# REFROUTEFR: DEMAND-DRIVEN SPARSE REFERENCE-TOKEN ROUTING FOR ONE-STEP FACE RESTORATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reference-based face restoration uses high-quality images of the same identity to recover personalized details from degraded observations. However, reference images exhibit substantial input-dependent redundancy: only a subset of reference regions is useful for a given low-quality (LQ) input, whereas dense processing wastes computation and may introduce mismatched textures or attributes. We propose **RefRouteFR**, a one-step framework centered on demand-driven sparse reference-token routing. RefRouteFR exploits the native multimodal interface of a pretrained Diffusion Transformer to incorporate reference guidance and text control without requiring auxiliary reference-injection networks. A lightweight router estimates the restoration-dependent usefulness of reference tokens from the current multimodal context and forwards only the most informative subset through each Transformer block, while redundant tokens bypass expensive computation. To enable stable one-step restoration, a learned latent calibrator mitigates the mismatch between degraded LQ latents and the pretrained flow prior. Experiments on multiple datasets demonstrate state-of-the-art restoration performance, with RefRouteFR matching the quality of its dense counterpart while reducing inference time by approximately 40%.

## 1 INTRODUCTION

Face restoration aims to reconstruct a high-quality (HQ) face image from a low-quality (LQ) observation affected by blur, noise, compression, and other unknown degradations. Unlike generic image restoration, face restoration is identity-critical: even a visually plausible result can be undesirable if the restored face no longer corresponds to the input person. Recent blind face restoration methods improve realism and robustness by introducing facial or generative priors (Gao et al., 2024; Kuai et al., 2025; Liang et al., 2025; Zhang et al., 2025b; Wu et al., 2025a; Do et al., 2025). However, these methods can only infer missing identity-specific details from the degraded input itself, often failing to recover personalized textures. Reference-based face restoration addresses this limitation by using one or more HQ images of the same person as external identity guidance (Hsiao et al., 2024; Zhang et al., 2025a; Ying et al., 2024; Zhou et al., 2026), which helps reduce identity hallucination and recover subject-specific details.

However, reference-based face restoration still faces a trade-off among identity fidelity, robustness, and efficiency. References are beneficial only when they match the restoration demand of the LQ input; otherwise, irrelevant content may introduce mismatched textures or attributes. Existing methods alleviate this issue through reference selection or adaptive conditioning (Yin et al., 2026; Wang et al., 2026b; Liu et al., 2025c), but often require auxiliary modules and dense feature extraction over entire reference images, limiting the number of references or increasing inference latency in multi-reference settings. Meanwhile, diffusion-based methods leverage strong generative priors for perceptual realism (Lin et al., 2024; Chen et al., 2024b; Yang et al., 2023; Li et al., 2025b; Miao et al., 2025), yet multi-step sampling incurs high inference cost and may introduce over-smoothed textures, semantic mismatch, or color artifacts (Wu et al., 2024a; Zheng et al., 2024; Wang et al., 2024; Ren et al., 2024). Therefore, an ideal restorer should exploit strong generative priors while preserving identity-specific information and avoiding redundant reference computation.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

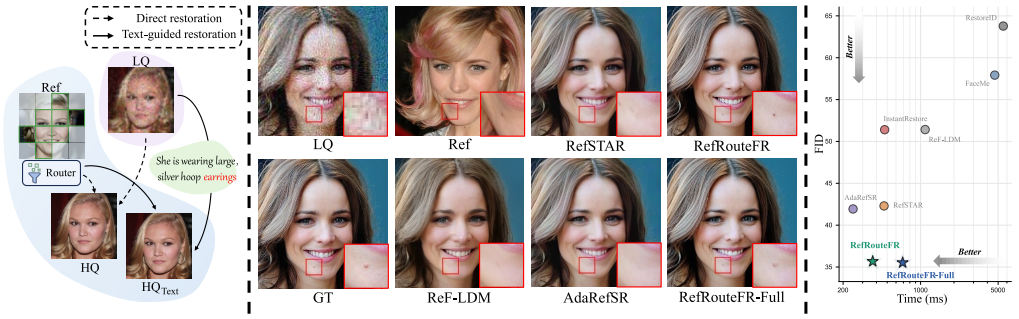


Figure 1: **Left**: reference routing and text-guided attribute control. Given an LQ input and same-identity references, the router selects informative reference regions for restoration, while the text prompt further controls specific facial attributes. **Middle**: visual comparison with existing face restoration methods, showing that RefRouteFR restores more faithful identity-related details and natural local textures. **Right**: FID–Time trade-off, where points closer to the lower-left corner indicate better performance. RefRouteFR-Full achieves the best FID, while RefRouteFR retains comparable quality with approximately 40% lower inference time through sparse reference-token routing.

In this paper, we propose **RefRouteFR**, an efficient reference-based face restoration framework built on a pretrained FLUX2 Diffusion Transformer (DiT) (Peebles & Xie, 2023; Black Forest Labs et al., 2025). Rather than introducing a dedicated reference-fusion branch, RefRouteFR represents the LQ image, reference images, and text prompt as a unified multimodal token sequence, thereby preserving prompt-based control over facial attributes. We introduce a lightweight router that selects reference tokens according to the current LQ–reference–text context. Unlike image-level reference selection or dense reference-feature processing, the router forwards only the reference tokens relevant to the current restoration demand, while the remaining tokens bypass expensive Transformer computation, avoiding redundant computation and unnecessary processing of weakly relevant reference regions. With a fixed routing budget, additional references mainly increase the lightweight routing cost rather than proportionally increasing backbone computation. We denote the counterpart that retains all reference tokens as RefRouteFR-Full. As illustrated in Fig. 1, RefRouteFR scales more efficiently to multiple references and preserves restoration quality close to RefRouteFR-Full while reducing inference time by approximately 40%.

To avoid the temporal overhead of iterative diffusion sampling, RefRouteFR formulates restoration as a single generative forward pass initialized directly from the LQ latent rather than random noise. Existing one-step restoration methods initialize from LQ latents directly or through prescribed noise injection (Wu et al., 2024b; Sun et al., 2025; Fang et al., 2026). However, unknown degradations may place this task-specific initialization off the flow-matching trajectory learned by the pretrained model (Lipman et al., 2023). We therefore learn an input-dependent latent correction to provide a suitable starting point for one-step face restoration. This design leverages the pretrained generative prior while eliminating iterative denoising.

Overall, our contributions are threefold:

- **Demand-driven sparse reference-token routing.** We identify input-dependent redundancy in reference tokens and propose a demand-driven sparse router. Under a fixed token budget, it limits the processing of weakly relevant reference tokens and prevents expensive backbone computation from growing proportionally with the number of references.
- **One-step restoration via calibrated flow initialization.** We formulate reference-based face restoration as a single flow update initialized from the LQ latent. A latent calibrator reduces the mismatch between degraded latents and the pretrained flow trajectory, eliminating iterative denoising while retaining strong generative priors.
- **Unified multimodal modeling for multi-reference and text-controllable restoration.** RefRouteFR naturally supports multi-reference and text-controllable restoration through unified token-space modeling. Experiments on multiple datasets demonstrate superior identity fidelity and perceptual quality with a favorable efficiency–quality trade-off.

# 2 Related Works

## 2.1 Reference-based Face Restoration

Reference-based face restoration leverages additional facial images of the same identity to recover personalized details that are difficult to infer from degraded inputs. Early methods introduce reference information through explicit warping of guidance images or external facial component dictionaries (Li et al., 2018; 2020), but their limited feature matching ability makes it difficult to preserve fine identity-specific textures. Recent approaches build on diffusion models and inject reference information through identity-aware adapters or cross-image attention (Hsiao et al., 2024; Ying et al., 2024; Zhou et al., 2026; Zhang et al., 2025a). Some methods further introduce explicit reference selection or adaptive conditioning modules to suppress harmful reference regions (Yin et al., 2026; Wang et al., 2026b). However, these designs usually require auxiliary modules and dense reference feature processing. In contrast, RefRouteFR integrates demand-driven routing directly into the pretrained backbone, selecting and processing only reference tokens relevant to the LQ input.

## 2.2 Efficient Diffusion Models

Diffusion models provide strong generative priors but incur high inference cost due to iterative denoising. Recent one-step restoration methods improve efficiency by replacing random-noise initialization with an LQ image or latent starting point, and compressing multi-step trajectories into a single restoration step (Dong et al., 2025; Wang et al., 2025; Li et al., 2025a). With the rise of Diffusion Transformers, another line of work reduces internal computational redundancy through feature reuse or token reduction. Caching-based methods reuse intermediate activations or attention outputs across denoising steps (Chen et al., 2024a; Selvaraju et al., 2024; Zhao et al., 2025; Liu et al., 2025b;a), while token reduction methods prune or merge redundant visual tokens to accelerate block-wise computation (Bolya & Hoffman, 2023; Lu et al., 2025; Wu et al., 2025b). These methods mainly target general image generation or non-reference restoration by reducing denoising steps or redundant visual tokens. They do not explicitly address reference-induced redundancy in multi-reference face restoration, where many reference regions may be irrelevant to the degraded LQ input but still incur substantial Transformer computation if processed densely. RefRouteFR fills this gap by combining one-step restoration with sparse reference-token routing, reducing both the temporal sampling cost and the block-wise computation caused by redundant reference tokens.

