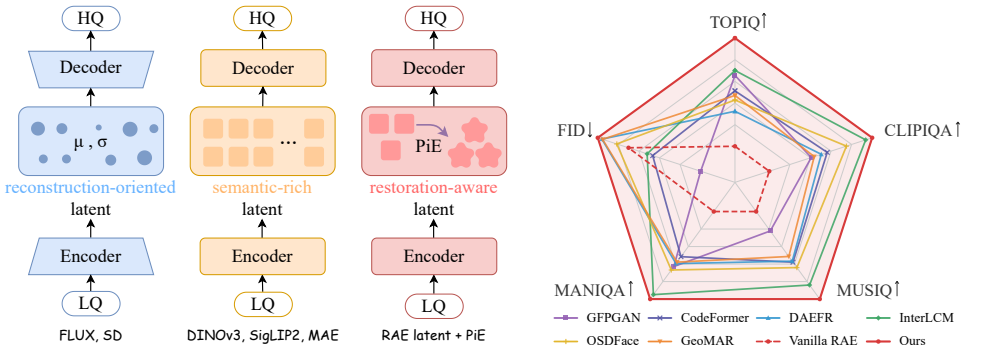
RAEFACE: RETHINKING LATENT REPRESENTATIONS FOR BLIND FACE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most state-of-the-art blind face restoration (BFR) methods, whether codebook- or diffusion-based, rely on VAE-style latent representations. Representation autoencoders (RAEs), built on pretrained visual encoders, have shown strong performance in image generation but remain largely unexplored in restoration-oriented low-level vision. We show that RAE latents better preserve facial semantics under severe degradation and exhibit stronger semantic compatibility between high-quality (HQ) and low-quality (LQ) observations than VAE latents. In light of this, we propose **RAEFace**, a framework that unifies HQ target modeling and LQ condition injection in a shared RAE space to produce semantically faithful base restorations. To seek perceptually stronger yet faithful solutions, we introduce **Perceptual Improvement via Extrapolation (PiE)**, a training-free inference strategy that performs stepwise extrapolation during flow-matching sampling. Using the base prediction as a fidelity anchor, PiE extrapolates beyond it along the direction from the degraded latent to its restored counterpart. Our analysis links the effectiveness of this direction to HQ–LQ semantic compatibility, which helps preserve facial content during extrapolation. Experiments on real-world and synthetic benchmarks demonstrate that RAEFace achieves state-of-the-art perceptual quality with competitive fidelity. Results on inpainting and colorization further suggest the broader potential of RAE representations for related low-level vision tasks.



(a) VAE (b) Vanilla RAE (c) RAEFace (d) Perceptual Quality on Real-World Benchmarks

Figure 1: Rethinking latent representations for BFR. (a) Reconstruction-oriented VAE latents lose facial semantics under severe degradation. (b) Vanilla RAE latents preserves rich semantics but remain perceptually limited. (c) Our RAEFace steers Perceptual Improvement via Extrapolation (PiE) to enhance perceptual quality while anchoring fidelity. (d) RAEFace achieves state-of-the-art average performance on three real-world benchmarks. The FID axis (↓) is reversed so outward is better.

1 INTRODUCTION

Blind face restoration (BFR) aims to recover a high-quality (HQ) face from a low-quality (LQ) observation. In real-world scenarios, the degradation is unknown and complex, often obscuring identity cues, facial attributes, and fine textures. This makes BFR inherently ill-posed, as a single

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.

LQ observation may correspond to multiple plausible HQ reconstructions. Thus, a restoration model must synthesize realistic details while preserving the visible facial information in the input.

Recent BFR methods rely on pretrained generators, codebooks, or diffusion priors (Wang et al., 2021a; Zhou et al., 2022; Tsai et al., 2024; Wang et al., 2025; Li et al., 2025) to recover missing facial content. Much of their design focuses on how to condition these priors on degraded input, while the underlying latent spaces are typically inherited from the generative models. Whether these representations are well suited to face restoration receives less attention. We argue that the latent representation shapes both the HQ target that the model learns to generate and the facial semantic information available in the LQ input. A suitable restoration space should therefore organize facial semantics in HQ targets and retain corresponding semantic evidence under degradation.

This motivates us to examine representation autoencoders (RAEs) (Zheng et al., 2026), whose latent spaces are built on pretrained visual representations. Our analysis shows that RAEs provide richer and more discriminative facial representations in HQ images than variational autoencoders (VAEs) (Rombach et al., 2022), while keeping facial semantics more accessible even under severe degradation. RAE latents also exhibit stronger HQ–LQ semantic compatibility, indicating that corresponding facial semantics remain consistently represented across clean and degraded inputs.

Based on these observations, we introduce **RAEFace**, a simple yet effective framework that uses a shared RAE representation space for target modeling and condition injection. Through conditional flow matching, the model learns to recover semantically organized HQ targets using the facial information retained in LQ inputs. The vanilla RAE-based model yields a semantically faithful base prediction, providing a reliable fidelity anchor but leaving its perceptual quality suboptimal. The inherent ambiguity of BFR nevertheless leaves room to explore nearby latent solutions with richer and more realistic details. We therefore seek a meaningful direction in the shared RAE space that moves beyond the base prediction while preserving the facial content supported by the input.

In the shared RAE space, the degraded latent provides a natural weak reference, while the base prediction supplies the strong output. Their difference defines an input-specific restoration direction. We introduce **Perceptual Improvement via Extrapolation (PiE)**, a training-free inference strategy that performs stepwise extrapolation during flow-matching sampling, which uses the base prediction as a fidelity anchor and extrapolates beyond it along this direction. We find that this simple operation effectively enhances perceptual details in RAE space, whereas applying the same operation in VAE space yields less consistent improvements. Our analysis relates this difference to stronger HQ–LQ semantic compatibility in RAE space, which provides a shared semantic reference that helps preserve facial content as local details change. By directly reusing the degraded latent and the current prediction, PiE requires no additional training and enables controllable refinement at inference time. Experiments on one synthetic and three real-world benchmarks show that RAEFace achieves state-of-the-art perceptual quality with competitive fidelity, highlighting semantically structured latent spaces as an effective foundation for faithful restoration and perceptual enhancement.

