.8925	0.6757	0.5757	0.5969	8.86
Ours	4.8927	0.6850	0.6528	0.5972	8.92

reference, and a VLM-as-a-judge score that prompts Gemini-3-Pro to rate, on a 1–10 scale, how closely the facial identity in the restored image matches that reference (higher is more similar; full prompt in the supplementary material). We use both because ArcFace embeddings become unreliable under the heavy, mixed degradation of in-the-wild mobile captures, whereas the VLM judge assesses holistic resemblance and is less sensitive to low-level corruption; their agreement provides a more trustworthy identity signal than either alone. This is also why the judge is introduced only here and not in Tab. 1, where CelebRef's pixel-aligned HQ ground truth already makes the ArcFace cosine reliable. As shown in Tab. 2, ID-PreFeR ranks first or second on every metric, achieving the best LIQE, MANIQA, and VLM identity scores while remaining competitive on CLIPIQA and ArcFace similarity. To assess identity stability, we further measure the standard deviation of the per-sample ArcFace similarity across the mobile dataset, which serves as a proxy for long-term identity consistency. ID-PreFeR attains a low variance of 0.0192, on par with RestorerID [59] (0.0210), InstantRestore [62] (0.0198), and ReF-LDM [19] (0.0181), indicating robust identity preservation across diverse in-the-wild captures.

4.4 Qualitative Results

Since human perception is sensitive to facial details, we present qualitative comparisons to complement quantitative metrics. For an unbiased and informed comparison, we include cases of various skin colors. To keep the figures legible, the qualitative panels focus on the most competitive baselines; CodeFormer [67] and DiffBIR [33] are omitted given their weaker quantitative performance in Tab. 1, while the additional personalized baselines (InstantRestore [62], RestorerID [59], and ReF-LDM [19]) are compared quantitatively in Tabs. 1 and 2.

In Fig. 4, we present qualitative comparisons on synthetic CelebRef [31] with mixed HQ/LQ references (average #HQ=3), where each row shows the input, two reference-free baselines, the reference set, three personalized methods, our

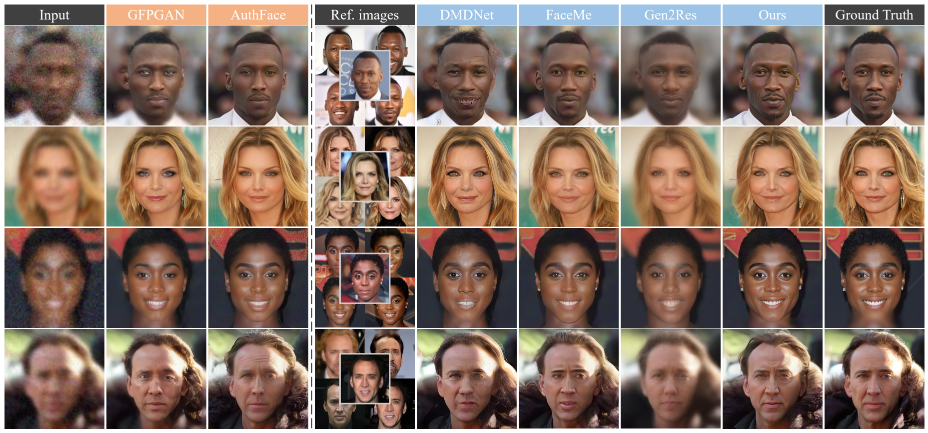

Fig. 4: Restoration results on synthetic degraded images, compared with reference-free face restoration baselines (orange titles) and personalized face restoration baselines (blue title, using a combination of HQ and LQ reference images).

result, and the ground truth. Reference-free methods (GFPGAN [47], AuthFace [32]) recover plausible textures but reshape identity-bearing landmarks (*e.g.*, jawline and lips) and wash out subject-specific details under strong degradation. Among personalized baselines, DMDNet [31] is geometrically unstable around mouth, ears, and hairline; FaceMe [34] introduces appearance inconsistencies under mixed quality and pose; and Gen2Res [11] overfits to reference texture and shifts identity-defining attributes. In contrast, our method is sharper and more consistent with the ground truth, better preserving facial geometry, hairline/ear placement, and fine-grained identity traits, without component alignment and while remaining robust to mixed-quality, pose-diverse references.