# 3 Methods

## 3.1 Preliminaries

**Notation and problem formulation.** Let $x_{\mathrm{LQ}}$ denote the degraded low-quality face image, $x_{\mathrm{HQ}}$ denote the target high-quality face image, and $\{x^i_{\mathrm{ref}}\}_{i=1}^n$ denote $n$ high-quality reference images of the same identity as $x_{\mathrm{LQ}}$. The text prompt is denoted by $p$, and its encoded text tokens are denoted by $\mathbf{c}_p$. We use $\mathcal{E}$ and $\mathcal{D}$ for the VAE encoder and decoder, respectively. The corresponding latents are $z_{\mathrm{LQ}} = \mathcal{E}(x_{\mathrm{LQ}})$, $z_{\mathrm{HQ}} = \mathcal{E}(x_{\mathrm{HQ}})$ and $z^i_{\mathrm{ref}} = \mathcal{E}(x^i_{\mathrm{ref}})$. The operators $\mathcal{P}$ and $\mathcal{P}^{-1}$ convert between spatial latents and image-token sequences.

**Flow matching.** Flow matching learns a continuous transport between a simple noise distribution and the data distribution by regressing the velocity field along a probability path (Lipman et al., 2023). Given a clean latent $z_0$, a Gaussian noise latent $z_1 \sim \mathcal{N}(0, I)$, and a time variable $t \in [0, 1]$, a commonly used linear path is defined as

$$z_t = (1 - t)z_0 + tz_1, \tag{1}$$

whose target velocity is $u_t = z_1 - z_0$. A conditional flow model $v_\theta(z_t, t, c)$ is trained to predict this velocity under condition $c$ by minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathrm{E}_{z_0,z_1,t}\left[\|v_\theta(z_t, t, c) - (z_1 - z_0)\|_2^2\right]. \tag{2}$$

At inference time, samples are generated by solving the corresponding ODE from the noise endpoint to the data endpoint. The pretrained FLUX2 model provides such a pretrained flow prior in latent

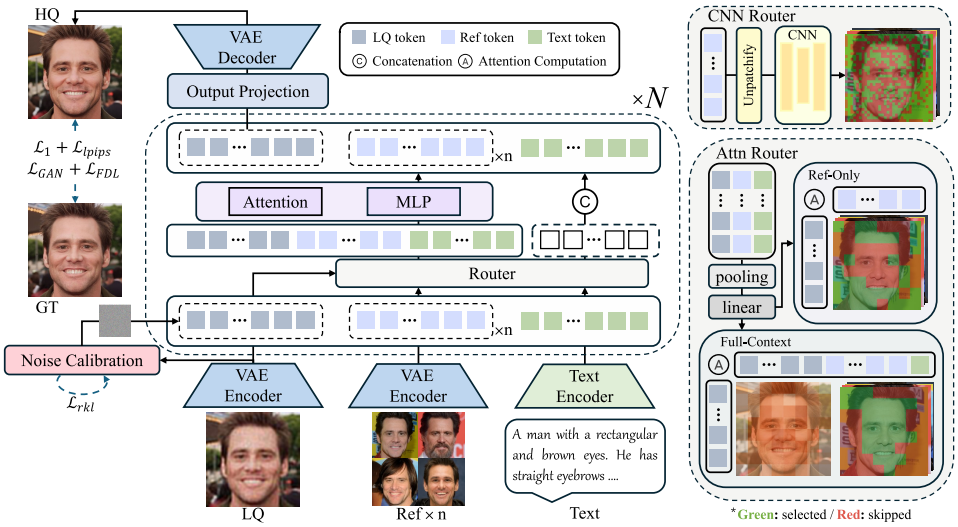


Figure 2: Architecture of RefRouteFR. The LQ image, multiple reference images, and text prompt are encoded into LQ, reference, and text tokens, respectively. The LQ latent is first calibrated and then concatenated with reference and text tokens for a single forward pass. The model predicts a restoration update for the LQ token branch, which is projected back to the latent space and decoded into the restored HQ image. The right inset illustrates alternative routing designs, where green regions are selected and red regions are skipped.

space, which serves as the generative backbone of our one-step restoration formulation. For restoration, the LQ latent is more informative than Gaussian noise because it preserves facial structure and identity cues; however, it may still deviate from the pretrained flow trajectory due to degradations. We therefore calibrate the LQ starting point before the single-step flow update to reduce this trajectory mismatch.

## 3.2 Model Architecture

### 3.2.1 Overview

RefRouteFR formulates reference-based face restoration as one forward pass in the native multimodal token space. As shown in Fig. 2, the LQ latent is used as the restoration branch, while the reference latents and prompt tokens act as conditions.

The overall inference process is summarized as follows:

$$\begin{aligned} \tilde{z}_{\mathrm{LQ}} &= z_{\mathrm{LQ}} + g_\phi(z_{\mathrm{LQ}}, t_0), \\ \mathbf{H}_0 &= \left[\mathcal{P}(\tilde{z}_{\mathrm{LQ}}); \mathcal{P}(z^1_{\mathrm{ref}}); \cdots; \mathcal{P}(z^n_{\mathrm{ref}})\right]. \end{aligned} \tag{3}$$

Here, $g_\phi$ predicts a correction to the LQ starting latent conditioned on $t_0$. The DiT backbone $F_\theta$ predicts a direct restoration update over the multimodal image-token sequence, and $\Pi_{\mathrm{LQ}}$ selects the LQ-token portion:

$$\hat{\mathbf{v}}_{\mathrm{LQ}} = \Pi_{\mathrm{LQ}} F_\theta(\mathbf{H}_0, \mathbf{c}_p, t_0). \tag{4}$$

The restored image is obtained with one update followed by VAE decoding:

$$\hat{x}_{\mathrm{HQ}} = \mathcal{D}\left(\mathcal{P}^{-1}\left(\mathcal{P}(\tilde{z}_{\mathrm{LQ}}) - \hat{\mathbf{v}}_{\mathrm{LQ}}\right)\right). \tag{5}$$

This formulation enables one-step restoration and avoids a hand-crafted reference injection branch, since all conditions are represented through the native token interface. Compared with iterative diffusion samplers, this single-update formulation removes the latency of multi-step denoising and keeps the restored latent close to the pretrained flow trajectory, retaining the strong generative prior of the pretrained model at the cost of a single forward pass.

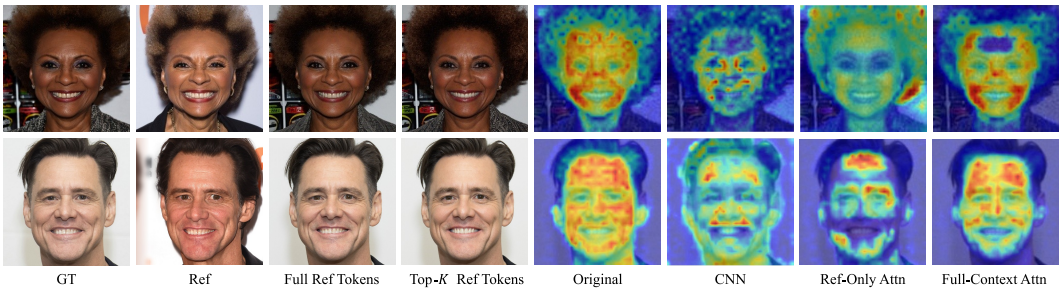


Figure 3: Analysis of reference-token redundancy and router attention behavior, shown for two examples. The ground-truth and reference images are followed by RefRouteFR-Full outputs using either full reference tokens or only the Top-$K$ reference tokens ranked by LQ attention, and by LQ-to-reference attention maps from the corresponding Transformer block of RefRouteFR-Full ("Original") and three router variants.

### 3.2.2 Demand-Driven Sparse Reference Token Routing

**Reference Redundancy.** The reference branch contains substantial token redundancy during restoration. To verify this, we perform a post-hoc sparse inference with RefRouteFR-Full by retaining only the Top-$K$ reference tokens receiving the highest LQ attention in each Transformer block. As shown in Fig. 3, despite this pruning, the result remains close to the full-reference output, indicating that reference-based face restoration depends on a small set of relevant reference regions rather than densely processing all reference tokens in every block.

**Router Design.** We therefore study three router designs, as illustrated in Fig. 2. The *CNN Router* predicts token importance from spatial feature maps using lightweight convolutional layers. Although efficient, it lacks explicit LQ–reference correspondence modeling, making its selection weakly conditioned on restoration demand. The *Reference-Only Attention Router* introduces this correspondence, but normalizes attention only over reference tokens. However, since each LQ token can only attend to reference tokens, the router is forced to assign attention to the reference branch even when reference guidance is unnecessary, often favoring large or visually dominant regions such as hair, forehead or jaw.

To address this issue, we propose the *Full-Context Attention Router*, which lets LQ tokens compare reference information with LQ self-context and text context. Before the $l$-th single-stream block, let $\mathbf{H}^l_{\mathrm{LQ}}$, $\mathbf{H}^l_{\mathrm{ref}}$ and $\mathbf{H}^l_{\mathrm{text}}$ denote the LQ, reference and text token states. We first aggregate image tokens into patch tokens using $4 \times 4$ average pooling:

$$\bar{\mathbf{H}}^l_{\mathrm{LQ}} = \mathrm{Pool}(\mathbf{H}^l_{\mathrm{LQ}}), \quad \bar{\mathbf{H}}^l_{\mathrm{ref}} = \mathrm{Pool}(\mathbf{H}^l_{\mathrm{ref}}). \tag{6}$$

The text tokens are compressed into one pooled text key. We then form LQ queries and full keys as:

$$\mathbf{Q}_{\mathrm{LQ}} = \psi_q(\bar{\mathbf{H}}^l_{\mathrm{LQ}}), \qquad \mathbf{K}_{\mathrm{Full}} = \begin{bmatrix} \psi_k(\bar{\mathbf{H}}^l_{\mathrm{LQ}}); \\ \psi_k(\bar{\mathbf{H}}^l_{\mathrm{ref}}); \\ \psi_k(\bar{\mathbf{H}}^l_{\mathrm{text}}) \end{bmatrix}, \tag{7}$$

where $\psi_q$ and $\psi_k$ are lightweight projections. The router attention is computed by:

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{\mathrm{LQ}}\mathbf{K}^\top_{\mathrm{Full}}}{\sqrt{d_r}}\right). \tag{8}$$

Reference-patch scores are obtained by averaging the attention assigned to each reference patch across all LQ queries. We retain the Top-$K$ patches and expand their mask back to token resolution. As shown in Fig. 3, the variant using Full-Context Attention Router produces an attention pattern closer to RefRouteFR-Full while avoiding diffuse responses to visually dominant salient regions. This indicates that providing non-reference alternatives makes reference selection demand-driven rather than compulsory. Moreover, with a fixed $K$, the expensive Transformer blocks process a constant number of reference tokens; additional references mainly increase the lightweight routing cost instead of proportionally scaling the backbone computation.