2 PRELIMINARIES

Representation Autoencoders. Representation Autoencoders (Zheng et al., 2026) replace the VAE encoder with a pretrained vision encoder and learn a decoder for image reconstruction. Let E_ϕ be an L-layer vision encoder. For an input image x, its layer-ℓ feature map is denoted by $z_0^\ell = E_\phi^\ell(x) \in \mathbb{R}^{N\times d}$, where N is the number of tokens and d is the latent dimension. Following RAEv2 (Singh et al., 2026), the RAE latent aggregates features from the last K layers as

$$z_0 = \sum_{\ell=L-K+1}^{L} z_0^\ell. \tag{1}$$

Since generative latents may be noisy and deviate from the clean encoder latents, directly decoding them can mismatch the decoder's training distribution. We therefore use noise-augmented latents $\tilde{z}_0 = z_0 + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, where σ controls the perturbation magnitude. Then, the ViT decoder reconstructs $\hat{x} = D_\theta(\tilde{z}_0)$ by minimizing

$$\mathcal{L}_{\text{rec}}(x, \hat{x}) = \mathcal{L}_{\text{pix}}(x, \hat{x}) + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}(x, \hat{x}) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(\hat{x}), \tag{2}$$

where $\mathcal{L}_{\text{pix}}$ is the pixel-wise ℓ_1 loss, and $\mathcal{L}_{\text{perc}}$ and $\mathcal{L}_{\text{adv}}$ are the perceptual (Zhang et al., 2018) and adversarial (Goodfellow et al., 2014) terms.

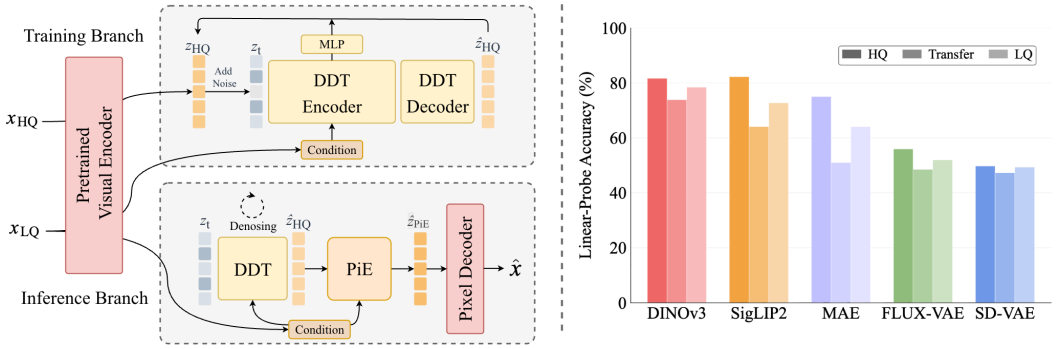


Figure 2: Overview of RAEFace. *Left*: RAEFace training and inference pipeline. *Right*: Linear probe accuracy across different representations.

Flow Matching. Following (Lipman et al., 2022), given a data sample z_0 and noise $z_1 \sim \mathcal{N}(0, I)$, the noise-to-data path is defined by linear interpolation

$$z_t = (1-t)z_0 + tz_1, \quad t \in [0,1], \tag{3}$$

with constant velocity $\mathrm{d}z_t/\mathrm{d}t = z_1 - z_0$. A diffusion transformer (Peebles & Xie, 2023) $v_\theta(z_t, t)$ is trained to predict this velocity by

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0, z_1, t}\left[\|v_\theta(z_t,t) - (z_1 - z_0)\|_2^2\right]. \tag{4}$$

For high-dimensional RAE latents, the time schedule is adjusted to the effective data dimension via the standard schedule-shifting strategy (Esser et al., 2024):

$$t_m = \frac{\alpha t_n}{1 + (\alpha - 1)t_n}, \qquad \alpha = \sqrt{\frac{m}{n}}, \tag{5}$$

where m and n denote the latent and reference dimensions. At inference, we initialize $z_1 \sim \mathcal{N}(0, I)$ and solve $\mathrm{d}z_t/\mathrm{d}t = v_\theta(z_t, t)$ from $t = 1$ to $t = 0$.

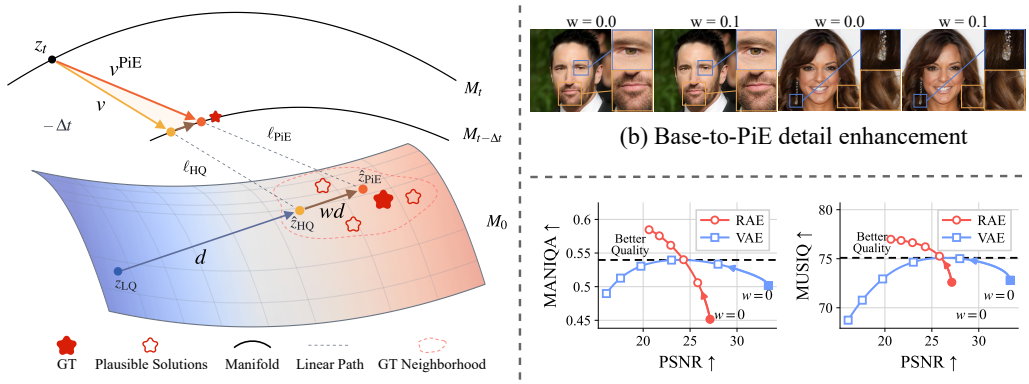
3 Method

3.1 RAEFace Overview

Existing BFR methods mainly study how to constrain the generative prior on the LQ observation, while treating the latent space as a fixed interface (Zhou et al., 2022; Wang et al., 2025; Li et al., 2025). Yet the latent representation itself receives less attention. We argue that it is a vital part of the BFR design because it determines both the HQ target learned by the model and the semantic cues unearthed from the LQ inputs. We hypothesize that a suitable restoration latent should organize facial semantics in HQ targets and preserve corresponding semantic evidence after degradation.

