FROM SCALE-EQUIVARIANT RECONSTRUCTION TO ONE-STEP DIFFUSION: LATENT ALIGNMENT FOR BLIND FACE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In blind face restoration, a central challenge lies in reconciling two distinct objectives: scale-equivariant reconstruction and photorealistic generation. Reconstruction models naturally handle arbitrary degradation scales and preserve facial structure, but fail to synthesize realistic high-frequency details. Diffusion models provide powerful generative priors for detail synthesis, yet their fixed-resolution latent sampling process is not inherently compatible with continuous scale variations and often depends on expensive multi-step denoising. To unify the structural reliability of reconstruction with the perceptual realism of generation, we introduce FaceDNO, a novel paradigm that decouples the restoration process into two native mathematical spaces. Specifically, we formulate reconstruction in a continuous function space by employing a resolution-equivariant neural operator to project degraded inputs into a scale-robust structural anchor. We then formulate generation in the discrete latent space, integrating this continuous anchor into a diffusion prior via a manifold-preserving Latent Spatial Feature Transform (Latent SFT). To seamlessly bridge these two spaces, we formulate the operator's reconstruction in the latent space as an isomorphic linear structure to the diffusion forward process. By matching their marginal distributions via Signal-to-Noise Ratios (SNR), we analytically identify the optimal mid-timestep for generative refinement. Experiments demonstrate that FaceDNO effectively combines scale-equivariant reconstruction with photorealistic generation, achieving state-of-the-art fidelity and realism across diverse degradations with one-step diffusion refinement.

1 INTRODUCTION

Blind face restoration (BFR) aims to recover identity-faithful, photorealistic faces from observations corrupted by unknown mixtures of blur, noise, compression, and resolution loss. When inference is restricted to a *single* diffusion refinement step, the bottleneck shifts from sampling length to *conditioning*, where the model must form a latent state that both supports structural recovery under continuous, arbitrary-scale degradations and occupies a noise level at which one-step denoising is valid. Existing one-step and few-step diffusion methods Wang et al. (2025); Wu et al. (2025); Lin et al. (2023) typically encode a fixed-grid low-quality input and rely on a heuristic or empirically chosen mid-timestep, leaving open how to construct a conditioning latent that is simultaneously scale-equivariant and schedule-compatible.

We argue that two failure modes make this conditioning problem concrete. First, a reconstruction-only structural anchor can suppress observation-specific residual cues that remain informative for identity and local texture, so feeding the anchor alone into a generative prior underuses the raw measurement Liang et al. (2021); Wang et al. (2021c). Second, even after fusing the anchor with the observation, the resulting latent may sit at an empirical Signal-to-Noise Ratio (SNR) that is poorly aligned with the diffusion schedule, resulting in a mismatched timestep that either under-refines textures or overwrites structure Wu et al. (2025). These failure modes motivate us to treat one-step BFR as *conditional latent construction*.

To this end, we introduce Face Diffusion Neural Operator (FaceDNO), a two-stage formulation that addresses the two failure modes separately. Stage 1 builds a scale-equivariant structural anchor with

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.

a Face Resolution Neural Operator (FRONet), extending continuous-coordinate neural operators Wei & Zhang (2023); Xu et al. (2025) to produce a resolution-robust estimate. Stage 2 encodes both the anchor and the upsampled observation, fuses them with a manifold-preserving Latent Spatial Feature Transform (Latent SFT) following spatial feature modulation Wang et al. (2018), and selects the refinement timestep by *dataset-level first-order SNR calibration* of the fused latent. Building on mid-timestep guidance and Latent Representation Refinement (LRR) Wu et al. (2025), the residual pathway restores cues suppressed by the anchor, and the SNR mapping places the fused state at a schedule-compatible working point for one-step refinement. Our contributions are threefold.

- **Dual-role conditional latent construction.** We show that a scale-equivariant structural anchor alone is insufficient for faithful one-step restoration, where reconstruction-based normalization can suppress observation-specific residual cues that remain useful for identity and texture. FaceDNO therefore constructs the conditioning latent from complementary sources, where a Face Resolution Neural Operator (FRONet) provides the geometric base, and manifold-preserving Latent SFT injects residual cues from the upsampled observation as a spatially varying modulation of refinement strength, rather than as a second structural prediction.
- **First-order SNR calibration.** Fusion changes the effective noise statistics of the conditioning state, so a heuristic mid-timestep chosen for a raw low-quality latent need not remain valid. Instead, we estimate a training-set signal scale, form the empirical SNR of the *fused* latent, and match it to the diffusion schedule to obtain a schedule-calibrated one-step working point, building on mid-timestep refinement and LRR Wu et al. (2025).
- **Mechanism-isolating evaluation.** To address the concern that gains may arise from backbone or loss stacking alone, we report capacity-controlled system attribution, Stage-2 input factorization (including anchor-only controls), timestep-prediction checks, and quality–latency trade-offs that isolate the proposed conditioning mechanisms on synthetic and real face benchmarks. Experiments show that our method achieves SOTA performance on both synthetic and real-world datasets.

2 Related Work

Structure Priors and Continuous Reconstruction. Blind face restoration balances geometric faithfulness with perceptual realism. Reconstruction-oriented CNN and Transformer models Dong et al. (2015); Li et al. (2020); Liang et al. (2021); Chen et al. (2023) retain structure but often under-recover texture, whereas GAN- and codebook-based priors Wang et al. (2021a); Yang et al. (2021); Zhou et al. (2022b); Wang et al. (2022); Gu et al. (2022) and diffusion restorers Yue & Loy (2022); Chen et al. (2024); Lin et al. (2023); Wang et al. (2025); Liang et al. (2025) improve realism at the cost of multi-step sampling or fixed-grid conditioning. Continuous-scale neural operators Wei & Zhang (2023); Luo et al. (2024); Liu & Tang (2025); Xu et al. (2025) learn resolution-flexible maps and have been coupled with diffusion for arbitrary-scale restoration. FaceDNO stays in this fidelity–perception setting: a Face Resolution Neural Operator (FRONet) provides a *scale-flexible structural anchor* that conditions a latent diffusion prior Wei & Zhang (2023); Xu et al. (2025).