The top two rows of Fig. 5 show Internet images with diverse compression artifacts differing from the synthetic training distribution, and the bottom two rows show real-world mobile-phone frames captured in low-light, dynamic scenes with mixed noise, blur, low resolution, and compression. Reference-free methods (GFPGAN [47], AuthFace [32]) and Gen2Res [11] produce realistic faces but shift identity cues (*e.g.*, beard and eye shape), while DMDNet [31] and FaceMe [34] can fail under alignment difficulty or strong compression. In comparison, our method restores identity and subtle expressions more faithfully; the eyes and glasses are preserved even when the glasses are absent in some references. Additional qualitative comparisons on synthetic and Internet images are provided in the supplementary material.

4.5 Ablation Study

To validate the effectiveness of the proposed method, we conduct an ablation study on different components.

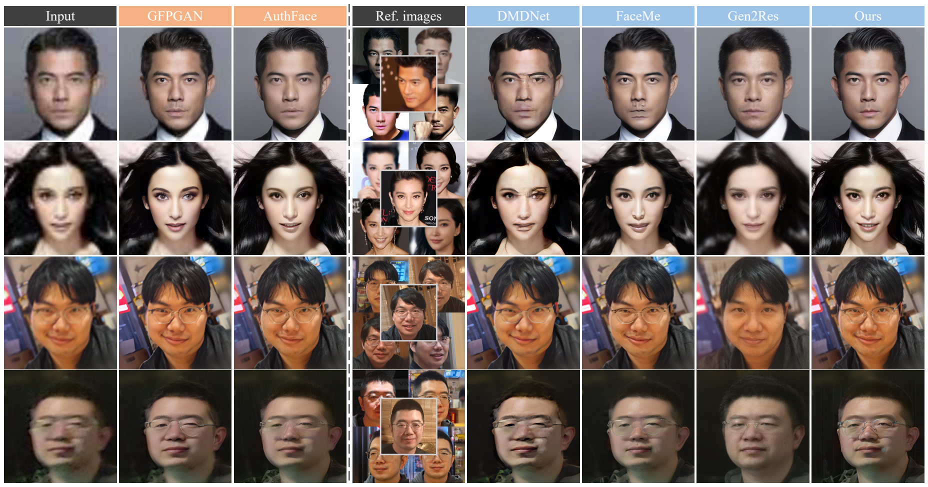

Fig. 5: Restoration results of degraded images from the Internet (top 2 rows) and captured by mobile phones (bottom 2 rows).

Table 3: Quantitative ablation study. Numbers in red indicate the best performance, while blue for the second best. "CA" and "SA" refers to the cross attention and self attention layers, respectively.

Method	#Params	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	ID↑
W/o prior preservation	6.41 M	25.52	0.731	0.245	0.699	0.624	0.746
W/o ID-quality disent.	6.41 M	25.57	0.698	0.279	0.674	0.555	0.766
W/o ID-preserving samp.	6.41 M	25.07	0.696	0.281	0.666	0.554	0.734
LoRA all CA	11.82 M	25.34	0.709	0.289	0.621	0.489	0.745
LoRA all CA & SA	12.57 M	25.45	0.720	0.290	0.599	0.472	0.756
Ours	6.41 M	25.51	0.729	0.242	0.701	0.649	0.767

As shown in the first row of Fig. 6, removing the latent prior loss causes color drift (often toward magenta) and more facial artifacts, especially with fewer HQ references; In the second row, removing quality cues and Min-SNR-γ produces visibly blurrier results under low-HQ settings; In the third row, removing ID-preserving sampling weakens salient identity cues (*e.g.*, iris color consistency).

These observations are corroborated by the quantitative comparisons in Tab. 3. The complete model uses only 6.41M parameters, far lighter than optimizing LoRA on all cross-attention (CA) or all attention (CA&SA) layers, yet performs stronger on most metrics. Since PSNR/SSIM do not always correlate with perceptual quality [64], the highest ID score is particularly important for validating identity preservation.