We introduce **RAEFace**, a simple yet effective framework that revisits blind face restoration from the perspective of latent representations. As illustrated in Fig. 2 *Right*, we observe that RAEs built on pretrained encoders such as DINO and SigLIP yield richer and more discriminative latent representations than the VAEs employed in generative models such as FLUX-VAE and SD-VAE. Specifically, RAE latents achieve higher facial-attribute prediction accuracy than VAE latents on both HQ and LQ images, making facial attributes more accessible even under severe degradation. Leveraging this property, RAEFace maps degraded inputs and HQ targets into a shared RAE space for condition injection and target modeling, respectively. This formulation produces a high-fidelity base prediction with well-preserved facial semantics.

Despite its semantic fidelity, the base prediction may lack realistic details because semantic alignment alone does not guarantee high perceptual quality. Since BFR is inherently ill-posed, a single LQ observation may correspond to multiple plausible HQ reconstructions, and any semantically correct and perceptually realistic solution can serve as a valid target without exactly matching the unknown GT. Based on this insight, we introduce **Perceptual Improvement via Extrapolation (PiE)** to



(a) Conceptual illustration of PiE

(b) Base-to-PiE detail enhancement

(c) Peception-Fidelity Tradeoff

Figure 3: Perceptual Improvement via Extrapolation (PiE) in the shared RAE space.

explore perceptually improved solutions in the shared RAE space. PiE constructs a restoration direction from the difference between the degraded and restored latents, and then extrapolates beyond the base prediction to **progressively and stably enhance perceptual details**, while retaining the base restoration as a fidelity anchor. Our analysis attributes the effectiveness of this direction to the semantic compatibility between HQ and LQ representations in the RAE space. PiE is simple, controllable, and training-free, enabling flexible perceptual enhancement at inference time.

3.2 Establishing a Fidelity Anchor in a Shared RAE Space

Conditional Flow Matching in the Shared RAE Space. To establish a semantically reliable fidelity anchor, we formulate BFR as conditional flow matching in a shared RAE space, using HQ latents as restoration targets and LQ latents as conditions. During training, the shared encoder E maps both the degraded input x_{LQ} and its paired HQ target x_{HQ} into the same latent space, i.e., $z_{\text{HQ}} = E(x_{\text{HQ}})$ and $z_{\text{LQ}} = E(x_{\text{LQ}})$. As illustrated in Fig. 2 *Left*, we project the features of z_{LQ} into the DDT Encoder (Wang et al., 2026a) space and inject them via cross-attention (see Appendix A.1 for details).

Following JiT (Li & He, 2026), given the flow state z_t and time t, the DDT predicts the HQ target directly under the LQ condition as $\hat{z}_{\text{HQ}} = f_\theta(z_t, t; z_{\text{LQ}})$, which, following the flow formulation in Sec. 2, defines the transport velocity

$$v_\theta(z_t, t; z_{\text{LQ}}) = \frac{z_t - \hat{z}_{\text{HQ}}}{t}. \tag{6}$$

We integrate the resulting ODE and decode its endpoint latent with the RAE decoder to obtain a base restoration with well-preserved facial semantics.

Training Objective. During training, we additionally apply a representation alignment objective following REPA (Yu et al., 2024). Intermediate DDT features are projected into the RAE space and aligned with the paired HQ representation, regularizing their token-to-token similarity structure. The final training objective combines flow matching and representation alignment:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda \mathcal{L}_{\text{REPA}}, \tag{7}$$

where λ weights representation alignment, empirically set to 0.5.

3.3 PiE: Perceptual Improvement via Extrapolation

The conditional flow model yields a semantically faithful base-restoration latent $\hat{z}_{\text{HQ}}$ in the shared RAE space, which serves as a fidelity anchor. Latent interpolation offers an effective mechanism to traverse semantic spaces by decoding intermediate points along paths between latent samples, yielding coherent variations (Xu et al., 2026). Weak-to-strong guidance further constructs an inference-time direction from the displacement between weak and strong outputs and extrapolates beyond the strong endpoint (Yuan et al., 2026).

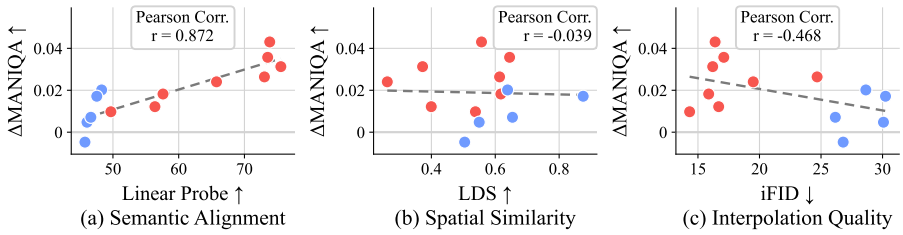


Figure 4: Correlation analysis across RAE and VAE representations, shown in red and blue, respectively. Linear-probe accuracy shows a strong positive correlation with perceptual gains, while LDS and iFID exhibit much weaker or negative correlations.

We adapt this weak-to-strong construction to BFR by treating the degraded latent z_{LQ} as the weak reference and the base prediction $\hat{z}_{\mathrm{HQ}}$ as the strong endpoint. Their displacement, $d = \hat{z}_{\mathrm{HQ}} - z_{\mathrm{LQ}}$, captures the input-specific transformation from the degraded representation to its restored counterpart. In light of this, we introduce **Perceptual Improvement via Extrapolation (PiE)** to enhance BFR, which extrapolates beyond $\hat{z}_{\mathrm{HQ}}$ along this direction to enhance perceptual details while retaining the facial semantics of the base prediction:

$$\hat{z}_{\mathrm{PiE}} = \hat{z}_{\mathrm{HQ}} + w(\hat{z}_{\mathrm{HQ}} - z_{\mathrm{LQ}}). \tag{8}$$

Here, w controls the extrapolation strength. As illustrated in Fig. 3(b), setting $w = 0$ recovers the base restoration, whereas increasing w extends the prediction further along the restoration direction. At each sampling step, $\hat{z}_{\mathrm{PiE}}$ replaces $\hat{z}_{\mathrm{HQ}}$ in the transport velocity computation, thereby steering the sampling trajectory toward a perceptually enhanced solution.