One-Step Conditioning and Timestep Calibration. One-step and few-step restorers Wang et al. (2025); Lin et al. (2023) typically condition on a low-quality encoding with a fixed short sampler. OMGSR Wu et al. (2025) shows that a single mid-timestep with Latent Representation Refinement (LRR) can be effective when a degraded latent is treated as an intermediate noisy state, but the timestep is still chosen by empirical heuristics or search on the raw low-quality pathway. FaceDNO adopts this one-step mid-timestep and LRR setting Wu et al. (2025), yet applies them to a *fused* conditioning latent rather than a raw low-quality encoding, and selects the working point by dataset-level first-order SNR calibration of that fused state.

Conditional Feature Modulation in Latent Space. Spatially adaptive normalization and feature transforms inject side information into generative networks without replacing the backbone prior Wang et al. (2018). In restoration, related designs concatenate control maps, attach side branches, or modulate intermediate features so that geometry and appearance cues can be steered at inference time. These mechanisms are effective for controllable generation, but they do not by themselves specify how a scale-flexible reconstruction anchor should share a latent state with observation-specific residuals under a single diffusion step. Latent SFT places Spatial Feature Transform Wang et al. (2018) in VAE space as a residual affine update around the anchor latent, so the observation stream

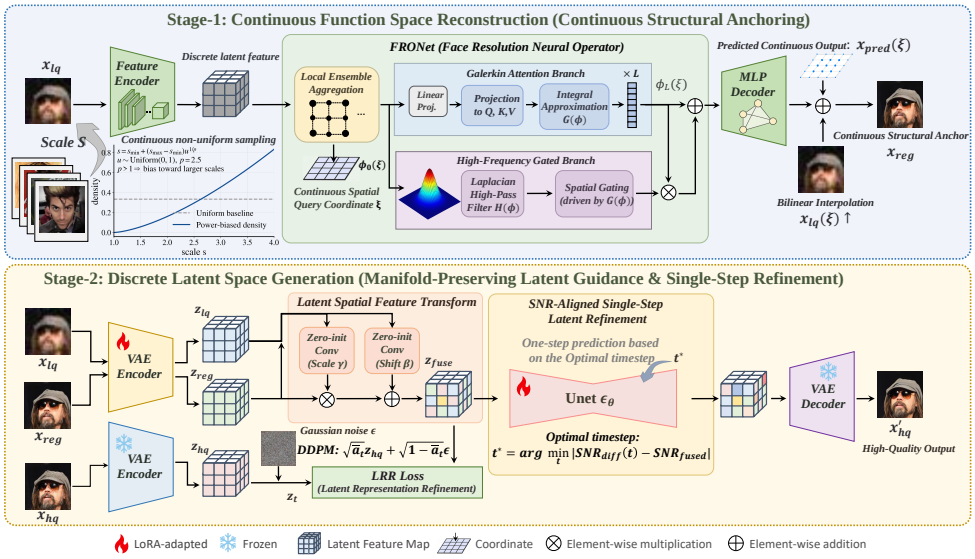

Figure 1: **Overview of FaceDNO.** Restoration is decoupled into continuous structural anchoring and one-step latent refinement. **(a) Stage-1.** FRONet maps x_{lq} to a scale-flexible structural anchor x_{reg}; a high-frequency gated Galerkin branch sharpens face-sensitive contours. **(b) Stage-2.** The anchor and upsampled observation are encoded and fused by manifold-preserving Latent SFT. A LoRA-adapted UNet then performs one-step refinement at the dataset-level SNR-calibrated timestep $\hat{t}$, under LRR supervision, to produce $\hat{x}_{hq}$.

modulates refinement strength while the structural manifold of the anchor remains the base state for one-step BFR.

3 Method

3.1 Problem Formulation and Overview

Given a low-quality face x_{lq}, we recover a high-quality face x_{hq} with a *single* diffusion refinement step. Under this budget, the bottleneck is *conditional latent construction*: the conditioning state must simultaneously (i) provide scale-equivariant structure, (ii) retain observation-specific residual cues, and (iii) occupy a noise level compatible with one-step denoising. These requirements are in tension. A strongly normalized structural anchor favors (i) but tends to suppress the residuals needed for (ii); fusing the observation back in restores (ii) but changes the effective noise statistics, so a heuristic timestep chosen for a raw low-quality latent need not satisfy (iii).

Dual-role conditioning. A continuous neural operator produces the structural anchor, and both the anchor and the upsampled observation are encoded into the VAE latent space of a pre-trained diffusion prior Rombach et al. (2022):

$$x_{reg} = \mathcal{F}_{\text{op}}(x_{lq}), \qquad z_{reg} = \mathcal{E}(x_{reg}), \qquad z_{lq} = \mathcal{E}(x_{lq}^{\uparrow}). \tag{1}$$

Here z_{reg} supplies the geometric base that Stage-2 should preserve, while z_{lq} carries degradation-dependent residual cues that reconstruction alone tends to suppress. We fuse them with manifold-preserving Latent SFT Wang et al. (2018):

$$z_{fused} = z_{reg} + \gamma(z_{lq}) \odot z_{reg} + \beta(z_{lq}). \tag{2}$$

First-order SNR calibration. Because fusion changes the working point of the conditioning state, we calibrate the *fused* latent rather than assuming that z_{reg} or a raw low-quality encoding already aligns with the diffusion schedule. On the training set, we estimate a scalar signal scale by least squares,

$$s^{\star} = \arg\min_{s} \mathbb{E}_{\mathcal{D}_{\text{train}}}\left[\|z_{fused} - s z_{hq}\|_2^2\right], \qquad n = z_{fused} - s^{\star} z_{hq}, \tag{3}$$

and define the empirical SNR

$$\widehat{\mathrm{SNR}}_{fused} = \frac{(s^\star)^2 \widehat{\mathrm{Var}}(z_{hq})}{\widehat{\mathrm{Var}}(n)}. \tag{4}$$

Matching $\widehat{\mathrm{SNR}}_{fused}$ to the diffusion schedule SNR yields a discrete candidate timestep $\hat{t}$ (Sec. 3.4). Scope of this first-order construction is stated in Sec. 5.

Pipeline. FaceDNO therefore assigns the three constraints to three mechanisms: FRONet builds x_{reg} for (i), Latent SFT forms z_{fused} for (ii), and dataset-level SNR calibration selects $\hat{t}$ for (iii). A LoRA-adapted UNet then refines z_{fused} in one step, and the denoised latent is decoded as $\hat{x}_{hq} = \mathcal{D}(\hat{z}_{hq})$.