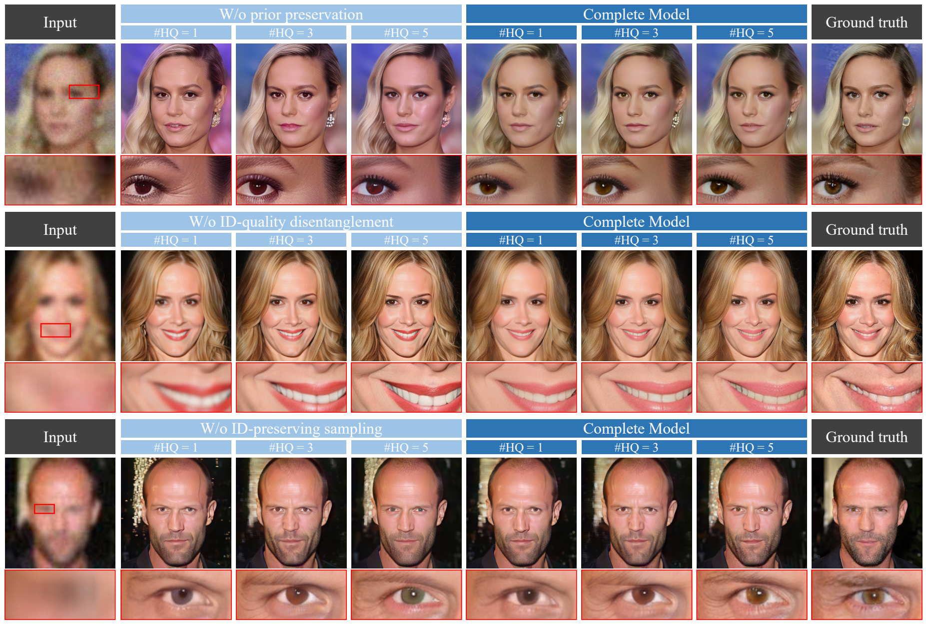

Fig. 6: Ablation study on the latent prior loss (first row), the ID-quality disentanglement strategy (second row), and the ID-preserving sampling (third row).

4.6 Robustness to Mixed Quality and Computational Cost

A central claim of ID-PreFeR is robustness to the number of high-quality references (#HQ). Tab. 4 sweeps #HQ$\in \{5, 3, 1\}$ while holding the reference-set size fixed and replacing the remaining references with degraded versions. ID-PreFeR maintains a nearly constant identity similarity (0.7671/0.7641/0.7613) as #HQ decreases. Removing ID-preserving sampling lowers identity similarity by a clear margin at every #HQ (to 0.7338/0.7325/0.7317), confirming that this component is the main driver of identity fidelity; removing ID-quality disentanglement leaves the ID score almost unchanged but, as the qualitative comparisons show, degrades perceptual quality and fine detail when HQ references are scarce, so both components contribute to mixed-quality robustness along complementary axes. Indeed, at the hardest #HQ=1 setting the disentanglement-free variant even edges out ours in ID (0.7632 *vs.* 0.7613), confirming that disentanglement buys perceptual quality under scarce HQ references rather than the ID score itself. The personalized baselines Gen2Res [11] and FaceMe [34] stay at a markedly lower identity level throughout.

We next analyze computational cost. Personalization is a one-time, per-identity procedure that trains only the LoRA injector (Sec. 3.2) with the backbone θ^* frozen; its 6.41M parameters ($\sim$0.25% of the SDXL backbone, $\sim$13 MB in FP16) are far lighter than storing a full diffusion model per identity, and personalization takes about 4.2 minutes, faster than the per-subject optimiza-

Table 4: Robustness to the number of HQ references (#HQ) and computational cost. ID↑ is reported under #HQ$\in \{5, 3, 1\}$. Training time is per identity; inference time and VRAM are per image. N/A denotes tuning-free methods without per-subject optimization. Red denotes the best value in each column.

Method	ID↑			Train	Inference	
	#HQ=5	#HQ=3	#HQ=1	time↓	time↓	VRAM↓
Gen2Res [11]	0.5576	0.5601	0.5676	11.4 min	8.9 s	3.26 GB
FaceMe [34]	0.5712	0.5722	0.5688	N/A	6.6 s	10.78 GB
Ours	0.7671	0.7641	0.7613	4.2 min	29.0 s	34.8 GB
w/o IDQ disent.	0.7664	0.7631	0.7632	4.2 min	29.2 s	34.8 GB
w/o IDPS	0.7338	0.7325	0.7317	4.3 min	5.6 s	24 GB

tion of Gen2Res. As the same table shows, the ID-preserving sampling accounts for most of the inference cost (*cf*. the "w/o IDPS" row): each refinement step in Eq. (7) decodes the intermediate latent through the VAE decoder $\mathcal{D}$ and backpropagates the cosine-similarity gradient through both $\mathcal{D}$ and the frozen face network f_{face}, so disabling it reduces inference from 29.0 s and 34.8 GB to 5.6 s and 24 GB. Restricting the refinement to the last few steps with K_{ID}=1 already recovers most of the identity gain, and reducing this cost for on-device deployment is a promising future direction.