Empirical Results. We next examine whether this restoration direction supports perceptual enhancement across different latent spaces. To exclude the influence of prediction errors, we conduct an oracle study in which $\hat{z}_{\mathrm{HQ}}$ is replaced with the ground-truth latent z_{HQ}. Starting from three representative RAE models and two representative VAE models, we expand the evaluation with additional variants, resulting in eight RAE and five VAE variants in total. Further details are provided in AppendixA.3. Fig. 3(c) presents the group-averaged extrapolation curves. Within a moderate range of extrapolation strengths , the RAE variants consistently achieve higher MANIQA and MUSIQ scores, indicating improved perceptual quality with a relatively gradual loss of fidelity. In contrast, the VAE variants exhibit smaller perceptual gains, larger fidelity losses, and a more pronounced decline in MANIQA and MUSIQ as the extrapolation strength increases (i.e., $w \in [0.6, 1]$). These results suggest that the effectiveness of PiE is closely tied to the semantic organization of the underlying latent representation.

Correlation Analysis. To test whether the benefit of extrapolation is governed by the semantic organization of the latent space, we quantify oracle performance as the MANIQA gain over each model's HQ reconstruction at a matched PSNR loss of 0.4 dB. Following the practice of Zhong et al. (2026), we compare three complementary representative metrics: linear-probe accuracy (trained on HQ latents and evaluated on LQ latents), LDS, and iFID. As shown in Fig. 4, linear-probe accuracy is strongly correlated with the MANIQA gain ($r = 0.872$), while LDS ($r = -0.039$) and iFID ($r = -0.468$) show much weaker correlations. The strong correlation of linear-probe accuracy indicates that semantic accessibility across HQ and LQ latents is the key factor for extrapolating along a meaningful restoration direction. Result of LDS suggests that spatial self-similarity within an individual face, does not directly characterize how facial details change along the restoration direction. While iFID evaluates the quality of interpolations between neighboring faces, such interpolations primarily reflect how smoothly one face can be transformed into another, rather than how a degraded face can be faithfully moved toward a perceptually better version. In a nutshell, these results indicate that HQ–LQ semantic compatibility provides the semantic anchor that governs perceptually beneficial extrapolation.

4 Experiments

4.1 Experimental Settings

Training Details. We train RAEFace on FFHQ (Karras et al., 2019), which contains 70,000 HQ face images. All images are resized to 512×512 pixels, and paired LQ inputs are generated online



Figure 5: Visual comparison on real-world benchmarks in challenging cases. Rows correspond to LFW-Test, WebPhoto-Test, and Wider-Test, respectively. Please zoom in for a better view.

Table 1: Quantitative comparison on real-world benchmarks. FID denotes FID(FFHQ) unless otherwise specified. Red and blue indicate the best and second-best results, respectively.

Type	Method	LFW-Test					WebPhoto-Test					Wider-Test				
		FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
GAN	GFPGAN	47.49	0.719	76.37	0.581	0.953	100.87	0.595	65.89	0.435	0.807	41.35	0.714	74.82	0.555	0.934
VAE	CodeFormer	52.30	0.697	75.47	0.523	0.901	83.91	0.692	74.00	0.503	0.861	39.23	0.698	73.41	0.496	0.888
	DAEFR	47.71	0.697	75.84	0.542	0.909	75.81	0.670	72.70	0.493	0.795	36.82	0.698	74.15	0.521	0.888
	InterLCM	55.36	0.722	75.89	0.578	0.925	78.60	0.751	75.82	0.572	0.881	39.84	0.756	75.41	0.560	0.902
	OSDFace	44.99	0.720	75.35	0.549	0.902	84.98	0.710	73.93	0.516	0.835	34.80	0.728	74.60	0.523	0.887
	GeoMAR	47.03	0.692	75.60	0.540	0.907	76.08	0.667	73.07	0.500	0.851	37.44	0.680	73.15	0.509	0.879
RAE	RAEFace	46.67	0.743	76.65	0.575	0.945	77.54	0.741	76.50	0.566	0.906	34.90	0.769	76.58	0.591	0.949

using the Real-ESRGAN (Wang et al., 2021b) degradation pipeline. The RAE decoder is trained on FFHQ for 40 epochs and kept fixed thereafter. The DDT is initialized with ImageNet-pretrained weights and trained for 50 epochs with a batch size of 32 on four NVIDIA RTX4090 48GB GPUs. We use GMuon (Zhang, 2026) as the optimizer with a learning rate of 2×10^{-4} for the first 25 epochs, followed by linear decay to 2×10^{-5} by epoch 50. We use an Euler ODE sampler with 50 steps. Each step combines both CFG and the proposed PiE, using CFG to establish the fidelity baseline and PiE to further enhance visual details. More sampling details are provided in Appendix A.1.

Evaluation. We evaluate on the synthetic CelebA-Test benchmark (Karras et al., 2017) and three real-world benchmarks: LFW-Test (Huang et al., 2008), WebPhoto-Test (Wang et al., 2021a), and Wider-Test (Yang et al., 2016). On CelebA-Test, we report PSNR, LPIPS (Zhang et al., 2018), IDS (Deng et al., 2019), FID (Heusel et al., 2017) against the HQ references and FFHQ, and four no-reference perceptual metrics: CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and TOPIQ (Chen et al., 2024). For the real-world benchmarks, we report FID(FFHQ) and the same no-reference perceptual metrics.

Baselines. We compare against GFPGAN (Wang et al., 2021a), CodeFormer (Zhou et al., 2022), DAEFR (Tsai et al., 2024), InterLCM (Li et al., 2025), OSDFace (Wang et al., 2025), and GeoMAR (Gan et al., 2026), which represent the state of the art in blind face restoration. All methods are evaluated using their official implementations.