**Straight-Through Top-$K$ Routing.** Since hard Top-$K$ selection is non-differentiable with respect to the routing scores, we adopt a straight-through estimator (STE) (Bengio et al., 2013) for joint optimization. Given routing scores $\mathbf{S}$, we first select Top-$K$ indices and construct a hard mask $\mathbf{M}_{\text{hard}}$. The differentiable routing mask is defined as:

$$\mathbf{M} = \mathbf{M}_{\text{hard}} - \text{sg}(\mathbf{S}) + \mathbf{S}, \tag{9}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. During the forward pass, $\mathbf{M}$ is equivalent to the hard Top-$K$ mask, preserving sparse execution. During the backward pass, gradients are passed through the continuous scores $\mathbf{S}$, allowing the router to learn which reference tokens should be selected.

### 3.3 One-Step Restoration with Latent Calibration

#### 3.3.1 Latent Calibration

Although the LQ latent is more informative than Gaussian noise, as exploited by recent one-step restoration methods that directly initialize from degraded observations (Wang et al., 2025; Wu et al., 2024b), unknown degradations may make it a suboptimal starting point for one-step restoration. Rather than directly using the LQ latent, we introduce a lightweight calibrator $g_\phi$ that predicts an input-dependent latent correction before the one-step update:

$$\begin{aligned} \Delta z_{\text{LQ}} &= g_\phi(z_{\text{LQ}}, t_0), \\ \tilde{z}_{\text{LQ}} &= z_{\text{LQ}} + \Delta z_{\text{LQ}}. \end{aligned} \tag{10}$$

The calibrator does not generate the HQ latent directly. Instead, it adjusts the restoration starting point so that the subsequent one-step update can better follow the pretrained flow prior.

To constrain the correction to a plausible restoration direction, we regularize the latent calibrator with residual-guided KL regularization. If the calibrator is trained only through reconstruction losses, $\Delta z_{\text{LQ}}$ may absorb image-generation details rather than remain a starting-point correction. In practice, the latent calibrator outputs a diagonal Gaussian distribution:

$$q_\phi(\Delta z \mid z_{\text{LQ}}, t_0) = \mathcal{N}\left(\mu_\phi, \text{diag}(\sigma_\phi^2)\right), \tag{11}$$

and we align it with a unit-variance Gaussian centered at the HQ–LQ latent residual. This residual-guided KL loss, denoted as $\mathcal{L}_{\text{rKL}}$, is defined as

$$\mathcal{L}_{\text{rKL}} = D_{\text{KL}}\left(q_\phi(\Delta z \mid z_{\text{LQ}}, t_0) \,\|\, \mathcal{N}(z_{\text{HQ}} - z_{\text{LQ}}, I)\right). \tag{12}$$

This loss encourages the predicted correction distribution to match the residual direction from the degraded latent to the clean latent, while discouraging unnecessary latent shifts. The calibrated latent $\tilde{z}_{\text{LQ}}$ is then used as the starting point for the one-step restoration update described in the overview.

#### 3.3.2 Training Objective

Training a one-step DiT restorer requires balancing reconstruction fidelity, perceptual realism, and trajectory alignment. Pixel-wise supervision alone tends to preserve coarse facial structure but often produces over-smoothed results or high-frequency periodic artifacts. We therefore optimize RefRouteFR with a multi-term generator objective. The reconstruction loss $\mathcal{L}_1$ preserves low-level fidelity, while $\mathcal{L}_{\text{LPIPS}}$ and the generator-side adversarial loss $\mathcal{L}_{\text{GAN}}$ encourage perceptual realism. We further adopt the frequency distribution loss (FDL) (Ni et al., 2024) to constrain frequency statistics:

$$\begin{aligned} \mathcal{L}_{\text{FDL}} = \;& \text{SW}\left(\mathcal{A}_{\Phi(\hat{x}_{\text{HQ}})}, \mathcal{A}_{\Phi(x_{\text{HQ}})}\right) \\ & + \lambda_{\text{phase}} \text{SW}\left(\mathcal{S}_{\Phi(\hat{x}_{\text{HQ}})}, \mathcal{S}_{\Phi(x_{\text{HQ}})}\right), \end{aligned} \tag{13}$$

where $\mathcal{A}$ and $\mathcal{S}$ denote the amplitude and phase components, respectively. This term discourages periodic texture artifacts and complements the image-space perceptual losses. Finally, $\mathcal{L}_{\text{rKL}}$ is applied only to the latent calibrator as defined in Latent Calibration. The overall generator objective is

$$\begin{aligned} \mathcal{L}_{\text{G}} = \;& \lambda_1 \mathcal{L}_1 + \lambda_{\text{lpips}} \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{gan}} \mathcal{L}_{\text{GAN}} \\ & + \lambda_{\text{fdl}} \mathcal{L}_{\text{FDL}} + \lambda_{\text{rKL}} \mathcal{L}_{\text{rKL}}. \end{aligned} \tag{14}$$

Table 1: Quantitative comparison with state-of-the-art methods on CelebRef-HQ and FFHQ-Ref. Best and second-best results are highlighted in bold and underlined, respectively. Methods marked with * support multiple reference images.

| Dataset | Method | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | IDS↑ | NIQE↓ | MUSIQ↑ | CLIP-IQA↑ |
|---|---|---|---|---|---|---|---|---|---|
| CelebRef-HQ | CodeFormer | 26.1161 | 0.7010 | 0.2389 | 53.5464 | 0.8135 | 4.4419 | **75.4505** | 0.6959 |
| | DMDNet* | 26.6770 | 0.7425 | 0.2440 | 55.6650 | 0.8024 | 4.6402 | 71.4155 | 0.6263 |
| | RestorerID | 25.8816 | 0.6974 | 0.2834 | 63.7824 | 0.8113 | 4.4475 | 73.5447 | **0.7231** |
| | InstantRestore* | 26.2172 | 0.7345 | 0.2071 | 51.4018 | 0.8592 | 5.9684 | 69.7801 | 0.5724 |
| | ReF-LDM* | 25.6431 | 0.7159 | 0.2452 | 51.4180 | 0.8067 | 4.6921 | 73.7572 | 0.6718 |
| | FaceMe* | 26.5592 | 0.7277 | 0.2574 | 57.9200 | 0.8067 | 4.9615 | 72.0050 | 0.6342 |
| | RefSTAR | 26.4645 | 0.7337 | 0.1823 | 42.2977 | 0.8523 | 4.7523 | 71.6428 | 0.6406 |
| | AdaRefSR | 27.0616 | 0.7476 | 0.1762 | 41.9339 | 0.8622 | 4.3176 | 74.4435 | 0.6472 |
| | Ours* | **27.2383** | 0.7489 | 0.1548 | 35.6720 | 0.8796 | 4.1439 | 74.1893 | 0.6200 |
| | Ours-Full* | 27.1844 | **0.7516** | **0.1486** | **35.5258** | **0.9021** | **4.0751** | 74.5981 | 0.6236 |
| FFHQ-Ref | CodeFormer | 25.6627 | 0.6825 | 0.2401 | 32.8341 | 0.8416 | 4.2341 | **75.2574** | **0.6945** |
| | DMDNet* | 26.2755 | 0.7139 | 0.2722 | 35.3261 | 0.8276 | 4.4043 | 71.2693 | 0.6287 |
| | RestorerID | 25.5616 | 0.6785 | 0.2905 | 33.5038 | 0.8341 | 4.2521 | 72.0061 | 0.6926 |
| | InstantRestore* | 25.7701 | 0.7116 | 0.2482 | 35.3102 | 0.8592 | 5.8380 | 69.4892 | 0.5709 |
| | ReF-LDM* | 25.4378 | 0.6981 | 0.2565 | 29.9068 | 0.8466 | 4.5221 | 73.0942 | 0.6635 |
| | FaceMe* | 26.3617 | 0.7191 | 0.2334 | 29.8711 | 0.8588 | 4.7733 | 71.4272 | 0.6199 |
| | RefSTAR | 25.8421 | 0.7065 | 0.2320 | 30.1595 | 0.8670 | 4.6176 | 72.5442 | 0.6415 |
| | AdaRefSR | 26.5725 | 0.7239 | 0.2132 | 27.5722 | 0.8781 | 4.3586 | 73.4873 | 0.6592 |
| | Ours* | 26.6149 | 0.7238 | 0.1918 | 24.8012 | 0.8945 | 4.1138 | 73.0259 | 0.6110 |
| | Ours-Full* | **26.6322** | **0.7291** | **0.1699** | **23.9486** | **0.9258** | **3.9052** | 74.1158 | 0.6360 |