5 Conclusion

In this work, we propose ID-PreFeR, a novel framework that personalizes the portrait restoration process, given a few reference images with mixed data quality. It disentangles content from varying image quality through joint optimization, enabling the model to focus solely on the shared identity information during finetuning. To support reproducibility, we will release the code, checkpoints, evaluation scripts, and the collected mobile dataset after acceptance.

Limitations. Variations across references in makeup, jewelry, and glasses may cause feature cancellation during training, as the model learns only the shared identity; ID-PreFeR may thus show diverse results in these attributes rather than adhering to any single reference. This can be partially mitigated by filtering the reference set, masking accessories and non-facial regions, or captioning the references so that such attributes are conditioned on explicitly rather than averaged away; we leave a thorough study to future work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 62136001, 62276007) and Beijing Major Science and Technology Project (Grant No. Z251100008125009).

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311 (2023)
2. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: Proc. of Computer Vision and Pattern Recognition (2023)
3. Bulat, A., Tzimiropoulos, G.: Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2018)
4. Chan, K.C., Xu, X., Wang, X., Gu, J., Loy, C.C.: GLEAN: Generative latent bank for image super-resolution and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3) (2022)
5. Chari, P., Ma, S., Ostashev, D., Kadambi, A., Krishnan, G., Wang, J., Aberman, K.: Personalized restoration via dual-pivot tuning. arXiv preprint arXiv:2312.17234 (2023)
6. Chen, C., Li, X., Yang, L., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: Proc. of Computer Vision and Pattern Recognition (2021)
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
8. 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)
9. Chesser, L., Carbone, T., Zahid, A.: Unsplash full dataset 1.2.2 (2020), unsplash.com/data, Last accessed on 2024-10-31
10. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: Proc. of Computer Vision and Pattern Recognition (2019)
11. Ding, Z., Zhang, X., Tu, Z., Xia, Z.: Restoration by generation with constrained priors. In: Proc. of Computer Vision and Pattern Recognition (2024)
12. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: Proc. of Computer Vision and Pattern Recognition Workshops (2019)
13. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proc. of Computer Vision and Pattern Recognition (2021)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
15. 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: Proc. of European Conference on Computer Vision (2022)
16. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-snr weighting strategy. In: Proc. of International Conference on Computer Vision (2023)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

18. Ho, M.M., Zhou, J.: Deep Preset: Blending and retouching photos with color style transfer. In: Proc. of Winter Conference on Applications of Computer Vision (2021)
19. Hsiao, C.W., Liu, Y.L., Yang, C.K., Kuo, S.P., Jou, K., Chen, C.P.: ReF-LDM: A latent diffusion model for reference-based face image restoration. In: Proc. of Advances in Neural Information Processing Systems (2024)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
21. Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., Menze, B., Liu, W.: Face super-resolution guided by 3D facial priors. In: Proc. of European Conference on Computer Vision (2020)
22. Hu, X., Ren, W., Yang, J., Cao, X., Wipf, D., Menze, B., Tong, X., Zha, H.: Face restoration via plug-and-play 3d facial priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12) (2021)
23. Hu, Y., Wang, Y., Zhang, J.: Dear-GAN: Degradation-aware face restoration with GAN prior. IEEE Transactions on Circuits and Systems for Video Technology **33**(9) (2023)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12) (2021)
25. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. arXiv preprint arXiv:1908.08239 (2019)
26. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proc. of Computer Vision and Pattern Recognition (2023)
27. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proc. of Computer Vision and Pattern Recognition (2017)
28. Lee, J., Cho, K., Kiela, D.: Countering language drift via visual grounding. In: Proc. of Conference on Empirical Methods in Natural Language Processing (2019)
29. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: Proc. of Computer Vision and Pattern Recognition (2020)
30. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proc. of European Conference on Computer Vision (2018)
31. Li, X., Zhang, S., Zhou, S., Zhang, L., Zuo, W.: Learning dual memory dictionaries for blind face restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
32. Liang, G., Fan, Q., Fu, B., Chen, J., Gu, H., Wang, L.: AuthFace: Towards authentic blind face restoration with face-oriented generative diffusion prior. arXiv preprint arXiv:2410.09864 (2024)
33. 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: Proc. of European Conference on Computer Vision (2024)
34. Liu, S., Duan, Z.P., OuYang, J., Fu, J., Park, H., Liu, Z., Guo, C., Li, C.: FaceMe: Robust blind face restoration with personal identification. In: Proc. of Association for the Advancement of Artificial Intelligence (2025)
35. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proc. of International Conference on Computer Vision (2015)

36. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: Proc. of Computer Vision and Pattern Recognition (2020)
37. Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., Cohen-Or, D.: Mystyle: A personalized generative prior. ACM Transactions on Graphics (2022)
38. 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)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of Computer Vision and Pattern Recognition (2022)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proc. of Computer Vision and Pattern Recognition (2023)
41. Shen, Y., Luo, P., Yan, J., Wang, X., Tang, X.: FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis. In: Proc. of Computer Vision and Pattern Recognition (2018)
42. Shen, Z., Lai, W.S., Xu, T., Kautz, J., Yang, M.H.: Exploiting semantics for face image deblurring. International Journal of Computer Vision **128**(7) (2020)
43. Varanka, T., Toivonen, T., Tripathy, S., Zhao, G., Acar, E.: Pfstorer: Personalized face restoration and super-resolution. In: Proc. of Computer Vision and Pattern Recognition (2024)
44. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
45. Wang, H., Teng, Z., Wu, C., Coleman, S.: Facial landmarks and generative priors guided blind face restoration. In: IEEE International Conference on Industrial Informatics (2022)
46. Wang, J., Chan, K.C., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: Proc. of Association for the Advancement of Artificial Intelligence (2023)
47. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proc. of Computer Vision and Pattern Recognition (2021)
48. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proc. of European Conference on Computer Vision (2018)
49. Wang, Y., Li, X., Xu, K., He, D., Zhang, Q., Li, F., Ding, E.: Neural color operators for sequential image retouching. In: Proc. of European Conference on Computer Vision (2022)
50. Wang, Z., Li, O., Wang, T., Wei, L., Hao, Y., Wang, X., Tian, Q.: Prior preserved text-to-image personalization without image regularization. IEEE Transactions on Circuits and Systems for Video Technology (2024)
51. 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: Proc. of Computer Vision and Pattern Recognition (2023)
52. Wang, Z., Zhang, J., Chen, R., Wang, W., Luo, P.: Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In: Proc. of Computer Vision and Pattern Recognition (2022)

53. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
54. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. arXiv preprint arXiv:2406.08177 (2024)
55. Yang, P., Zhou, S., Tao, Q., Loy, C.C.: PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In: Proc. of Advances in Neural Information Processing Systems (2024)
56. 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: Proc. of Computer Vision and Pattern Recognition (2022)
57. Yang, T., Ren, P., Xie, X., Zhang, L.: GAN prior embedded network for blind face restoration in the wild. In: Proc. of Computer Vision and Pattern Recognition (2021)
58. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
59. Ying, J., Liu, M., Wu, Z., Zhang, R., Yu, Z., Fu, S., Cao, S.Y., Wu, C., Yu, Y., Shen, H.L.: RestorerID: Towards tuning-free face restoration with id preservation. arXiv preprint arXiv:2411.14125 (2024)
60. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
61. Yue, Z., Loy, C.C.: Difface: Blind face restoration with diffused error contraction. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12) (2024)
62. Zhang, H., Alaluf, Y., Ma, S., Kadambi, A., Wang, J., Aberman, K.: InstantRestore: Single-step personalized face restoration with shared-image attention. In: Proc. of ACM SIGGRAPH (2025)
63. 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. pp. 3836–3847 (2023)
64. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of Computer Vision and Pattern Recognition (2018)
65. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proc. of Computer Vision and Pattern Recognition (2023)
66. Zhang, Y., Yang, M., Zhou, Q., Wang, Z.: Attention calibration for disentangled text-to-image personalization. In: Proc. of Computer Vision and Pattern Recognition (2024)
67. Zhou, S., Chan, K.C., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: Proc. of Advances in Neural Information Processing Systems (2022)
68. Zhu, F., Zhu, J., Chu, W., Zhang, X., Ji, X., Wang, C., Tai, Y.: Blind face restoration via integrating face shape and generative priors. In: Proc. of Computer Vision and Pattern Recognition (2022)