4.2 Main Results

Evaluation on Real-World Benchmarks. As shown in Tab. 1, RAEFace performs consistently well on real-world datasets. It achieves superior performance in the no-reference perceptual comparisons, securing all metrics on Wider-Test, and maintains the second-best FID across LFW-Test and Wider-Test. Additionally, we present the qualitative comparisons in Fig. 5. It is observed that RAEFace demonstrates exceptional robustness to challenging real-world conditions, including subtle facial expressions, old-photo artifacts, occlusions with large poses, while producing visually pleasing and reliable results. Notably, it preserves coherent facial structures and identity, produces high perceptual quality results with rich details.

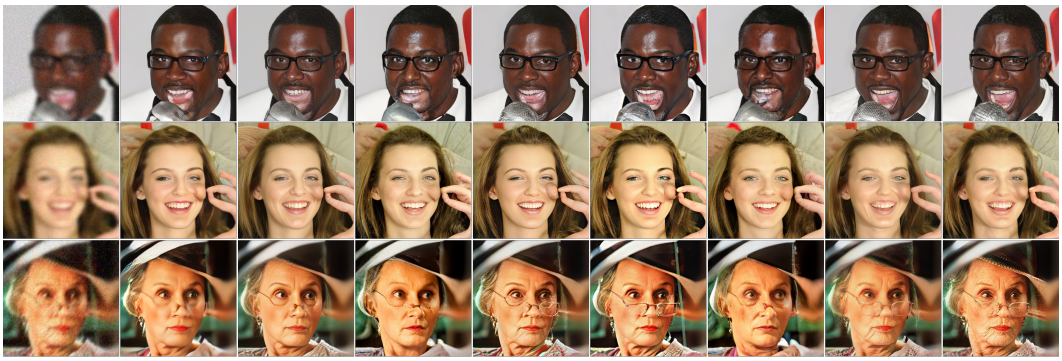


Figure 6: Visual comparison on the synthetic CelebA-Test dataset. Please zoom in for a better view.

Table 2: Quantitative comparison on the synthetic CelebA-Test dataset.

Type	Method	PSNR↑	LPIPS↓	IDS↑	FID(HQ)↓	FID(FFHQ)↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
GAN	GFPGAN	24.98	0.240	0.855	14.32	46.74	0.685	75.29	0.565	0.934
VAE	CodeFormer	25.14	0.227	0.847	14.98	53.10	0.671	75.55	0.537	0.913
	DAEFR	22.65	0.243	0.765	13.10	40.64	0.668	75.27	0.544	0.904
	InterLCM	23.88	0.243	0.873	14.56	51.70	0.718	76.06	0.592	0.925
	OSDFace	23.38	0.239	0.874	11.72	44.18	0.694	75.45	0.549	0.898
	GeoMAR	21.94	0.271	0.700	14.92	39.96	0.658	74.82	0.539	0.898
RAE	RAEFace	24.34	0.226	0.861	10.71	38.59	0.701	76.44	0.573	0.943



Figure 7: Zero-shot results of cartoons.

Evaluation on Synthetic Benchmarks. As shown in Tab. 2, on CelebA-Test, RAEFace achieves the best LPIPS, FID(FFHQ), and FID(HQ) along with leading no-reference perceptual scores, while maintaining competitive reconstruction fidelity. From the visual comparison in Fig. 6, the compared methods often produce visually unappealing restoration results and distort facial structures under challenging shadows. In contrast, RAEFace more naturally preserves eyeglass geometry, shadows cast by hands, and age-related details in elderly faces. We attribute this structural fidelity partly to the semantic and spatial cues captured by the pretrained encoder, which provide global context for distinguishing facial components from occlusions and surrounding objects when local evidence is unreliable. We also test its zero-shot ability on cartoon images in Fig. 7, demonstrating its cross-domain generalization.

4.3 Ablation Study

We ablate the encoder choice, target and condition representations, and guidance mechanism on CelebA-Test. In ablations, all restoration models are trained from scratch on FFHQ for 20 epochs.

Pretrained Encoder Selection. To examine whether the oracle findings extend to learned restoration models with different RAE representations, we compare DINOv3, SigLIP2, and MAE. Since this experiment aims to characterize how different encoders respond to varying extrapolation strengths, rather than to report final benchmark performance, we conduct the PiE sweeps on the first 64 images of CelebA-Test. This subset already reveals clear relative trends and is used consistently across all encoders, with the remaining sampling settings fixed. As shown in Fig. 8, all three encoders benefit from moderate extrapolation, indicating that PiE is effective across different RAE representations. DINOv3 offers a favorable balance between reconstruction fidelity and perceptual quality. We therefore adopt it as the default RAE encoder for the following ablations.

Effect of Target and Condition Representations. Tab. 3 evaluates PiE under different combinations of HQ target and LQ conditioning representations. Here, RAE and VAE denote the DINOv3-based RAE and SD-VAE, respectively. To isolate the effect of PiE, we disable all other guidance mechanisms. Across all three configurations, enabling PiE consistently reduces FID and improves all four no-reference perceptual metrics, demonstrating its broad applicability. Notably, these gains are substantially larger when the restoration target is represented in the RAE space, whereas the VAE target yields only modest improvements. This indicates that the semantic structure of the target space

Table 3: Effect of target and conditioning representations without and with PiE.

Target Rep.	Cond Rep.	Latent Interpolation	PSNR↑	LPIPS↓	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
VAE	RAE	w/o PiE	24.77	0.233	38.45	0.664	73.83	0.539	0.906
		w/ PiE	23.94	0.244	**37.96**	0.693	75.37	0.545	0.912
RAE	VAE	w/o PiE	**25.12**	0.222	43.30	0.590	72.04	0.459	0.864
		w/ PiE	24.28	0.240	40.40	0.701	76.63	0.594	0.953
RAE	RAE	w/o PiE	24.84	**0.220**	42.50	0.601	72.92	0.468	0.875
		w/ PiE	24.04	0.238	39.84	**0.704**	**76.77**	**0.596**	**0.956**

is critical to effective extrapolation. With the RAE target fixed, replacing VAE conditioning with RAE conditioning further improves LPIPS, FID, and all four no-reference metrics, at the cost of only a 0.24 dB decrease (i.e., from 24.28 to 24.04 dB) in PSNR. Consequently, the RAE–RAE configuration with PiE achieves the best scores on all four no-reference metrics while retaining competitive PSNR, LPIPS and FID, providing the strongest overall perception–fidelity balance.