3.2 Stage-1: Scale-equivariant Neural Operator Anchor

Blind degradations vary continuously in scale. A fixed-grid interpolation or discrete reconstruction anchor ties the structural estimate to one sampled resolution; the resulting geometric error is then amplified by a single generative step that cannot iteratively correct it. Stage-1 therefore builds a scale-equivariant structural anchor before latent refinement, which we later compare against bicubic and learned discrete anchors (Sec. 4.4).

We instantiate a Face Resolution Neural Operator (FRONet) on the continuous-coordinate query formulation of SRNO Wei & Zhang (2023) and the Galerkin attention operator used in NeurOp-Diff Xu et al. (2025). Given x_{lq}, an encoder extracts a discrete latent feature map. For any query coordinate $\xi \in \Omega \subset \mathbb{R}^2$, local ensemble aggregation yields an initial representation $\phi_0(\xi) \in \mathbb{R}^d$. Continuous querying treats structure as a function on Ω, so the same operator can be evaluated at the target resolution without rebinding the anchor to a single discrete grid.

The operator updates this representation with a kernel integral $\mathcal{K}(\phi)(\xi) = \int_\Omega K(\phi(\xi), \phi(\eta))\phi(\eta)\, d\eta$, approximated by Galerkin-type attention Xu et al. (2025): $G(\phi) = Q(K^\top V)/N$, where $N = H \times W$ is the latent grid size.

Global Galerkin mixing is an effective smoother: it aggregates long-range context but attenuates high-curvature contours such as eyelids and lip boundaries. Those contours are precisely the geometric cues that Stage-2 must respect as a structural manifold. As a supporting design, FRONet therefore adds a high-frequency gated branch: a Laplacian-like high-pass map $H(\phi)$ is spatially gated by $G(\phi)$,

$$Y = G(\phi) + \alpha \cdot (\sigma(W_g G(\phi)) \odot W_h H(\phi)), \tag{5}$$

where σ is the sigmoid, W_g and W_h are learnable projections, and α is a residual scale. An MLP decoder with a bilinear residual skip then produces the continuous anchor

$$x_{reg}(\xi) = \mathcal{D}_{mlp}(Y(\xi)) + x_{lq}(\xi). \tag{6}$$

The continuous query path targets cross-scale structural reliability; the gated high-frequency path reduces contour over-smoothing so that Stage-2 inherits a sharper geometric base.

3.3 Stage-2: Manifold-Preserving Latent Guidance

A reconstruction-only anchor is not a lossless encoding of the measurement. FRONet maps diverse scales and degradations toward a canonical, structurally restored face: this normalization yields a reliable geometric base, but can suppress observation-specific residual cues such as where blur, noise, or compression is spatially severe. Conditioning Stage-2 on z_{reg} alone therefore underuses the raw observation; we keep the upsampled measurement as a second stream and use it to modulate *where* and *how strongly* to refine, rather than to re-predict structure (Sec. 4.4).

Both streams are mapped into the discrete latent space with a LoRA-adapted VAE encoder $\mathcal{E}_{\mathrm{lora}}$:

$$z_{reg} = \mathcal{E}_{\mathrm{lora}}(x_{reg}), \qquad z_{lq} = \mathcal{E}_{\mathrm{lora}}(x_{lq}^{\uparrow}). \tag{7}$$

Thus z_{reg} specifies the geometry to preserve, while z_{lq} is neither another structural prediction nor an explicit blur or noise parameter. Following Spatial Feature Transform Wang et al. (2018), Latent SFT applies a residual affine modulation on the anchor latent,

$$z_{fused} = z_{reg} + \gamma(z_{lq}) \odot z_{reg} + \beta(z_{lq}), \tag{8}$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are convolutional scale and shift maps. The residual form keeps z_{reg} as the manifold coordinate and lets z_{lq} control local amplitude and bias, which is less likely than naive concatenation or ControlNet-style side branches to pull the state off the structured VAE geometry of the pre-trained prior. Manifold preservation is carried by this residual parameterization together with LRR and the Stage-2 training constraints.

We encourage z_{fused} to occupy a schedule-compatible noisy state of the clean latent $z_{hq} = \mathcal{E}_{\text{fixed}}(x_{hq})$ at the calibrated timestep $\hat{t}$ via Latent Representation Refinement (LRR) Wu et al. (2025):

$$z_{fused} \approx z_{\hat{t}} = \sqrt{\bar{\alpha}_{\hat{t}}} z_{hq} + \sqrt{1-\bar{\alpha}_{\hat{t}}}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{9}$$

LRR pins the fused state to the working point selected in Sec. 3.4, so that a one-step noise predictor ϵ_θ operates in a regime familiar to the pre-trained denoiser. A LoRA-adapted UNet then performs one-step noise prediction on z_{fused} with an empty text prompt $\emptyset$, and the frozen VAE decoder $\mathcal{D}_{\text{fixed}}$ yields $\hat{x}_{hq} = \mathcal{D}_{\text{fixed}}(\hat{z}_{hq})$.

3.4 First-Order SNR Calibration

Multi-step reverse sampling is expensive for restoration, yet a poorly chosen single timestep either under-refines texture or overwrites the structural anchor. OMGSR Wu et al. (2025) shows that a single mid-timestep with LRR can be effective when a low-quality latent is treated as a noisy state. FaceDNO adopts this one-step setting, but the conditioning state is now the *fused* latent z_{fused}: observation-conditioned modulation changes its effective noise structure, so a timestep chosen for a raw low-quality encoding need not remain appropriate.

Encoding and fusion also introduce an amplitude bias relative to z_{hq}. If one estimates an empirical SNR from unscaled residuals, that bias is absorbed into the residual term and systematically shifts the matched timestep. We therefore first remove a training-set scalar gain. The diffusion forward process defines

$$z_t = \sqrt{\bar{\alpha}_t} z_{hq} + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{10}$$

with schedule SNR $\mathrm{SNR}_{\text{diff}}(t) = \bar{\alpha}_t \widehat{\mathrm{Var}}(z_{hq})/(1-\bar{\alpha}_t)$. Using the training-set scale $s^\star$ from Sec. 3.1, the empirical fused SNR is

$$\widehat{\mathrm{SNR}}_{fused} = \frac{(s^\star)^2 \widehat{\mathrm{Var}}(z_{hq})}{\widehat{\mathrm{Var}}(z_{fused} - s^\star z_{hq})}. \tag{11}$$

We then select a discrete schedule-calibrated timestep

$$\hat{t} = \arg\min_{t \in \mathcal{T}} \left| \mathrm{SNR}_{\text{diff}}(t) - \widehat{\mathrm{SNR}}_{fused} \right|. \tag{12}$$

The construction is a dataset-level first-order amplitude calibration followed by nearest-neighbor matching on the diffusion schedule: equal SNR does not imply equal conditional distributions, and $\hat{t}$ is a working-point choice rather than a proof of identifiability (Sec. 5). The timestep is estimated from training statistics and then fixed for Stage-2 training and inference; held-out residual dependence and timestep scans are reported in Sec. 4.4. Derivation details of $s^\star$ are deferred to Appendix B.