ID-PreFeR: ID-Preserving Face Restoration with Mixed Data Quality Supplementary Material

Chengxuan Zhu[1†], Yuchen Hong[1], Qi Zhang[2], Bingtao Fu[3], Jinxiu Liang[4], Jinwei Chen[2], Huaqi Zhang[2], Chao Xu[1], Boxin Shi[1*], Qingnan Fan[2*]

[1]Peking University [2]vivo BlueImage Lab
[3]Ant Group [4]National Institute of Informatics, Japan

A Justification for Min-SNR-γ Weighting

This section gives the formal argument behind the Min-SNR-γ weighted diffusion loss of the main paper: *why* this weighting prevents reference-image degradation from leaking into the identity injector. The key idea is to track how a degradation component propagates into the ϵ-prediction objective at each timestep, and to show that the weighting bounds its contribution.

Identity–degradation decomposition. Consider the K references of a single identity. Following the main paper, we decompose the latent $z_{\mathrm{LQ}}^{(i)}$ of the i-th reference into an *identity-consistent* component z_{HQ}, shared across all references of that identity, and a *sample-specific* degradation residual $\epsilon_{\mathrm{deg}}^{(i)}$:

$$z_{\mathrm{LQ}}^{(i)} = z_{\mathrm{HQ}} + \epsilon_{\mathrm{deg}}^{(i)}. \tag{9}$$

Here z_{HQ} is the component the injector is intended to encode, whereas $\epsilon_{\mathrm{deg}}^{(i)}$ is a per-sample nuisance that should *not* be attributed to the identity. Substituting Eq. (9) into the forward process gives the diffused latent

$$z_t^{(i)} = \sqrt{\bar{\alpha}_t}\, z_{\mathrm{HQ}} + \sqrt{1-\bar{\alpha}_t}\,\big(\epsilon + \omega_t\, \epsilon_{\mathrm{deg}}^{(i)}\big), \qquad \omega_t^2 = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} = \mathrm{SNR}(t), \tag{10}$$

so that, in the ϵ-prediction parameterization, the degradation manifests as an additive residual $\Delta\epsilon_t = \omega_t\, \epsilon_{\mathrm{deg}}^{(i)} = \sqrt{\mathrm{SNR}(t)}\, \epsilon_{\mathrm{deg}}^{(i)}$ in the regression target.

* Corresponding authors. † Work done during their internship at vivo.
[1]Chengxuan Zhu and Chao Xu are with National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. Yuchen Hong and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

The degradation term grows at high SNR. The (unweighted) contribution of this residual to the diffusion loss therefore scales as $\|\Delta\epsilon_t\|_2^2 = \mathrm{SNR}(t)\,\|\epsilon_{\mathrm{deg}}^{(i)}\|_2^2$, which *grows* with $\mathrm{SNR}(t)$ and is largest at low-noise (high-SNR) timesteps. Consequently, without re-weighting, these high-SNR steps dominate the gradient and push the injector to reproduce the per-sample degradation $\epsilon_{\mathrm{deg}}^{(i)}$ as if it were part of the identity.

Min-SNR-γ caps the degradation gradient. Min-SNR-γ [1] applies the weight $w_t = \min\big(\gamma/\mathrm{SNR}(t),\, 1\big)$, which is exactly the factor $\min(\gamma\,\omega_t^{-2}, 1)$ used in the main paper's weighted loss. Its effect on the degradation contribution is to bound it by a constant independent of t:

$$\begin{aligned} w_t\,\big\|\omega_t\,\epsilon_{\mathrm{deg}}^{(i)}\big\|_2^2 &= \min\Big(\tfrac{\gamma}{\mathrm{SNR}(t)},\, 1\Big)\,\mathrm{SNR}(t)\,\big\|\epsilon_{\mathrm{deg}}^{(i)}\big\|_2^2 \\ &= \min\big(\gamma,\, \mathrm{SNR}(t)\big)\,\big\|\epsilon_{\mathrm{deg}}^{(i)}\big\|_2^2 \;\le\; \gamma\,\big\|\epsilon_{\mathrm{deg}}^{(i)}\big\|_2^2, \end{aligned} \tag{11}$$