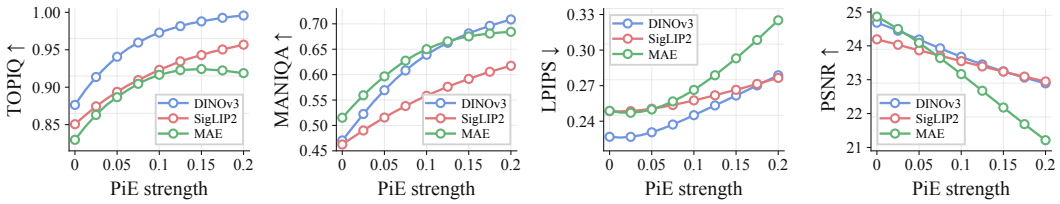


Figure 8: Effect of PiE strength on restoration performance for DINOv3, SigLIP2, and MAE.

Guidance Mechanism. Using the RAE target and RAE condition selected in Tab. 3, we compare five sampling settings in Tab. 4: (i) no guidance, (ii) CFG, (iii) REPA guidance (Singh et al. (2026)), (iv) PiE, and (v) PiE + CFG. CFG achieves the highest pixel-fidelity but provides only marginal perceptual improvements, while REPA guidance yields modest perceptual gains at a larger fidelity cost. In contrast, PiE substantially increases MANIQA from 0.468 to 0.596 and TOPIQ from 0.875 to 0.956 without requiring an additional guidance model. Building on PiE, adding CFG improves PSNR from 24.04 to 24.43 and LPIPS from 0.238 to 0.226, while largely retaining its perceptual gains. We therefore adopt PiE combined with CFG as the default sampling strategy.

Table 4: Ablation on guidance mechanisms.

Guidance	PSNR↑	LPIPS↓	MANIQA↑	TOPIQ↑
w/o Guidance	24.84	0.220	0.468	0.875
CFG	**25.11**	**0.213**	0.471	0.882
REPA Guidance	23.93	0.235	0.482	0.886
PiE Only	24.04	0.238	**0.596**	0.956
PiE + CFG	24.43	0.226	0.593	**0.957**

4.4 Applications

We further validate the generalization of RAEFace on facial colorization and inpainting tasks. Detailed task setup and additional results are provided in A.2.1 and A.2.2.

Face Colorization. For face colorization, we use grayscale faces as the input condition and their corresponding color images as restoration targets. The model recovers missing color information from the facial content preserved in the grayscale input. As shown in Fig. 9 *Left*, RAEFace produces natural facial colors while preserving facial structure and fine details.

Face Inpainting. RAEFace can also be easily extended to face inpainting, demonstrating exceptional restoration capabilities in complex scenarios where key facial structures suffer from local occlusions. As shown in Fig. 9 *Right*, representative methods such as CodeFormer and PixelHacker(Xu et al., 2025) may introduce noticeable structural inconsistencies, including mismatched pupil colors and misaligned eyeglasses lenses, reflecting the difficulty of balancing fine-grained recovery and cross-region consistency. In contrast, benefiting from the powerful semantic prior and structural organization of the RAE latent space, our RAEFace ensures both meticulous detail recovery and rigorous cross-region semantic consistency, effectively preventing facial feature distortions, yielding natural and coherent completions.



Figure 9: Visual comparisons on CelebA-Test. *Left*: facial colorization. *Right*: inpainting.

5 Related Work

Blind Face Restoration with Generative Priors. Generative priors provide facial structure and texture information for BFR under unknown degradations. GAN-based methods such as GFPGAN (Wang et al., 2021a) use pretrained face generators to recover plausible facial content. Codebook-based methods represent HQ faces with discrete visual tokens, following visual tokenization approaches such as VQGAN (Esser et al., 2021). CodeFormer (Zhou et al. (2022)) predicts HQ code sequences from degraded inputs, while DAEFR (Tsai et al., 2024) uses associated encoders to improve the matching between LQ features and codebook entries. GeoMAR (Gan et al., 2026) incorporates geometric conditioning into masked autoregressive code prediction. Diffusion-based approaches use denoising priors for conditional restoration. OSDFace (Wang et al., 2025) conditions a one-step diffusion model on codebook-based visual prompts from the LQ input. InterLCM (Li et al., 2025) treats the LQ image as an intermediate state of a latent-consistency model and learns a transition toward the HQ face. These approaches primarily explore the use of generative priors and their conditioning on degraded inputs. *Our work revisits the latent representation for blind face restoration, examining how semantic organization and HQ–LQ compatibility support restoration and enable perceptual enhancement through latent extrapolation.*

Representation Autoencoders. Recent work explores pretrained visual representations to build semantically structured latent spaces for image generation. RAE (Zheng et al., 2026) and RAEv2 (Singh et al., 2026) train diffusion models directly in these spaces, with learned decoders reconstructing images. FAE (Gao et al., 2026) further explores compact representations through lightweight feature compression. To improve detail reconstruction, SVG (Shi et al., 2026) augments frozen DINO features with a learned residual encoding that captures complementary fine-grained information. DINO-Tok (Jia et al., 2025) combines shallow features containing local textures with deep features encoding global semantics. DecQ (Wang et al., 2026b) introduces learnable queries that gather fine-grained information from intermediate encoder layers, complementing semantic patch tokens for reconstruction and generation. These approaches enrich latent representations with detail information available in the clean source image. However, representation autoencoders have been developed and evaluated primarily for high-quality reconstruction and generation; their behavior on degraded LQ observations and their suitability for restoration remain largely unexplored. This gap is pronounced in BFR where fine details are missing in the degraded input. *Our RAEFace uses the facial semantics retained under degradation to condition restoration and enhances recovered details through PiE, without adding a dedicated branch to extract low-level details from the LQ input.*

6 Conclusion

In this work, we revisit blind face restoration from the perspective of latent representations and introduce **RAEFace**. By modeling HQ targets and injecting LQ conditions in a shared RAE space, RAEFace establishes a semantically reliable fidelity anchor. **Perceptual Improvement via Extrapolation (PiE)** then extrapolates beyond this anchor along the LQ-to-HQ restoration direction, enabling training-free exploration of perceptually improved solutions. Our analysis attributes its effectiveness to HQ–LQ semantic compatibility. RAEFace achieves state-of-the-art perceptual quality with competitive fidelity on synthetic and real-world benchmarks, underscoring semantic latent organization as a key design principle for blind face restoration.