Following the one-step predictor of OMGSR Wu et al. (2025), we treat z_{fused} as the noisy state at $\hat{t}$ and recover

$$\hat{z}_{hq} = \frac{z_{fused} - \sqrt{1-\bar{\alpha}_{\hat{t}}}\,\epsilon_\theta(z_{fused}, \hat{t}, \emptyset)}{\sqrt{\bar{\alpha}_{\hat{t}}}}. \tag{13}$$

Training and inference both lock this fixed $\hat{t}$, turning Stage-2 into a one-step refinement around a schedule-compatible working point.

3.5 Training Objectives

FaceDNO is trained in two phases that mirror the mechanism split above. Stage-1 optimizes FRONet with reconstruction losses so that x_{reg} provides a stable geometric base. Stage-2 freezes Stage-1 and jointly trains Latent SFT and the LoRA-adapted diffusion UNet at the calibrated timestep $\hat{t}$.

Table 1: Quantitative comparison on CelebA-Test Karras et al. (2018). CodeFormer Zhou et al. (2022b) and DiffBIR Lin et al. (2023) are newly retrained and evaluated under the same protocol as FaceDNO. Best and second-best results are highlighted in red and blue. Inference time is omitted when not measured under the same timer.

Methods	Venue	Inference time	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	NIQE↓	FID↓
SwinIR Liang et al. (2021)	*ICCV21'*	0.486	25.51	0.6589	0.3347	0.3130	42.07	8.90	46.38
RestoreFormer Wang et al. (2022)	*CVPR22'*	0.273	24.48	0.6157	0.2742	0.3419	46.00	8.16	43.22
CodeFormer Zhou et al. (2022b)	*NeurIPS22'*	0.551	23.72	0.6310	0.3197	0.5548	72.85	4.72	24.31
GCFSR He et al. (2022)	*CVPR22'*	0.1286	24.48	0.6912	0.2326	0.6101	70.67	4.94	23.65
VQFR Gu et al. (2022)	*ECCV22'*	0.3047	24.06	0.6527	0.2458	0.6683	74.29	4.61	19.93
DiffBIR Lin et al. (2023)	*arXiv23'*	89.863	22.96	0.6071	0.3244	0.5957	71.68	6.41	28.84
BFRffusion Chen et al. (2024)	*TCSVT24'*	17.374	24.56	0.6629	0.353	0.6732	71.69	4.90	10.49
PGDiff Yang et al. (2023) (multi-step)†	*NeurIPS23'*	152.441	23.36	0.6702	0.2975	0.5267	69.54	4.72	12.44
DynFace Do et al. (2025) (multi-step)†	*ICCV25'*	125.328	23.08	0.6583	0.2910	0.6908	71.66	4.23	11.91
OSDFace Wang et al. (2025) (single-step)†	*CVPR25'*	0.685	20.05	0.5623	0.3405	0.6810	75.51	3.92	14.49
FaceDNO (single-step)†	–	0.821	24.67	0.6743	0.2136	0.7422	76.84	3.51	8.72

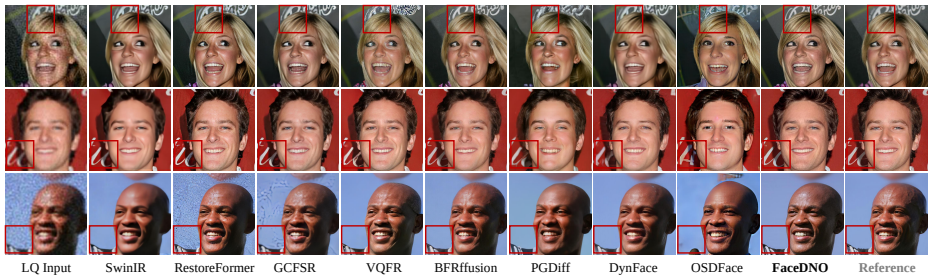

Figure 2: Qualitative comparison on synthetic faces in CelebA-Test Karras et al. (2018). FaceDNO recovers sharper facial contours and more coherent eye/mouth textures under heavy degradation. (**Zoom in for details**)

The primary Stage-2 latent objective is LRR on the fused state Wu et al. (2025):

$$\mathcal{L}_{\mathrm{LRR}} = \mathbb{E}_{x_{hq},\epsilon}\left[\left\| z_{fused} - \left(\sqrt{\bar{\alpha}_{\hat{t}}} z_{hq} + \sqrt{1-\bar{\alpha}_{\hat{t}}}\,\epsilon\right)\right\|_2^2\right]. \tag{14}$$

This ℓ_2 alignment pins z_{fused} to a sampled noisy state at $\hat{t}$, i.e., it supervises the schedule working point rather than pixel appearance. The decoded image $\hat{x}_{hq}$ is further supervised by

$$\mathcal{L}_{\mathrm{img}} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{Dv3D}\mathcal{L}_{Dv3D} + \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{id}\mathcal{L}_{id} + \lambda_{lq}\mathcal{L}_{lq_cons}, \tag{15}$$

where $\mathcal{L}_{L1}$ is the main image-space fidelity term and the remaining losses provide perceptual, adversarial, identity, and observation-consistency cues at smaller relative weight. The full Stage-2 objective is $\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{LRR}}\mathcal{L}_{\mathrm{LRR}} + \mathcal{L}_{\mathrm{img}}$. Detailed loss forms and weights are given in Appendix C.