Table 2: Quantitative comparison with state-of-the-art methods on the real-world test set. Best and second-best results are highlighted in bold and underlined, respectively. Methods marked with * support multiple reference images.

| Metric | CodeFormer | DMDNet* | RestorerID | InstantRestore* | ReF-LDM* | FaceMe* | RefSTAR | AdaRefSR | Ours* | Ours-Full* |
|---|---|---|---|---|---|---|---|---|---|---|
| $ID_{ref}$↑ | 0.3767 | 0.3672 | 0.4004 | 0.4653 | 0.4211 | 0.4379 | **0.5192** | 0.4223 | 0.4837 | 0.5047 |
| NIQE↓ | **4.4064** | 5.4114 | 4.9816 | 6.1071 | 4.7943 | 5.0632 | 5.0314 | 6.2070 | 5.1455 | 5.0617 |
| MUSIQ↑ | 72.9365 | 58.3552 | 70.1072 | 70.7345 | 70.9052 | 71.2512 | 71.8228 | 71.1040 | **73.6157** | 73.6137 |
| CLIP-IQA↑ | 0.6684 | 0.4939 | 0.6610 | 0.6068 | 0.6590 | 0.6456 | 0.6272 | **0.6978** | 0.6712 | 0.6652 |

# 4 EXPERIMENTS

## 4.1 EXPERIMENT SETTING

**Datasets and Metrics.** We train our model on CelebRef-HQ (Li et al., 2023), which contains 10,555 aligned $512\times512$ face images from 1,005 identities. LQ inputs are synthesized using the two-stage Real-ESRGAN degradation pipeline (Wang et al., 2021). We hold out 40 identities for testing and further evaluate the model on FFHQ-Ref (Hsiao et al., 2024). To assess robustness to authentic degradations, we further collect a real-world test set covering 10 identities. We report full-reference metrics including PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), together with the no-reference metrics NIQE, CLIP-IQA (Wang et al., 2023), and MUSIQ (Ke et al., 2021). IDS and $ID_{ref}$ are computed as the cosine similarities between ArcFace (Deng et al., 2019) features of the restored image and those of the corresponding GT and reference images, respectively, thereby measuring identity consistency with the target face.

**Implementation Details.** Both RefRouteFR and RefRouteFR-Full are built on FLUX2-klein-9B and fine-tuned with LoRA of rank 32. During training, we randomly sample one to four same-identity HQ images as references, while four reference images are used during evaluation. We optimize the model with AdamW using a learning rate of $1 \times 10^{-4}$, an effective batch size of 8, and a reference dropout ratio of 10%. We first train RefRouteFR-Full for 10K steps with full reference tokens, then initialize RefRouteFR from it and train it for another 10K steps with sparse reference-token routing. In RefRouteFR, the router keeps 512 reference tokens in each Transformer block. All experiments are conducted on NVIDIA A100 GPU.

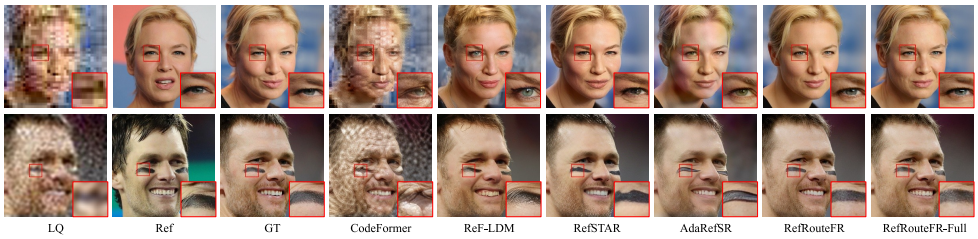


Figure 4: Visual comparison with state-of-the-art methods on reference-based face restoration benchmarks. RefRouteFR can take multiple references and route informative reference tokens in the unified token space; for clarity, only one representative reference image is shown for each example.

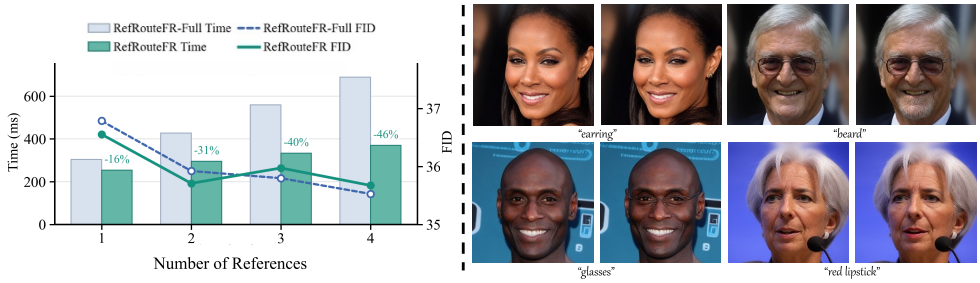


Figure 5: **Left:** Efficiency–quality trade-off on CelebRef-HQ as the number of references varies. Bars report inference time on $512 \times 512$ images, and curves report FID. **Right:** Text-controllable restoration results. Within each pair, the left result is restored without an attribute-specific prompt, while the right uses the prompt shown below to recover the corresponding facial attribute.

### 4.2 Comparison with State-of-the-Arts

**Quantitative Comparison.** We compare RefRouteFR and RefRouteFR-Full with representative blind and reference-based face restoration methods, including CodeFormer (Zhou et al., 2022), DMDNet (Li et al., 2023), RestorerID (Ying et al., 2024), InstantRestore (Zhang et al., 2025a), ReF-LDM (Hsiao et al., 2024), FaceMe (Liu et al., 2025c), RefSTAR (Yin et al., 2026), and AdaRefSR (Wang et al., 2026b). As shown in Table 1, both variants achieve the best or second-best performance on most fidelity, identity, and perceptual metrics across CelebRef-HQ and FFHQ-Ref. RefRouteFR-Full delivers the strongest overall performance, while RefRouteFR remains competitive despite processing only sparsely routed reference tokens. Results on the real-world test set in Table 2 further show that both variants maintain competitive identity consistency in terms of $\mathrm{ID}_{\mathrm{ref}}$ while achieving the best MUSIQ scores, indicating strong visual quality under authentic degradations. Together, these results demonstrate a favorable balance between reference identity consistency and perceptual quality across both synthetic and real-world settings.

**Qualitative Comparison.** The visual comparisons in Fig. 4 show that both RefRouteFR variants effectively exploit reference information. Compared with existing methods, RefRouteFR better restores identity-discriminative textures and fine-grained details such as eye color, wrinkles, and facial marks, while preserving natural facial structure and texture. Competing methods often miss or mismatch these reference-specific attributes. Additional qualitative comparisons are provided in Appendix C.

### 4.3 Efficiency–Quality Trade-off under Multiple References

Fig. 5 (left) evaluates the efficiency–quality trade-off as the number of reference images increases. RefRouteFR-Full processes all reference tokens in every Transformer block, so its cost grows rapidly with more references. RefRouteFR instead forwards only a fixed Top-$K$ token set, keeping the extra cost limited and yielding increasing efficiency gains at comparable FID. Thus, sparse routing scales reference-based restoration to multi-reference settings without redundant dense processing.

Table 3: End-to-end inference time (ms) on a single NVIDIA A100 GPU at $512 \times 512$ resolution. * denotes multi-reference; † multi-step sampling.

| Method | Time (ms) | Method | Time (ms) |
|---|---|---|---|
| RestorerID† | 5563 | RefSTAR | 468 |
| FaceMe*,† | 4661 | AdaRefSR | 246 |
| ReF-LDM*,† | 1099 | Ours-Full* | 689 |
| InstantRestore* | 474 | Ours* | 370 |

Table 4: Ablation of router design, latent calibrator ("Calib."), and $\mathcal{L}_{\mathrm{rKL}}$ regularization on CelebRef-HQ; best and second-best results are in bold and underlined.

| Router | Calib. | rKL | LPIPS↓ | FID↓ | IDS↑ | NIQE↓ |
|---|---|---|---|---|---|---|
| CNN | Yes | Yes | <u>0.1598</u> | 37.7036 | 0.8725 | 4.1448 |
| Ref-Only | Yes | Yes | 0.1618 | <u>37.2832</u> | 0.8681 | **4.1102** |
| Full-Context | Yes | No | 0.1635 | 39.0704 | 0.8707 | 4.1589 |
| Full-Context | No | – | 0.1667 | 38.5373 | <u>0.8749</u> | 4.2182 |
| Full-Context | Yes | Yes | **0.1548** | **35.6720** | **0.8796** | <u>4.1439</u> |

### 4.4 Text Controllability

Besides reference guidance, RefRouteFR inherits the text-controllable ability of the pretrained FLUX2 backbone. As shown in Fig. 5 (right), text prompts can guide the restoration of specific facial attributes when local details are ambiguous or missing in severely degraded inputs. By complementing visual references with semantic cues, RefRouteFR enables controllable attribute recovery while preserving natural facial structure and identity consistency.

### 4.5 Inference Time Statistics

Table 3 reports the end-to-end inference time on a single NVIDIA A100 GPU at $512\times512$ resolution, including the backbone forward pass, VAE encoding and decoding, and all auxiliary modules (averaged over 50 runs after three warm-up iterations, with 50 sampling steps for multi-step methods). Combined with the restoration results above, these results demonstrate a favorable quality–efficiency trade-off while supporting multiple references. This efficiency stems from modeling the LQ image, references, and text in a unified token space, avoiding costly dedicated reference-injection branches, while sparse routing compresses reference tokens before expensive backbone computation.