AI USE STATEMENT

A large language model was used to assist with language editing of the manuscript, including refining the wording, correcting basic errors, and improving fluency.

ETHICS STATEMENT

This work studies blind face restoration. Our restoration models are trained on FFHQ and evaluated on existing face-image benchmarks. All datasets used in this study are publicly available. We use these datasets in accordance with their applicable licenses and usage restrictions.

REPRODUCIBILITY STATEMENT

The RAEFace architecture, training objectives, and PiE formulation are described in Section 3. The experimental section specifies the datasets, evaluation metrics, and ablation settings. Appendix A.1 provides the implementation details, including model configurations, training schedules, condition injection, and sampling procedures. The measurement protocols for linear probing, LDS, and iFID are provided in Appendix A.3. Task-specific data construction, training settings, and evaluation procedures for face colorization and inpainting are described in Appendices A.2.1 and A.2.2, respectively. We will release the code and model weights soon. All datasets and pretrained models used in this work are publicly available or available upon request, with some RAE encoders requiring an application.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 4690–4699, 2019.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. Int. Conf. Mach. Learn.*, 2024.

Lu Gan, Hanyu Yan, Chaofeng Chen, Junqi Hu, and Dan Zeng. Geomar: Unleashing geometrically aligned features for masked autoregressive blind face restoration. *arXiv preprint arXiv:2608.03923*, 2026.

Yuan Gao, Chen Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 4688–4697, 2026.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Inf. Process. Syst.*, 27, 2014.

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 Inf. Process. Syst.*, 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.

Mingkai Jia, Mingxiao Li, Zhijian Shu, Anlin Zheng, Liaoyuan Fan, Jiaxin Guo, Tianxing Shi, Dongyue Lu, Zeming Li, Xiaoyang Guo, et al. Dino-tok: Adapting dino for visual tokenizers. *arXiv preprint arXiv:2511.20565*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 4396–4405. IEEE, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 5128–5137. IEEE, 2021.

Senmao Li, Kai Wang, Joost van de Weijer, Fahad Khan, Chun-Le Guo, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, et al. $interlcm$: Low-quality images as intermediate states of latent consistency models for effective blind face restoration. In *Proc. Int. Conf. Learn. Representations*, volume 2025, pp. 59165–59197, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 36115–36125, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. IEEE Int. Conf. Comp. Vis.*, pp. 4172–4182. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 10674–10685. ieee, 2022.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *Proc. Int. Conf. Learn. Representations*, volume 2026, pp. 154506–154537, 2026.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In *Proc. Int. Conf. Learn. Representations*, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration. In *Proc. Int. Conf. Learn. Representations*, volume 2024, pp. 44867–44890, 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proc. Conf. AAAI*, 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 12626–12636. IEEE, 2025.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 40633–40642, 2026a.

Tianhang Wang, Yitong Chen, Wei Song, Zuxuan Wu, Min Li, and Jiaqi Wang. Decq: Detail-condensing queries for enhanced reconstruction and generation in representation autoencoders. *arXiv preprint arXiv:2605.22777*, 2026b.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 9164–9174. IEEE, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proc. IEEE Int. Conf. Comp. Vis. Workshops*, pp. 1905–1914. IEEE, 2021b.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid. *arXiv preprint arXiv:2603.05630*, 2026.

Ziyang Xu, Kangsheng Duan, Xiaolei Shen, Zhifeng Ding, Wenyu Liu, Xiaohu Ruan, Xiaoxin Chen, and Xinggang Wang. Pixelhacker: Image inpainting with structural and semantic consistency. *arXiv preprint arXiv:2504.20438*, 2025.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 5525–5533, 2016.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshops*, pp. 1190–1199. IEEE, 2022.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Liangyu Yuan, Yufei Huang, Mingkun Lei, Tong Zhao, Ruoyu Wang, Changxi Chi, Yiwei Wang, and Chi Zhang. Improving diffusion generalization with weak-to-strong segmented guidance. *arXiv preprint arXiv:2603.20584*, 2026.

Jack C Zhang. Gram newton-schulz: A fast, hardware-aware muon optimizer. 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pp. 586–595. IEEE, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *Proc. Int. Conf. Learn. Representations*, volume 2026, pp. 35791–35820, 2026.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

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

A Appendix

A.1 Implementation Details

Training. Tab. 5 summarizes the architecture and training configurations for the main model and encoder ablations. Before restoration training, the ViT-XL decoders are fine-tuned on FFHQ for 40 epochs using AdamW with a base learning rate of 2×10^{-5}, while the pretrained visual encoders remain frozen. The resulting encoder–decoder pairs are then fixed throughout DDT training. Only the DDT parameters, including the added conditioning branches and REPA projector, are optimized for restoration. For the main model, the shared backbone parameters are loaded from the ImageNet-pretrained checkpoint, while the newly introduced conditioning and alignment parameters are initialized separately.