4 Experiments

4.1 Experimental Setup

Datasets. We train on FFHQ Karras et al. (2019) (512×512) with LQ-HQ pairs synthesized by the Real-ESRGAN degradation pipeline Wang et al. (2021c), following recent face restoration protocols Wang et al. (2025); Liang et al. (2025). Evaluation uses synthetic CelebA-Test Karras et al. (2018) (3,000 images under the same pipeline) and real-world LFW-Test Huang et al. (2008), WebPhoto-Test Wang et al. (2021b), and Wider-Test Zhou et al. (2022a), spanning mild to severe degradations Zhou et al. (2022b); Wang et al. (2025).

Metrics. On CelebA-Test we report PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018a), CLIPIQA Wang et al. (2023), MUSIQ Ke et al. (2021), NIQE Zhang et al. (2015), and FID Heusel et al. (2017) to the HQ references. On real-world sets without paired HQ faces, we follow OSDFace Wang et al. (2025) and report FID, CLIPIQA, MUSIQ, and NIQE.

Implementation. FaceDNO is implemented in PyTorch and trained in two stages on NVIDIA A100 GPUs. Stage-1 supervision uses non-uniform continuous degradation-scale sampling on

Table 2: Quantitative comparison on real-world datasets (LFW-Test Huang et al. (2008), WebPhoto-Test Wang et al. (2021b), and Wider-Test Zhou et al. (2022a)) with non-diffusion, multi-step diffusion, and one-step diffusion methods. The best and second best results are colored in red and blue.

Methods	Wider-Test				LFW-Test				WebPhoto-Test			
	C-IQA↑	MUSIQ↑	NIQE↓	FID↓	C-IQA↑	MUSIQ↑	NIQE↓	FID↓	C-IQA↑	MUSIQ↑	NIQE↓	FID↓
SwinIR Liang et al. (2021)	0.5492	63.33	4.71	69.98	0.5945	61.20	4.88	80.44	0.6050	64.48	4.00	134.07
RestoreFormer Wang et al. (2022)	0.5182	61.47	4.57	75.86	0.6083	64.38	4.86	85.80	0.5890	60.06	4.12	140.76
GCFSR He et al. (2022)	0.5990	69.62	4.11	58.75	0.6892	69.44	4.03	73.34	0.6418	67.77	4.91	120.18
VQFR Gu et al. (2022)	0.6989	70.14	3.90	60.01	0.6943	70.83	3.88	80.57	0.6605	70.68	4.93	107.47
BFRffusion Chen et al. (2024)	0.5964	69.13	3.90	55.82	0.6275	71.44	3.99	75.36	0.5994	67.59	5.00	110.94
PGDiff Yang et al. (2023)	0.5804	67.90	3.83	67.09	0.6012	70.61	4.04	87.12	0.5733	65.11	5.22	125.03
DynFace Do et al. (2025)	0.6884	70.32	5.59	50.34	0.7294	71.42	3.1996	70.32	0.6598	68.59	4.57	117.23
OSDFace Wang et al. (2025)	0.7045	73.25	4.07	47.22	0.7143	73.27	3.9104	68.44	0.6986	71.41	4.95	107.62
FaceDNO (ours, s=1)†	0.7424	74.69	3.71	37.67	0.7754	76.32	3.5011	60.09	0.7169	73.47	4.35	102.01



Figure 3: Qualitative comparison on real-world faces. Top: WebPhoto-Test Wang et al. (2021b). Middle: Wider-Test Zhou et al. (2022a). Bottom: LFW-Test Huang et al. (2008). FaceDNO better preserves geometry under non-uniform blur and compression. (**Zoom in for details**)

$(s_{\min}, s_{\max}]$: each minibatch draws $u \sim \mathcal{U}(0,1)$ and sets $s = s_{\min} + (s_{\max} - s_{\min})\, u^{1/p}$ with $s_{\min} = 1$, $s_{\max} = 4$, and $p = 2.5$, concentrating mass near larger downsampling factors. Ablations over p are in Appendix E (Table 10). FRONet is trained from scratch with an RDN Zhang et al. (2018b) encoder (no upsampling), Adam at 4×10^{-5} for 400 epochs, 10-epoch linear warmup, batch size 4 per GPU on 4 GPUs, and L1 reconstruction as the primary loss. Stage-2 freezes FRONet and adapts Stable Diffusion 2.1-base Rombach et al. (2022) with LoRA on the VAE encoder (rank 16) and UNet (rank 32); Latent SFT uses 32 hidden channels, and the dataset-level SNR-calibrated timestep is fixed at $\hat{t} = 312$. Optimization uses AdamW Loshchilov & Hutter at 4×10^{-5}, a constant scheduler with 500 warmup steps, 10,000 steps, `bf16`, and batch size 1 per GPU. Loss weights are $\lambda_{LRR} = 5.0$, $\lambda_{Dv3D} = 5.0$ (DINOv3-ConvNeXt-Large), $\lambda_{L1} = 0.5$, $\lambda_{GAN} = 0.5$, $\lambda_{id} = 0.5$ (ArcFace ResNet-18), and $\lambda_{lq_cons} = 0.3$.

4.2 Comparisons on Synthetic and Real Benchmarks

We compare FaceDNO with reconstruction baselines Liang et al. (2021); Wang et al. (2022), codebook/GAN restorers He et al. (2022); Gu et al. (2022); Zhou et al. (2022b), and diffusion restorers Yang et al. (2023); Chen et al. (2024); Lin et al. (2023); Do et al. (2025); Wang et al. (2025). CodeFormer Zhou et al. (2022b) and DiffBIR Lin et al. (2023) are newly retrained under the same CelebA-Test degradation and metric scripts as FaceDNO.

Synthetic results. Table 1 shows that FaceDNO improves over the matched one-step baseline OSDFace Wang et al. (2025) on both fidelity and perceptual metrics (e.g., PSNR 24.67 vs. 20.05, FID 8.72 vs. 14.49), and also surpasses protocol-matched CodeFormer and DiffBIR on all seven reported scores. Relative to multi-step diffusion restorers such as PGDiff and DynFace, FaceDNO remains competitive in FID and no-reference IQA while using a single refinement step instead of hundreds of reverse iterations. Pure reconstruction methods such as SwinIR retain high PSNR but lag far behind on LPIPS/FID, illustrating the familiar fidelity–perception trade-off that FaceDNO targets with dual-role conditioning. Fig. 2 shows clearer eyelid and mouth boundaries and fewer stochastic artifacts than unconstrained multi-step diffusion sampling under severe synthetic degradations.