where the first equality uses $\omega_t^2 = \mathrm{SNR}(t)$. At high SNR ($\mathrm{SNR}(t) > \gamma$) the weight $\gamma/\mathrm{SNR}(t)$ exactly cancels the $\mathrm{SNR}(t)$ amplification and clamps the term to $\gamma\,\|\epsilon_{\mathrm{deg}}^{(i)}\|_2^2$; at low SNR ($\mathrm{SNR}(t) \le \gamma$) the weight is 1 and the term is already small. The weight scales the whole residual at a given timestep, not the degradation term alone; identity survives this because the shared component z_{HQ} accumulates coherently across the K references, whereas the capped per-sample residuals $\epsilon_{\mathrm{deg}}^{(i)}$ are uncorrelated and cannot. Min-SNR-γ thus bounds how much any single (high-SNR) timestep drives the injector toward the degradation residual, keeping quality-specific artifacts out of the learned identity and complementing the quality-token disentanglement in the main paper.

B Full Results on FFHQ-Ref-Moderate

To verify that our conclusions are not specific to CelebRef, we additionally evaluate on the larger FFHQ-Ref-Moderate benchmark (857 subjects) proposed by ReF-LDM [2]. As reported in Tab. 5, where † denotes the numbers reported by ReF-LDM [2], ID-PreFeR improves identity similarity (IDS) and overall perceptual similarity (LPIPS) by a clear margin. On the tightly cropped facial area (fLPIPS) it trails the two baselines slightly (0.105 *vs*. 0.088/0.096), as our gains concentrate on whole-image restoration and identity rather than on the cropped-patch metric. The restored images are also better aligned with the natural HQ image distribution, as reflected by the lowest FID. A qualitative example from this benchmark is shown in Fig. 7: ID-PreFeR yields the closest appearance to the ground truth, especially when restoring the subject's eyes and mouth. Assessing fairness across demographic subgroups on the more demographically balanced FairFace dataset [3] is a natural next step for a finer-grained subgroup analysis.

Table 5: Quantitative comparison on the larger FFHQ-Ref-Moderate benchmark [2] (857 subjects). † denotes results reported by ReF-LDM [2]. Red and blue denote the best and second-best.

Method	IDS↑	fLPIPS↓	LPIPS↓	FID↓
DMDNet† [4]	0.810	0.096	0.348	36.60
ReF-LDM† [2]	0.840	0.088	0.332	33.05
Ours	0.891	0.105	0.202	32.63

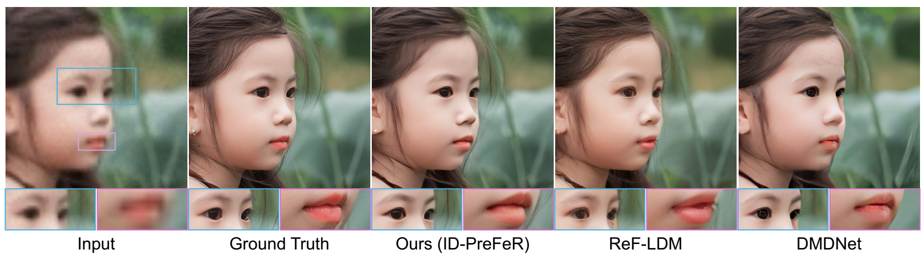

Fig. 7: Qualitative comparison on the FFHQ-Ref-Moderate benchmark [2] (reference images omitted). ID-PreFeR restores identity-bearing details, especially around the eyes and mouth, most faithfully to the ground truth.

C VLM-as-a-Judge Identity Score

The *VLM* column of Tab. 2 uses **Gemini-3-Pro** as the judge. For each restored image it is shown the restored face and the held-out HQ reference of the same subject and returns an integer 1–10 facial-identity score with the exact prompt below:

```
You are an expert face-identity verifier. You are given two
images of a face: (1) a restored image and (2) a high-quality
reference image of the same person. Rate, on an integer scale
from 1 to 10, how likely the two images show the SAME identity,
where 1 = clearly a different person and 10 = definitely the
same person. Judge only identity-bearing facial structure (face
shape/contour, eyes and brows, nose and mouth geometry, and
other person-specific traits). Ignore differences in pose,
expression, lighting, background, and image quality. Output
only the integer score.
```

The same prompt and image order are used for every method, and the reported value is the mean over the mobile dataset.