### 4.6 Ablation Study

#### 4.6.1 Comparison between Different Router Methods

Table 4 compares different routing strategies under the same calibration settings. CNN and Ref-Only achieve reasonable performance, but relying only on spatial saliency or pairwise LQ–reference similarity remains suboptimal. Although Ref-Only performs well on NIQE, its lower IDS suggests that strong LQ–reference correspondence alone does not guarantee recovering identity-specific details. By additionally considering LQ self-context and text context, Full-Context better evaluates tokens for the current restoration demand.

#### 4.6.2 Effect of Latent Calibration and rKL Regularization

We further analyze the latent calibrator and $\mathcal{L}_{\mathrm{rKL}}$ under the Full-Context router. Removing the calibrator degrades performance, supporting the need to correct the LQ initialization before the one-step update. Retaining it without $\mathcal{L}_{\mathrm{rKL}}$ also degrades clearly, showing that the gain relies on regularization rather than unconstrained fitting: $\mathcal{L}_{\mathrm{rKL}}$ pulls the predicted distribution toward a Gaussian centered at the HQ–LQ latent residual, encouraging trajectory-consistent initialization.

## 5 Conclusion

In this paper, we presented **RefRouteFR**, a one-step reference-based face restoration framework centered on demand-driven sparse reference-token routing. It jointly models LQ, reference, and text tokens in the native multimodal space of a pretrained model without auxiliary reference-injection modules, while its lightweight router forwards only restoration-relevant reference tokens through the Transformer blocks, cutting redundant computation in multi-reference scenarios. With LQ latent calibration, it enables one-step restoration without iterative denoising. Experiments show strong restoration quality with flexible text control and a favorable quality–efficiency trade-off.

### AI USE STATEMENT

In this work, we used generative AI tools to generate the identity-related training prompts (synthetic training annotations), to assist in implementing parts of our methods, to propose and refine research hypotheses, and to draft, polish, and format the manuscript. We have reviewed all AI-assisted work: AI-assisted code was tested by the authors, and all drafted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We provide implementation and evaluation details to facilitate reproducibility. The model architecture and demand-driven routing mechanism are described in Sec. 3.2, and the training objectives are defined in Sec. 3.3 and Appendix A.4. The training data construction, evaluation benchmarks, and metrics are described in Sec. 4.1 and Appendix A.5. Implementation details, including the training configuration, prompt generation, and router initialization, are reported in Appendices A.1–A.3. Additional experiments on the routed-token budget and the number of references, together with router behavior analyses, are provided in Appendix B, and qualitative comparisons on all benchmarks are included in Appendix C.

## REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4599–4603, 2023. doi: 10.1109/CVPRW59228.2023.00484.

Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. $\Delta$-DiT: A training-free acceleration method tailored for diffusion transformers, 2024a. URL https://arxiv.org/abs/2406.01125.

Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. Toward real-world blind face restoration with generative diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):8494–8508, 2024b. doi: 10.1109/TCSVT.2024.3383659.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Huu-Phu Do, Yu-Wei Chen, Yi-Cheng Liao, Chi-Wei Hsiao, Han-Yang Wang, Wei-Chen Chiu, and Ching-Chun Huang. DynFaceRestore: Balancing fidelity and quality in diffusion-guided blind face restoration with dynamic blur-level mapping and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10432–10441, 2025.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23174–23184, 2025. doi: 10.1109/CVPR52734.2025.02158.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23440–23450, 2026.

Nan Gao, Jia Li, Huaibo Huang, Zhi Zeng, Ke Shang, Shuwu Zhang, and Ran He. DiffMAC: Diffusion manifold hallucination correction for high generalization blind face restoration, 2024. URL https://arxiv.org/abs/2403.10098.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Chi-Wei Hsiao, Yu-Lun Liu, Cheng-Kun Yang, Sheng-Po Kuo, Yucheun Kevin Jou, and Chia-Ping Chen. ReF-LDM: A latent diffusion model for reference-based face image restoration. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2380.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Tianshu Kuai, Sina Honari, Igor Gilitschenski, and Alex Levinshtein. Towards unsupervised blind face restoration using diffusion prior. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1839–1849, 2025. doi: 10.1109/WACV61041.2025.00186.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34044–34053, 2025a.

Runyi Li, Bin Chen, Jian Zhang, and Radu Timofte. LAFR: Efficient diffusion-based blind face restoration via latent codebook alignment adapter, 2025b. URL https://arxiv.org/abs/2505.23462.

Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Computer Vision – ECCV 2018*, volume 11217 of *Lecture Notes in Computer Science*, pp. 278–296, 2018. doi: 10.1007/978-3-030-01261-8_17.

Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, 2020.

Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5904–5917, 2023.

Guoqiang Liang, Qingnan Fan, Bingtao Fu, Jinwei Chen, Hong Gu, and Lin Wang. AuthFace: Towards authentic blind face restoration with face-oriented generative diffusion prior. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Computer Vision–ECCV 2024*, pp. 430–448, 2024. doi: 10.1007/978-3-031-73202-7_25.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with TaylorSeers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15853–15863, 2025a.

Joseph Liu, Joshua Geddes, Ziyu Guo, Haomiao Jiang, and Mahesh Kumar Nandwana. Smooth Cache: A universal inference acceleration technique for diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3220–3229, 2025b. doi: 10.1109/CVPRW67362.2025.00305.

Siyu Liu, Zheng-Peng Duan, Jia OuYang, Jiayi Fu, Hyunhee Park, Zikun Liu, Chun-Le Guo, and Chongyi Li. FaceMe: Robust blind face restoration with personal identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5567–5575, 2025c. doi: 10.1609/aaai.v39i5.32593.

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. ToMA: Token merge with attention for diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40930–40951, 2025.

Yunqi Miao, Zhiyu Qu, Mingqi Gao, Changrui Chen, Jifei Song, Jungong Han, and Jiankang Deng. Unlocking the potential of diffusion priors in blind face restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13471–13480, 2025.

Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Zhiyao Ren, Yibing Zhan, Liang Ding, Gaoang Wang, Chaoyue Wang, Zhongyi Fan, and Dacheng Tao. Multi-step denoising scheduled sampling: Towards alleviating exposure bias for diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4667–4675, 2024. doi: 10.1609/aaai.v38i5.28267.

Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. FORA: Fast-forward caching in diffusion transformer acceleration, 2024. URL https://arxiv.org/abs/2407.01425.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025. doi: 10.1109/CVPR52734.2025.00223.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. doi: 10.1007/s11263-024-02168-7.

Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. OSDFace: One-step diffusion model for face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12626–12636, 2025.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and trajectory regularization for diffusion transformer super-resolution, 2026a. URL https://arxiv.org/abs/2603.06275.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Bin Wu, Wei Wang, Yahui Liu, Zixiang Li, and Yao Zhao. DiffusionReward: Enhancing blind face restoration through reward feedback learning, 2025a. URL https://arxiv.org/abs/2505.17910.

Haoyu Wu, Jingyi Xu, Hieu Le, and Dimitris Samaras. Importance-based token merging for efficient image and video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4983–4995, 2025b.

Hongjie Wu, Linchao He, Mingqin Zhang, Dongdong Chen, Kunming Luo, Mengting Luo, Ji-Zhe Zhou, Hu Chen, and Jiancheng Lv. Diffusion posterior proximal sampling for image restoration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 214–223, 2024a. doi: 10.1145/3664647.3681556.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92529–92553, 2024b. doi: 10.52202/079017-2938.

Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhicun Yin, Junjie Chen, Ming Liu, Zhixin Wang, Fan Li, Renjing Pei, Xiaoming Li, Rynson WH Lau, and Wangmeng Zuo. RefSTAR: Blind face image restoration with reference selection, transfer, and reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12053–12062, 2026.

Jiacheng Ying, Mushui Liu, Zhe Wu, Runming Zhang, Zhu Yu, Siming Fu, Si-Yuan Cao, Chao Wu, Yunlong Yu, and Hui-Liang Shen. RestorerID: Towards tuning-free face restoration with ID preservation, 2024. URL https://arxiv.org/abs/2411.14125.

Howard Zhang, Yuval Alaluf, Sizhuo Ma, Achuta Kadambi, Jian Wang, and Kfir Aberman. InstantRestore: Single-step personalized face restoration with shared-image attention. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 145:1–145:10, 2025a. doi: 10.1145/3721238.3730628.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1724–1732, 2024.

Ziying Zhang, Xiang Gao, Zhixin Wang, Qiang Hu, and Xiaoyun Zhang. TD-BFR: Truncated diffusion model for efficient blind face restoration. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025b.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In *International Conference on Learning Representations*, 2025.

Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25806–25816, 2024.

Mo Zhou, Keren Ye, Viraj Shah, Kangfu Mei, Mauricio Delbracio, Peyman Milanfar, Vishal M. Patel, and Hossein Talebi. Reference-guided identity preserving face restoration. *Transactions on Machine Learning Research*, 2026.

Shangchen Zhou, Kelvin C. K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30599–30611, 2022.

This appendix complements the main paper with additional details and analyses. Specifically, we provide:

- additional implementation details, including training configuration, prompt generation, router initialization, loss functions, and metric computation;
- additional experiments on component complexity, the routed-token budget, and the number of reference images;
- router behavior analyses on reference-token redundancy, layer-wise reference utilization, and router selection quality;
- an additional ablation of the frequency distribution loss (FDL);
- additional visual comparisons on CelebRef-HQ, FFHQ-Ref, and the real-world test set.