Table 3: **System attribution** on CelebA-Test. Minimal uses $\mathcal{L}_{L1} + \mathcal{L}_{\mathrm{LRR}}$ only.

Method	PSNR↑	FID↓
OSDFace Wang et al. (2025)	20.05	14.49
OSDFace + FRONet	22.47	12.26
FaceDNO-Minimal	24.82	10.25
FaceDNO-Full	24.67	8.72

Table 4: **Quality–latency Pareto** under a matched timer.

Method	Time (s)↓	FID↓
OSDFace Wang et al. (2025)	0.685	14.49
FaceDNO-Fast	0.649	9.45
FaceDNO-Full	0.821	8.72

Table 5: **A1: Stage-1 structural anchor** under the same Stage-2. Bicubic uses only PSNR/LPIPS/FID from the matched protocol.

Stage-1 variant	PSNR↑	SSIM↑	LPIPS↓	ArcFace Deg.↓	CLIPIQA↑	FID↓
Bicubic upsampling	21.37	–	0.3326	–	–	15.08
SwinIR-style anchor	24.43	0.6782	0.2435	36.62	0.7109	10.03
FRONet w/o HF gated branch	24.02	0.6670	0.2244	34.56	0.7302	9.45
FRONet (full)	**24.67**	**0.6743**	**0.2136**	**28.44**	**0.7422**	**8.72**

Real-world results. Table 2 evaluates unpaired restoration on LFW-Test, WebPhoto-Test, and Wider-Test. FaceDNO improves no-reference perceptual scores and FID over the matched one-step baseline OSDFace Wang et al. (2025) on all three sets, while remaining competitive with multi-step diffusion restorers at a single refinement step. Fig. 3 shows that facial contours and eye/mouth regions remain more coherent under non-uniform blur or compression, consistent with using FRONet as a structural conditioner for one-step refinement.

4.3 System Attribution and Quality–Latency Trade-off

A natural alternative explanation is that FaceDNO merely swaps in a stronger Stage-1 backbone or stacks more losses. Table 3 isolates this concern on CelebA-Test under the same evaluation script. OSDFace+FRONet replaces the OSDFace structural pathway with our Stage-1 anchor while retaining the OSDFace Stage-2 recipe. FaceDNO-Minimal keeps FRONet, Latent SFT, and one-step refinement at $\hat{t}$, but trains Stage-2 with only $\mathcal{L}_{L1} + \mathcal{L}_{\mathrm{LRR}}$ (no GAN/identity/auxiliary terms). FaceDNO-Full uses the complete Stage-2 objective.

Replacing the OSDFace backbone with FRONet improves FID from 14.49 to 12.26, yet remains behind FaceDNO-Minimal (10.25). Minimal already exceeds both OSDFace variants; Full further improves FID to 8.72 at a slight PSNR cost, reflecting the usual perception–distortion trade-off. We therefore attribute the main gain to the two-stage conditioning and alignment design.

Under a matched timer, Stage-1 costs 0.352 s and Stage-2 0.469 s (total 0.821 s). Halving FRONet channels yields FaceDNO-Fast (0.649 s, FID 9.45), which is faster than OSDFace (0.685 s, FID 14.49) while remaining substantially stronger in FID (Table 4). Full FaceDNO trades additional latency for the best FID among these operating points.

4.4 Does x_{lq} contribute beyond x_{reg}, and does SNR calibration help?

Q1: Stage-1 anchor. Table 5 compares Bicubic, a SwinIR-style discrete anchor, FRONet without the HF gate, and full FRONet under the same Stage-2. Learned anchors beat interpolation by a large margin in FID and LPIPS; removing the HF gated branch raises FID from 8.72 to 9.45 and ArcFace degree from 28.44 to 34.56. Full FRONet is strongest among the tested Stage-1 variants.

Q2: Observation residual. Table 6 removes x_{lq}/Latent SFT and feeds only the encoded anchor to one-step refinement. Anchor-only drops PSNR by 0.85 dB and worsens LPIPS/FID, supporting the dual-role view that z_{lq} supplies residual cues suppressed by Stage-1 normalization. Table 8 further compares fusion mechanisms; Latent SFT configurations outperform concatenation, additive fusion, and ControlNet-style branches.

Q3: Timestep and first-order SNR. Table 7 shows that the training-set calibrated $\hat{t} = 312$ matches or beats fixed heuristics and is close to a validation grid-search optimum ($t = 320$). On held-out

Table 6: **Anchor-only vs. full conditioning**.

Stage-2 input	PSNR↑	LPIPS↓	FID↓
Anchor only	23.82	0.2854	11.80
Full (Latent SFT)	**24.67**	**0.2136**	**8.72**

Table 7: **A3: Timestep selection.** SNR-calibrated $\hat{t}$ vs. heuristics.

Strategy	t	PSNR↑	LPIPS↓	CLIPIQA↑	FID↓
Fixed early	100	24.05	0.2491	0.7016	10.07
Fixed middle	250	24.42	0.2288	0.7269	9.19
Fixed late	500	23.91	0.2576	0.6887	10.45
Grid-search best	320	24.61	0.2179	0.7388	8.91
SNR-calibrated	**312**	**24.67**	**0.2136**	**0.7422**	**8.72**

Table 8: **A2: Latent guidance strategy** for fusing z_{reg} and z_{lq}.

Guidance / fusion scheme	CLIPIQA↑	MUSIQ↑	NIQE↓	ArcFace Deg.↓	FID↓
Concat + Conv projection	0.6984	72.35	3.93	35.35	10.21
Additive fusion ($z_{reg} + z_{lq}$)	0.7112	73.41	3.84	30.43	9.73
ControlNet-style side branch	0.7246	74.68	3.72	30.76	9.26
Latent SFT	0.7358	75.91	3.61	29.90	8.98
Latent SFT (full)	**0.7422**	**76.84**	**3.51**	**28.44**	**8.72**

CelebA-Test, the absolute Pearson correlation between z_{hq} and the residual falls from 0.37 (naive $s = 1$) to 0.14 after OLS scaling with $s^\star \approx 0.71$.

OOD corruptions. Under extreme motion blur and severe noise (protocol in Appendix A), FaceDNO-Full attains FID 15.06 and 13.48 versus OSDFace FID 28.67 and 25.33 (Table 9).