D More Results

Fig. 8 compares the performance of the baselines and ID-PreFeR on synthetic scenes, where the proposed method achieves the restoration closest to the ground truth. Compared with the baselines, which often generate visually plausible yet slightly identity-drifting faces, our method better preserves identity-bearing facial geometry and fine local traits, yielding restorations that are consistently closer to the ground truth in face contour, eye–brow configuration, nose–mouth structure, and age-related texture.

Fig. 9 presents Internet photographs of celebrities that have undergone unknown mixed compression. The proposed method restores their recognizable facial identity. Compared with the baselines, which often generate plausible yet slightly identity-drifting faces, our method better preserves identity-bearing facial geometry and fine local traits, yielding restorations that are consistently closer to the ground truth in face contour, eye–brow configuration, nose–mouth structure, and age-related texture.

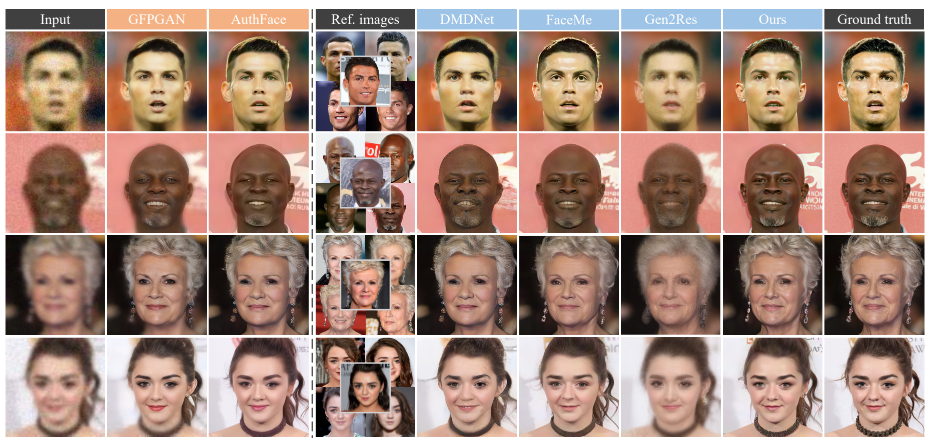

Fig. 8: More qualitative comparisons on ID-preserving face restoration (ground truth undergone synthetic degradation).

E Ethical Impact

E.1 Demographic Fairness

Face restoration systems are highly sensitive to visual identity cues that vary with ethnicity, age, and other demographic factors. Without careful consideration, such systems may show uneven performance across population subgroups. Note that ID-PreFeR is designed to focus on identity-preserving features while minimizing dependence on lighting and skin-tone–related signals, as validated in the diverse qualitative cases throughout the paper. It is able to maintain consistent restoration quality across ethnic groups. However, small residual disparities remain, indicating the need for continued research in this area.

Fig. 9: More qualitative comparisons on ID-preserving face restoration (Internet images with compression and various degradations).

E.2 Privacy and Misuse Concerns

Technical advances are often accompanied by heightened concerns about misuse. The capacity to reconstruct high-fidelity facial images from degraded inputs introduces potential misuse risks. In particular, applications involving surveillance, de-anonymization, or forensic prediction without consent are explicitly discouraged. We urge users to adhere to legal frameworks and ethical standards, especially in contexts involving vulnerable or marginalized populations.

E.3 Transparency and Accountability

We support responsible deployment by releasing model evaluation protocols and encouraging third-party audits on fairness and generalization. Further, we advocate for the development of community benchmarks that better capture demographic variability and downstream societal effects.

References

1. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-snr weighting strategy. In: Proc. of International Conference on Computer Vision (2023)
2. Hsiao, C.W., Liu, Y.L., Yang, C.K., Kuo, S.P., Jou, K., Chen, C.P.: ReF-LDM: A latent diffusion model for reference-based face image restoration. In: Proc. of Advances in Neural Information Processing Systems (2024)
3. Kärkkäinen, K., Joo, J.: FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proc. of Winter Conference on Applications of Computer Vision (2021)
4. Li, X., Zhang, S., Zhou, S., Zhang, L., Zuo, W.: Learning dual memory dictionaries for blind face restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)