# A Additional Implementation Details

## A.1 Training Configuration

Table A.1: Core training configuration of RefRouteFR.

| Setting | Configuration |
|---|---|
| Backbone | FLUX2-klein-9B |
| LoRA rank | 32 |
| LoRA alpha | 16 |
| LoRA target | Attn., FFN, output proj. |
| Starting timestep $t_0$ | 0.601 |
| Router channels | 256 |
| References | 1–4 same-identity images; dropout 0.1 |
| Optimizer | AdamW; cosine decay; 500-step warm-up |
| Learning rates | $10^{-4}$ (Generator); $2 \times 10^{-5}$ (discriminator) |
| Batch size | 8 |
| Precision | BF16 |
| Loss weights | $(\lambda_1, \lambda_{\mathrm{LPIPS}}, \lambda_{\mathrm{GAN}}, \lambda_{\mathrm{FDL}}, \lambda_{\mathrm{rKL}}) = (1, 3, 0.1, 0.002, 0.01)$ |
| Training stages | RefRouteFR-Full: 10K steps;<br>RefRouteFR: +10K steps from Full |

Table A.1 summarizes the core configuration used to train RefRouteFR.

## A.2 Prompt Generation

We employ two strategies to generate training prompts, both of which are encoded by the FLUX2 text encoder to provide semantic conditioning. First, we apply the Recognize Anything Model (RAM) (Zhang et al., 2024) online to each HQ training target and convert the recognized visual concepts into a concise description of facial appearance, hairstyle, facial attributes, and visible accessories. Second, we generate prompts offline using the multimodal Gemma-4-12B-IT model (Gemma Team, 2026) to generate richer identity-related descriptions. Given the GT image as the primary input and another same-identity HQ image as auxiliary context, the model produces a concise and objective description in two or three sentences. The description is restricted to attributes visible in the GT image, excluding details observed only in the auxiliary reference.

## A.3 Sparse Router Initialization

Each attention router is placed immediately before a Transformer block. Instead of using random initialization, we initialize its compact query and key projections from the corresponding downstream block, allowing the initial routing scores to inherit its dominant attention subspaces. For the $i$-th block, we extract the query and key matrices $\mathbf{W}_Q^i$ and $\mathbf{W}_K^i$ from the first two partitions of its fused QKV projection. For $a \in \{Q, K\}$, we compute

$$\mathbf{W}_a^i = \mathbf{U}_a^i \mathbf{\Sigma}_a^i (\mathbf{V}_a^i)^\top, \qquad \mathbf{P}_a^i = (\mathbf{\Sigma}_{a,1:d_r}^i)^{1/2} (\mathbf{V}_{a,1:d_r}^i)^\top, \tag{A.1}$$

where $d_r = 256$ is the router projection dimension. The truncated matrices $\mathbf{P}_Q^i$ and $\mathbf{P}_K^i$ initialize the router query and key projections, respectively; the text-key projection shares the same initialization as $\mathbf{P}_Q^i$. The square-root singular-value scaling emphasizes the dominant input directions while mapping them into the compact router space.

We initialize the query and key RMSNorm scale parameters of each router from the corresponding RMSNorm parameters in the pretrained Transformer block. When their dimensions differ, the pretrained parameters are repeated or truncated to match the router dimension. This initialization is performed layer by layer, with router $i$ paired with Transformer block $i$.

### A.4 Loss Function

Our model is trained with the following weighted objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{G}} = \quad & \lambda_1 \mathcal{L}_1 + \lambda_{\mathrm{lpips}} \mathcal{L}_{\mathrm{LPIPS}} + \lambda_{\mathrm{gan}} \mathcal{L}_{\mathrm{GAN}} \\ & + \lambda_{\mathrm{fdl}} \mathcal{L}_{\mathrm{FDL}} + \lambda_{\mathrm{rKL}} \mathcal{L}_{\mathrm{rKL}}.\end{aligned} \tag{A.2}$$

The L1 loss preserves pixel-level fidelity by measuring the mean absolute difference between the restored image and its HQ target:

$$\mathcal{L}_1 = \frac{1}{3HW} \left\| \hat{x}_{\mathrm{HQ}} - x_{\mathrm{HQ}} \right\|_1 . \tag{A.3}$$

The VGG-based LPIPS loss (Zhang et al., 2018) improves perceptual similarity and is averaged over the training batch:

$$\mathcal{L}_{\mathrm{LPIPS}} = \frac{1}{B} \sum_{b=1}^{B} d_{\mathrm{LPIPS}}^{\mathrm{VGG}}(\hat{x}_{\mathrm{HQ}}^b, x_{\mathrm{HQ}}^b). \tag{A.4}$$

The GAN loss encourages realistic facial details. We apply a vision-aided discriminator with a frozen DINO backbone to random $224 \times 224$ crops, denoted by $\mathcal{C}(\cdot)$. Let $D_\ell(\cdot)$ be the logit at the $\ell$-th output level. The non-saturating logistic losses averaged over all $L$ levels are

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}} = \quad & -\frac{1}{L} \sum_{\ell=1}^{L} \mathrm{E}[\log \sigma(D_\ell(\mathcal{C}(\hat{x}_{\mathrm{HQ}})))] \,, \\ \mathcal{L}_D = \quad & -\frac{1}{L} \sum_{\ell=1}^{L} \Big( \mathrm{E}[\log \sigma(D_\ell(\mathcal{C}(x_{\mathrm{HQ}})))] \\ & \qquad + \mathrm{E}[\log \sigma(-D_\ell(\mathcal{C}(\hat{x}_{\mathrm{HQ}})))] \Big).\end{aligned} \tag{A.5}$$

$\mathcal{L}_D$ is optimized separately from $\mathcal{L}_{\mathrm{G}}$.

The FDL loss (Ni et al., 2024) reduces periodic artifacts and is evaluated after mapping both images to $[0, 1]$:

$$\begin{aligned}\mathcal{L}_{\mathrm{FDL}} = \quad & \mathrm{SW}\Big( \mathcal{A}_{\Phi(\bar{\hat{x}}_{\mathrm{HQ}})}, \mathcal{A}_{\Phi(\bar{x}_{\mathrm{HQ}})} \Big) \\ & + 0.5\, \mathrm{SW}\Big( \mathcal{S}_{\Phi(\bar{\hat{x}}_{\mathrm{HQ}})}, \mathcal{S}_{\Phi(\bar{x}_{\mathrm{HQ}})} \Big) \,,\end{aligned} \tag{A.6}$$

where $\bar{x} = (x + 1)/2$. We use $5 \times 5$ patches, stride 1, and 256 sliced-Wasserstein projections.

Following the main paper, the rKL loss regularizes the predicted calibration distribution toward a unit-covariance Gaussian centered at the HQ–LQ latent residual:

$$\mathcal{L}_{\mathrm{rKL}} = D_{\mathrm{KL}}(q_\phi(\Delta z \mid z_{\mathrm{LQ}}, t_0) \,\|\, \mathcal{N}(z_{\mathrm{HQ}} - z_{\mathrm{LQ}}, I)) \,. \tag{A.7}$$

### A.5 Metric Computation

We evaluate identity preservation using the normalized $d$-dimensional ArcFace embedding $f(\cdot)$ (Deng et al., 2019). Before feature extraction, each image is resized to $112 \times 112$ and normalized to the range $[-1, 1]$. We compute two identity similarity metrics:

$$\begin{aligned}\mathrm{IDS} \quad &= \quad f(\hat{x}_{\mathrm{HQ}})^\top f(x_{\mathrm{HQ}}), \\ \mathrm{ID}_{\mathrm{ref}} \quad &= \quad \frac{1}{n} \sum_{i=1}^{n} f(\hat{x}_{\mathrm{HQ}})^\top f(x_{\mathrm{ref}}^i),\end{aligned} \tag{A.8}$$

where $\{x_{\text{ref}}^i\}_{i=1}^n$ denotes the set of available same-identity reference images. Since the embeddings are $\ell_2$-normalized, the inner product corresponds to cosine similarity. We report IDS on synthetic benchmarks with paired ground-truth images, and $\text{ID}_{\text{ref}}$ on the real-world benchmark where ground truth is unavailable.

# B Additional Experimental Results

## B.1 Component Parameters and Computational Cost

Table B.1: Parameter count and computational cost of the backbone and auxiliary components. FLOPs are measured at $512 \times 512$ resolution images. Percentages are computed relative to the dense Transformer backbone.

| Component | Params | Relative Params | FLOPs | Relative FLOPs |
|---|---|---|---|---|
| Dense Transformer Backbone | 9.079B | 100% | 52.294T | 100% |
| Routed Transformer Backbone | 9.079B | 100% | 33.381T | 63.8% |
| Latent Calibrator | 10.03M | 0.11% | 5.61G | 0.011% |
| CNN Router | 53.50M | 0.59% | 21.90G | 0.042% |
| Ref-Only Attn. Router | 75.52M | 0.83% | 9.82G | 0.019% |
| Full-Context Attn. Router | 75.52M | 0.83% | 13.09G | 0.025% |

Table B.1 summarizes the parameter and computational overhead of the backbone and auxiliary components. Sparse reference-token routing substantially reduces the computation of the Transformer backbone, while the added routing and calibration modules contribute only a small fraction of the overall model size and computation. This indicates that the efficiency gain mainly comes from avoiding redundant reference-token processing rather than from adding lightweight auxiliary computation.