5 Limitations

Dataset-level SNR calibration is a first-order schedule match under training statistics: it is not an identifiability proof, and equal SNR does not imply equal conditional distributions; a fixed $\hat{t}$ may also need recomputation under substantially different degradations. We have not yet run a human preference study separating realism, identity faithfulness, and local artifacts, and extremely low-SNR inputs can still fail on irises, teeth, skin, or identity. The study is face-only under the reported benchmarks (no trained non-face transfer claim) and provides no video temporal consistency guarantee.

6 Conclusion

This paper studies one-step blind face restoration as *conditional latent construction*: under a single diffusion refinement step, the conditioning state must recover scale-flexible structure, retain observation-specific residual cues, and sit at a schedule-compatible noise level. We propose FaceDNO, a two-stage formulation that addresses these requirements separately. A Face Resolution Neural Operator builds a scale-flexible structural anchor; manifold-preserving Latent SFT then fuses the encoded anchor with the upsampled observation so that geometry and residual modulation play complementary roles; and dataset-level first-order SNR calibration selects a one-step working point for the fused latent, building on mid-timestep refinement and LRR. Matched synthetic and real-world benchmarks, together with attribution and Stage-2 factorizations, show that dual-role conditioning outperforms anchor-only inputs, that SNR calibration predicts a useful timestep among the tested candidates, and that FaceDNO-Fast/Full offer a concrete quality–latency trade-off. A remaining scope limitation is extremely low-SNR or heavily unconstrained in-the-wild inputs, including tiny faces, strong occlusions, and large non-frontal poses, where local failures around irises, teeth, or identity can still appear. Extending the framework with adaptive, uncertainty-aware refinement for such extreme cases—and complementing automatic metrics with human preference studies—is an important next step.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks that require disclosure under the ICLR 2027 AI Policy for Authors, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, writing proofs, proposing or refining research hypotheses, designing or providing feedback on research methodology or experiments, implementing the proposed methods, translating the manuscript, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting experimental results. Additionally, we used generative AI tools for recommended-disclosure tasks: drafting and editing manuscript text to improve readability, suggesting phrasing and section structure, assisting with LaTeX formatting and reference formatting, summarizing publicly available related work to support literature organization, and helping prepare analysis or visualization scripts. We have reviewed all AI-assisted work: AI-suggested text and references were checked against primary sources; quantitative claims and tables were verified against experimental logs; and any AI-assisted scripts were inspected and tested by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies blind face restoration on publicly available face image benchmarks (FFHQ for training; CelebA-Test, LFW-Test, WebPhoto-Test, and Wider-Test for evaluation) and does not collect new human-subject data or personally identifiable information beyond what these datasets already contain. Training pairs are synthesized with a standard degradation pipeline rather than by collecting new private photographs. Face restoration methods can be misused to enhance deceptive media or to alter identity cues; we therefore report identity-related metrics where applicable and emphasize that FaceDNO is intended for research and legitimate restoration settings. The reported results concern restoration quality under the specified protocols and do not establish safety guarantees for downstream biometric or security-critical applications. The authors adhered to the ICLR Code of Ethics and are responsible for the scientific claims, methodology, experiments, and presentation of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe datasets, metrics, and training hyperparameters (scale sampling, optimizers, LoRA ranks, loss weights, and $\hat{t} = 312$) in Section 4.1. The conditional-latent formulation, Latent SFT, and dataset-level first-order SNR calibration are detailed in Sections 3.1–3.5. Appendix B derives the least-squares scaling factor $s^\star$ used for SNR calibration, and Appendix C expands the Stage-2 loss terms. Additional ablations and qualitative examples are provided in Appendices E and D. Complete experiment logs, configurations, and checkpoints are retained on the remote training servers and will be released together with anonymized source code as supplementary material.

REFERENCES

Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22367–22377, 2023.

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, 2024.

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.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pp. 126–143. Springer, 2022.

Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1889–1898, 2022.

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. *Advances in neural information processing systems*, 30, 2017.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

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*, pp. 5148–5157, 2021.

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*, pp. 399–415. Springer, 2020.

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*, pp. 9842–9851, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Xinqi Lin, Jingwen He, Zhaoyang Chen, Zhaoyang Lyu, Chao Dong, and Yuchao Dai. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

Xiaoyi Liu and Hao Tang. Difffno: Diffusion fourier neural operator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 150–160, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xihaier Luo, Xiaoning Qian, and Byung-Jun Yoon. Hierarchical neural operator transformer with learnable frequency-aware loss prior for arbitrary-scale super-resolution. *arXiv preprint arXiv:2405.12202*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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 Computer Vision and Pattern Recognition Conference*, pp. 12626–12636, 2025.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 606–615. IEEE, 2018.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178, 2021a.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178, 2021b.

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*, pp. 1905–1914, 2021c.

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.

Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17512–17521, 2022.

Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18247–18256, 2023.

Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. *arXiv preprint arXiv:2508.08227*, 2025.

Zihao Xu, Yuzhi Tang, Bowen Xu, and Qingquan Li. Neurop-diff: Continuous remote sensing image super-resolution via neural operator diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12491–12501, 2025.

Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36:32194–32214, 2023.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 672–681, 2021.

Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022.

Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018b.

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

Shangchen Zhou, Kelvin CK 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, 2022b.

A OOD Corruption Protocol and Results

We evaluate two controlled out-of-distribution corruptions on CelebA-Test-style images: extreme motion blur and severe noise. FaceDNO remains substantially stronger than OSDFace Wang et al. (2025) in both settings (Table 9). Detailed kernel/noise parameter ranges will be released with the remote experiment configs.

Table 9: **OOD FID** under extreme motion blur or severe noise.

Method	In-dist. FID↓	Extreme motion blur↓	Severe noise↓
OSDFace Wang et al. (2025)	14.49	28.67	25.33
FaceDNO-Full	8.72	15.06	13.48

B Derivation of the Scaling Factor $s^\star$

In Section 3.4, we treat the fused latent as a scalar-signal plus residual under training-set statistics:

$$z_{fused} = s \cdot z_{hq} + n. \tag{16}$$

We choose s by least squares on the training distribution,

$$s^\star = \arg\min_s \mathbb{E}_{\mathcal{D}_{\text{train}}}\left[\|z_{fused} - s z_{hq}\|_2^2\right], \tag{17}$$

which yields the closed form

$$s^\star = \frac{\mathrm{Cov}(z_{fused}, z_{hq})}{\mathrm{Var}(z_{hq})} \tag{18}$$

when covariances and variances are estimated with the same empirical measure used for $\widehat{\mathrm{SNR}}_{fused}$. Under this scalar projection, the residual $n = z_{fused} - s^\star z_{hq}$ is orthogonal to z_{hq} in the corresponding one-dimensional inner-product sense. See Sec. 5 for the scope of this first-order calibration.

C Detailed Loss Formulations

In this section, we provide the detailed mathematical formulations for the loss functions used during the Stage-2 training of FaceDNO.

Latent Representation Refinement (LRR) Loss. As introduced in the main text, the LRR loss constrains the manifold-preserving fused latent z_{fused} to approximate a sampled noisy state of the clean latent at the SNR-calibrated timestep $\hat{t}$:

$$\mathcal{L}_{LRR} = \mathbb{E}_{x_{hq},\epsilon}\left[\|z_{fused} - (\sqrt{\bar{\alpha}_{\hat{t}}}\mathcal{E}_{fixed}(x_{hq}) + \sqrt{1-\bar{\alpha}_{\hat{t}}}\epsilon)\|_2^2\right] \tag{19}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\mathcal{E}_{fixed}$ is the frozen VAE encoder. This is an ℓ_2 alignment objective in latent space.

Pixel Reconstruction Loss. We employ the standard L1 loss to enforce basic pixel-level fidelity between the generated image $\hat{x}_{hq}$ and the ground truth x_{hq}:

$$\mathcal{L}_{L1} = \|\hat{x}_{hq} - x_{hq}\|_1 \tag{20}$$

DINOv3-ConvNeXt DISTS Loss ($\mathcal{L}_{Dv3D}$). To address the resolution mismatch and enhance structural perception, we adopt a loss inspired by the Dv3CD loss from OMGSR Wu et al. (2025). We extract multi-level features $\{F^l(\cdot)\}$ from a pre-trained DINOv3-ConvNeXt model, which possesses

powerful extrapolation capabilities. The loss computes both feature-level L1 distances and structural similarities (SSIM) across different scales:

$$\mathcal{L}_{Dv3D} = \sum_{l} \left(\alpha_l \|F^l(\hat{x}_{hq}) - F^l(x_{hq})\|_1 + \beta_l (1 - \text{SSIM}(F^l(\hat{x}_{hq}), F^l(x_{hq}))) \right) \tag{21}$$

where α_l and β_l are scale-specific weighting coefficients.

Adversarial Loss ($\mathcal{L}_{GAN}$). To drive the generation of high-frequency photorealistic details, we employ a multi-level discriminator D built upon the frozen DINOv3-ConvNeXt feature extractor, following the architecture design in OMGSR Wu et al. (2025). The generator is trained to fool this discriminator using the non-saturating logistic loss:

$$\mathcal{L}_{GAN} = -\mathbb{E}[\log D(\hat{x}_{hq})] \tag{22}$$

Correspondingly, the discriminator is optimized to distinguish between the generated images and the real high-quality images.

Identity Preservation Loss ($\mathcal{L}_{id}$). For face super-resolution, preserving the identity of the subject is paramount. We utilize a pre-trained ArcFace network R_{id} to extract identity embeddings and compute the cosine similarity loss:

$$\mathcal{L}_{id} = 1 - \frac{R_{id}(\hat{x}_{hq}) \cdot R_{id}(x_{hq})}{\|R_{id}(\hat{x}_{hq})\|_2 \|R_{id}(x_{hq})\|_2} \tag{23}$$

LQ Consistency Loss ($\mathcal{L}_{lq_cons}$). To ensure that the generated high-quality image does not hallucinate structures that contradict the original low-frequency information, we enforce a degradation consistency constraint. We pass the generated image through a degradation pipeline $\mathcal{D}_{deg}$ (e.g., Real-ESRGAN degradation) and compute the L1 distance to the original low-quality input:

$$\mathcal{L}_{lq_cons} = \|\mathcal{D}_{deg}(\hat{x}_{hq}) - x_{lq}\|_1 \tag{24}$$

D Additional Visual Results

This section provides additional qualitative comparisons to complement the quantitative results in the main text. We evaluate representative challenging cases spanning heavy blur, compression artifacts, severe downsampling, mixed real-world degradations, and large pose variations. Across these settings, FaceDNO consistently recovers clearer facial contours and more stable local textures (e.g., eyes, eyebrows, and mouth corners), while maintaining a natural global appearance.

Observed behavior. As shown in Fig. 4, compared with non-diffusion models, diffusion-based methods generally produce more realistic fine-grained details, but may introduce local inconsistency under severe degradations. Multi-step diffusion methods can improve perceptual sharpness, yet often incur unstable detail synthesis and substantially higher latency. Our one-step FaceDNO yields a better fidelity–realism balance: it reduces over-smoothing, suppresses checkerboard-like artifacts, and preserves more identity-relevant cues around facial components.

Failure cases and limitations. We still observe failure modes on extremely tiny faces, heavy occlusions, and cases with large non-frontal poses under very low signal-to-noise conditions. Typical artifacts include slight identity drift and occasional texture repetition in highly ambiguous regions. These observations motivate future work on adaptive uncertainty-aware refinement and stronger identity constraints for extremely hard in-the-wild samples.

E Additional Ablation Studies

A4: Non-uniform degradation-scale sampling. Using the inverse-transform map in Sec. 4.1, we compare uniform sampling on $(1, 4]$ with power-law sampling at different exponents p. As shown in Table 10, moderate bias ($p = 2.5$) best balances hard-case robustness and overall perceptual quality; overly aggressive bias can slightly hurt mild degradations.



Figure 4: Qualitative comparison on synthetic faces in CelebA-Test set Karras et al. (2018).Our FaceDNO is able to restore high-quality faces, showing robustness to the heavy degradation. (**Zoom in for details**)

Table 10: **A4: Ablation on Stage-1 scale sampling.** Comparison of uniform and power-law sampling strategies.

Sampling strategy	p	PSNR↑	LPIPS↓	CLIPIQA↑	NIQE↓	FID↓
Uniform on $(1, 4]$	–	24.39	0.2287	0.7249	3.78	9.34
Power-law	1.5	24.52	0.2211	0.7334	3.67	8.98
Power-law	2.0	24.61	0.2175	0.7389	3.58	8.83
Power-law	2.5	24.67	0.2136	0.7422	3.51	8.72
Power-law	3.0	24.58	0.2192	0.7361	3.60	8.89