## B.2 Effect of the Number of Routed Reference Tokens

Table B.2: Effect of the number of routed tokens on CelebRef-HQ. Best and second-best results are highlighted in bold and underlined, respectively.

| Num. of Routed Tokens | LPIPS↓ | FID↓ | IDS↑ | NIQE↓ |
|---|---|---|---|---|
| 256 | 0.1581 | 36.1873 | 0.8677 | 4.1727 |
| 512 | 0.1548 | **35.6720** | <u>0.8796</u> | <u>4.1439</u> |
| 768 | **0.1528** | 35.9521 | 0.8774 | **4.1027** |
| 1024 | <u>0.1530</u> | <u>35.7520</u> | **0.8845** | 4.2103 |

Table B.2 studies how the Top-$K$ routing budget affects restoration quality while keeping the other settings fixed. Even with only 256 routed reference tokens, the model remains competitive with the larger-budget variants, providing further evidence that the reference branch is highly redundant. Nevertheless, its higher FID and lower IDS than the 512-token setting indicate that an overly restrictive bottleneck can discard subtle identity cues that are useful for perceptually faithful restoration. Increasing the budget to 512 recovers these cues and yields the best FID together with competitive fidelity, identity, and no-reference quality scores.

The gains become marginal beyond 512 tokens and, importantly, are not monotonic across metrics. Although 768 or 1024 tokens slightly improve several distortion metrics, they do not consistently improve FID. This suggests that once the most responsive facial regions have been selected, lower-ranked tokens are more likely to contain duplicated evidence or reference details that are less relevant to the current LQ input. Routing these tokens therefore adds computation without a proportional quality gain and may even introduce mild interference. We consequently use 512 tokens as the default setting: it lies near the knee of the quality–sparsity trade-off, rather than optimizing one individual metric at the expense of routing efficiency.

### B.3 Effect of the Number of Reference Images

Table B.3 evaluates whether RefRouteFR can exploit a varying number of same-identity references. With no reference, the model must rely on the degraded observation, and the generative prior of the pretrained backbone. The resulting performance is still reasonable, which is consistent with the use of reference dropout during training. Adding just one reference reduces FID from 38.0889 to 36.5527 and increases IDS from 0.8542 to 0.8773, showing that the improvement comes from identity evidence that cannot be reliably inferred from the LQ input alone.

Table B.3: Effect of the number of reference images on CelebRef-HQ. Best and second-best results are highlighted in bold and underlined, respectively.

| Num. of Ref. Images | LPIPS↓ | FID↓ | IDS↑ | NIQE↓ |
|---|---|---|---|---|
| 0 | 0.1635 | 38.0889 | 0.8542 | 4.1942 |
| 1 | 0.1566 | 36.5527 | 0.8773 | 4.2021 |
| 2 | 0.1551 | <u>35.7076</u> | 0.8778 | 4.1730 |
| 3 | <u>0.1550</u> | 35.9726 | <u>0.8783</u> | <u>4.1520</u> |
| 4 | **0.1548** | **35.6720** | **0.8796** | **4.1439** |

Additional references provide complementary views of the same identity: a detail obscured by pose, expression, or illumination in one image may be visible in another. The router can compare this larger candidate pool against the current restoration demand and select the most relevant local evidence. Accordingly, four references achieve the best results across all four reported metrics. The improvement is not strictly monotonic, however. Because the Transformer receives a fixed Top-$K$ token budget, cues from different references compete for the same routing capacity, and extra images may contribute redundant or less compatible details. This explains both the diminishing returns after the first few references and small metric fluctuations. Rather than blindly accumulating reference content, the router uses multiple images as a candidate bank and keeps a bounded amount of evidence. Randomly sampling one to four references during training further makes this behavior robust to the number of references available at test time.

### B.4 Router Behavior Analysis

#### B.4.1 Reference-Token Redundancy Analysis

We use an attention-guided two-pass diagnostic on CelebRef-HQ to probe reference-token redundancy. A full pass first ranks each reference token by the attention mass it receives, summed over all queries and averaged across heads. For a given budget, the Top-$K$ tokens retain their current-layer outputs, while lower-ranked tokens bypass the current block and reuse their incoming features. We then compare the resulting image with both the ground truth and the full output in Table B.4.

Table B.4: PSNR comparison under different reference-token budgets on CelebRef-HQ. "PSNR vs. Full" uses the densely processed output without token pruning as the reference.

| Kept Ref. Tokens | 256 | 512 | 768 | 1024 |
|---|---|---|---|---|
| PSNR vs. GT | 21.8362 | 24.5083 | 25.0476 | 25.3891 |
| PSNR vs. Full | 22.8580 | 28.3328 | 32.9789 | 35.1789 |

The two PSNR measures disentangle restoration error from routing-induced deviation: comparison with the ground truth evaluates output fidelity, whereas comparison with RefRouteFR-Full controls for the dense model and isolates sensitivity to replacing lower-ranked token updates. Their joint behavior indicates that reference redundancy is rank-structured rather than uniform. A compact set of highly attended tokens carries the dominant reference-conditioned response, whereas the lower-attention tail is dominated by duplicated or weakly relevant evidence. At moderate budgets, bypassing these current-layer updates causes limited deviation because the retained tokens already cover the restoration-relevant cues. Once the budget becomes too restrictive, however, selection crosses from redundancy removal into information loss: complementary evidence from different facial re-

gions and reference views is discarded, and the resulting errors accumulate through subsequent blocks. The observed capacity threshold thus marks the boundary between effective redundancy pruning and harmful over-pruning. Accordingly, we retain 512 reference tokens in our experiments, consistent with the task-level comparison in Table B.2.

### B.4.2 Layer-Wise Attention and Router Selection Quality

To further understand how reference information is utilized throughout the backbone, Fig. B.1 compares the attention received by different reference-token sets across Transformer blocks. For each block, we compute the softmax attention from the LQ-token queries to all keys. The reference attention ratio is defined as the fraction of the total attention mass assigned to a given reference-token set, averaged over attention heads and test samples.

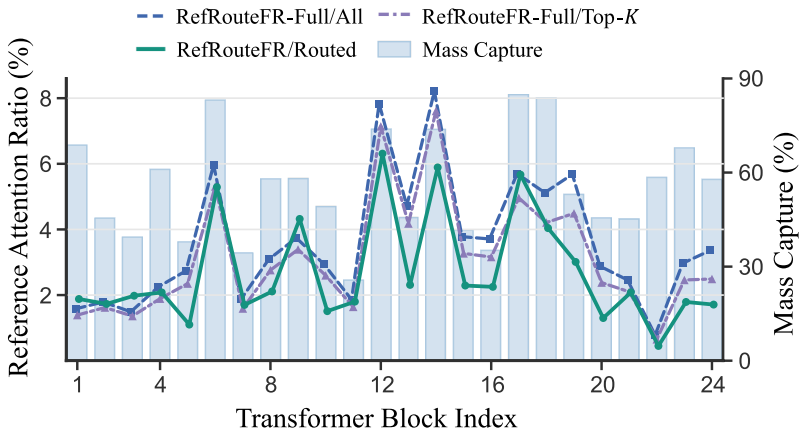


Figure B.1: Layer-wise reference attention and router selection quality on CelebRef-HQ. The curves show the reference attention ratios of all reference tokens in RefRouteFR-Full (Full/All), its 512 most attended tokens selected post hoc (Full/Top-$K$), and the tokens retained by RefRouteFR (RefRouteFR/Routed). The bars report the fraction of the Full/Top-$K$ attention mass captured by the routed token set, evaluated using RefRouteFR-Full attention scores.

We additionally evaluate the quality of router selection using attention-mass capture. Let $\mathcal{R}_l$ denote the $K$ reference tokens selected by the RefRouteFR router at block $l$, and let $\mathcal{T}_l$ denote the post-hoc Top-$K$ reference tokens ranked by their attention in RefRouteFR-Full. The attention-mass capture is defined as

$$\text{Capture}_l = \frac{\sum_{j\in\mathcal{R}_l} a_l^{\text{Full}}(j)}{\sum_{j\in\mathcal{T}_l} a_l^{\text{Full}}(j)}, \tag{B.1}$$

where $a_l^{\text{Full}}(j)$ denotes the attention mass received by reference token $j$ in RefRouteFR-Full, averaged over all LQ queries and attention heads. The resulting capture ratios are further averaged across the test samples. Both the numerator and denominator are evaluated using RefRouteFR-Full attention scores, thereby isolating token-selection quality from changes in attention normalization after sparse routing.

As shown in Fig. B.1, reference attention varies substantially across Transformer blocks, indicating that the demand for reference information is strongly layer-dependent. The RefRouteFR-Full/Top-$K$ curve closely follows the major peaks and valleys of Full/All and retains most of the attention assigned to the complete reference set. This concentration provides further evidence that the reference branch contains substantial token redundancy.

The RefRouteFR/Routed curve exhibits a similar layer-wise pattern, preserving the principal reference-attention peaks of RefRouteFR-Full. Moreover, the routed token sets capture a substantial fraction of the post-hoc Top-$K$ attention mass, with particularly strong capture around blocks exhibiting high reference responses. Although the capture ratio is lower in several weak-reference blocks, these blocks also assign relatively little overall attention to reference information. These results indicate that the router preferentially retains high-value reference tokens when reference

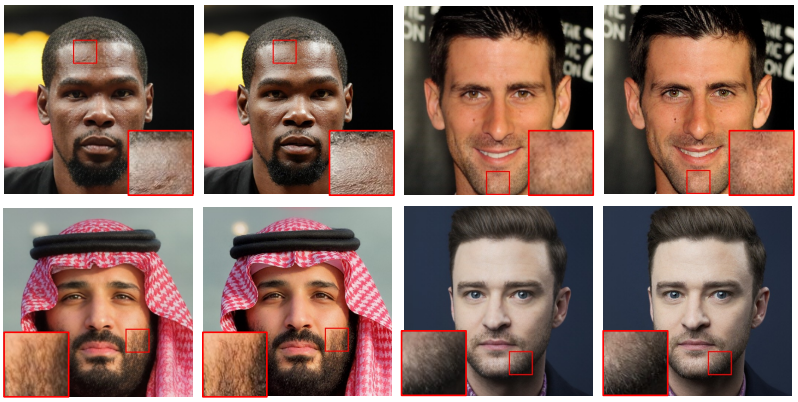

Figure B.2: Qualitative ablation of FDL. Within each pair, the left image is produced without FDL and the right image with FDL. Red boxes and enlarged crops highlight the repetitive textures on skin and facial-hair regions; FDL suppresses these artifacts and recovers more natural details.

guidance is most important, while filtering out redundant tokens and preserving the layer-dependent reference demand of the dense model.

### B.5 Ablation Study on FDL Loss

Figure B.2 shows the effect of FDL. Without FDL, the restored faces contain grid-like, rippled, or repetitive textures, particularly on smooth forehead and cheek regions and along dense facial hair. These patterns are emphasized in the enlarged crops. In contrast, training with FDL produces more irregular and natural skin and hair details. Following the analysis in StrSR (Wang et al., 2026a), we attribute these periodic artifacts to high-frequency spectral leakage in one-step DiTs: a single large update aggravates discretization errors, while patchification induces spatially periodic correlations. Since the L1, LPIPS, and GAN losses do not explicitly match the frequency statistics of the prediction and target, FDL supplies this missing constraint by minimizing sliced-Wasserstein distances between their amplitude and phase distributions, suppressing periodic artifacts without removing fine facial details.

## C Additional Visual Comparisons

We select representative examples from CelebRef-HQ, FFHQ-Ref, and the real-world test set. The corresponding visual comparisons are presented in Figures C.1, C.2, and C.3, respectively.

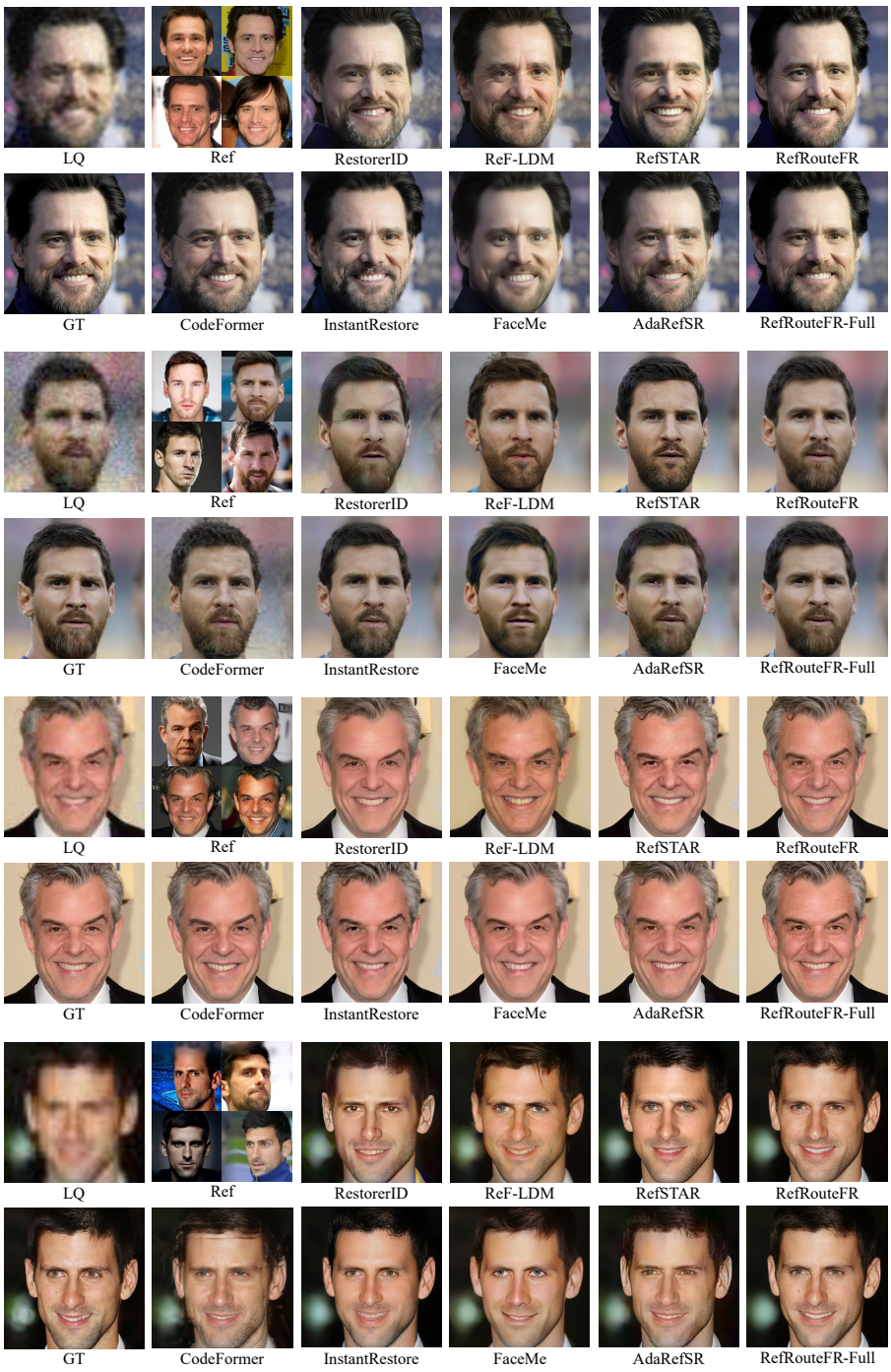


Figure C.1: Additional visual comparisons on CelebRef-HQ.

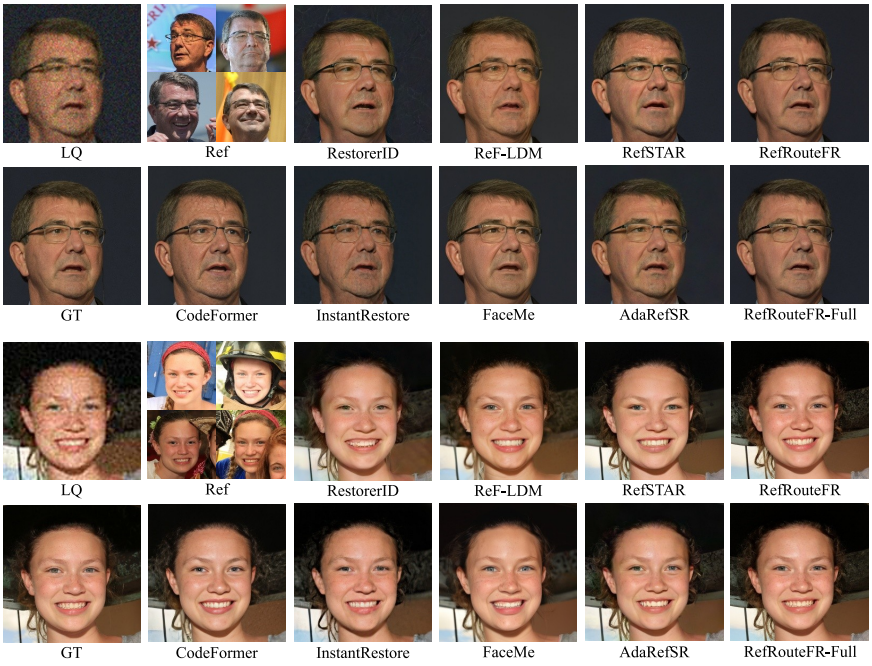


Figure C.2: Additional visual comparisons on FFHQ-Ref.

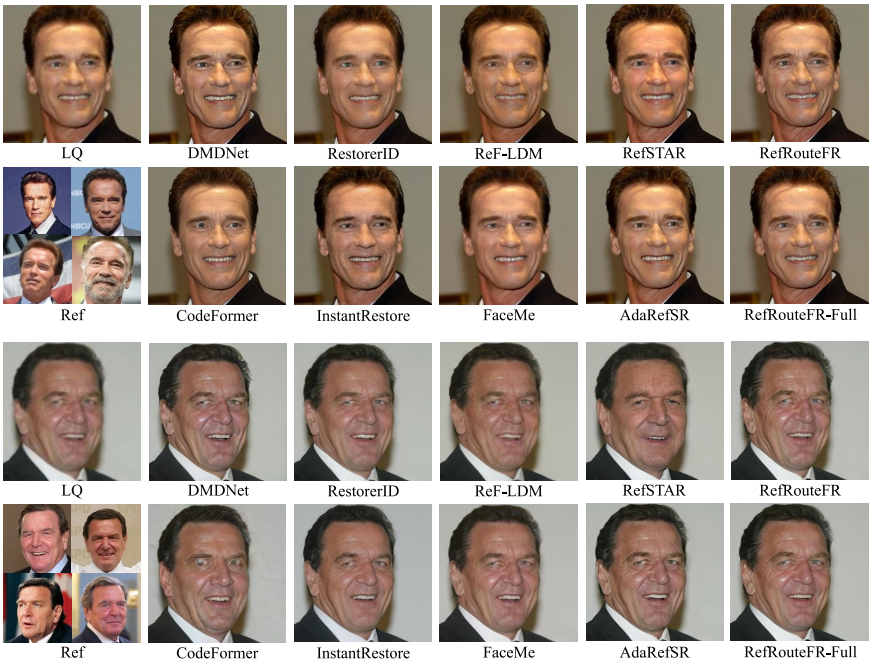


Figure C.3: Additional visual comparisons on the real-world test set.