Condition injection. We follow the DiT^{DH}-XL architecture in RAEv2 (Singh et al., 2026) and add cross-attention after each DDT encoder block, as illustrated in Fig. 10. A 3×3 convolution projects the LQ latent into the DDT hidden dimension without changing its spatial resolution. The projected features are flattened into conditioning tokens and kept separate from the main transformer sequence. These LQ tokens provide the keys and values, while the spatial DDT tokens provide the queries. The same LQ tokens are reused across encoder blocks, allowing the evolving restoration features to access the degraded observation at each depth. Cross-attention outputs are added through residual connections with zero-initialized output projections, preserving the pretrained backbone's behavior at initialization. For CFG training, the entire LQ latent is replaced with a learned null representation with probability 0.1.

Sampling. We evaluate the EMA model using 50 Euler ODE steps, with the dimension-adjusted time schedule described in Sec. 2. On CelebA-Test, the CFG scale is $s = 2.25$ and the PiE strength is $w = 0.065$. For LFW-Test, WebPhoto-Test, and Wider-Test, we use $s = 1.5$ and $w = 0.1$. At each step, CFG first combines the conditional and unconditional HQ predictions. PiE then uses the CFG-adjusted prediction and the encoded LQ latent to define the restoration direction. The extrapolated prediction is converted into the transport velocity for the next Euler update, so PiE influences the sampling trajectory throughout inference.

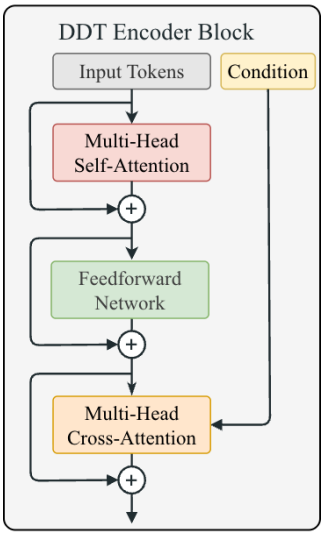


Figure 10: DDT encoder with LQ conditioning.

A.2 Applications

We extend RAEFace to face colorization and inpainting by constructing task-specific input–target pairs. Both tasks use the frozen DINOv3-L encoder and the FFHQ-trained RAE decoder. We initialize separate DDT models from ImageNet-pretrained weights and train each on FFHQ for 20 epochs. Training uses GMuon with a learning rate of 2×10^{-4} and a global batch size of 32. At inference

Table 5: Architecture, DDT training, and REPA configurations for the main model and encoder ablations.

Configuration	DINOv3 (Main)	DINOv3 (Ablation)	SigLIP2 (Ablation)	MAE (Ablation)
Architecture				
Backbone	DiT^{DH}-XL	DiT^{DH}-XL	DiT^{DH}-XL	DiT^{DH}-XL
Encoder blocks / Hidden dim / Heads	28 / 1440 / 20	28 / 1440 / 20	28 / 1440 / 20	28 / 1440 / 20
Decoder blocks / Hidden dim / Heads	2 / 2048 / 16	2 / 2048 / 16	2 / 2048 / 16	2 / 2048 / 16
RAE encoder	DINOv3-L	DINOv3-L	SigLIP2-L	MAE-L
Input channels	1024	1024	1024	1024
Latent spatial resolution	32×32	32×32	32×32	32×32
Latent Patch size	1	1	1	1
LQ projection	3×3 Conv.	3×3 Conv.	3×3 Conv.	3×3 Conv.
DDT training				
Dataset	FFHQ	FFHQ	FFHQ	FFHQ
Online degradation	Real-ESRGAN	Real-ESRGAN	Real-ESRGAN	Real-ESRGAN
Initialization	ImageNet-pretrained	From scratch	From scratch	From scratch
Global batch size	32	32	32	32
Optimizer	GMuon	GMuon	GMuon	GMuon
Base learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
Weight decay	0	0	0	0
Mixed precision	bfloat16	bfloat16	bfloat16	bfloat16
Gradient clipping (max norm)	1.0	1.0	1.0	1.0
EMA decay	0.9995	0.9995	0.9995	0.9995
CFG dropout probability	0.1	0.1	0.1	0.1
Prediction type	x-prediction	x-prediction	x-prediction	x-prediction
Time distribution	Logit-normal	Logit-normal	Logit-normal	Logit-normal
DDT training schedule				
Training epochs	**50**	20	20	20
LR schedule	**Linear decay**	Constant	Constant	Constant
Constant-LR epochs before decay	25	—	—	—
Final learning rate	2×10^{-5}	2×10^{-4}	2×10^{-4}	2×10^{-4}
REPA				
Target encoder	Same as RAE encoder	Same as RAE encoder	Same as RAE encoder	Same as RAE encoder
Alignment layer depth	8	8	8	8
REPA coefficient (λ)	0.5	0.5	0.5	0.5

time, we use the EMA models with 50 Euler ODE steps. The CFG scale is set to 1.0 and the PiE strength to 0 for both tasks.

Training and sampling settings. Both tasks use the frozen DINOv3-L encoder and the RAE decoder trained on FFHQ. The encoder and decoder remain fixed during restoration training. Separate DDT models are initialized from ImageNet-pretrained weights and trained on FFHQ for 20 epochs using GMuon, with a learning rate of 2×10^{-4} and a global batch size of 32. At inference time, we evaluate the EMA models using 50 Euler ODE steps. For both tasks, the CFG scale is $s = 1.0$ and the PiE strength is $w = 0$, so the results directly evaluate the task-specific conditional restoration models.

A.2.1 Face Colorization

Training pairs. We convert HQ color images to grayscale and retain the original color images as reconstruction targets. The grayscale inputs are replicated across three channels before being encoded by the shared RAE encoder. Their latents provide the LQ conditions, while the paired color-image latents serve as the restoration targets.

Evaluation. We evaluate on CelebA-Test by converting the test images to grayscale and comparing the restored color images with CodeFormer outputs. The original color images provide the paired ground truth. As shown in Fig. 11, RAEFace consistently generates more natural and realistic facial colors, while faithfully preserving the facial structure. In contrast, CodeFormer often introduces color bleeding or structural distortions, due to the quantization errors in its discrete codebook, exacerbated by limited codebook capacity. These observations highlight the advantage of RAEFace in producing visually pleasing and structurally consistent colorizations.

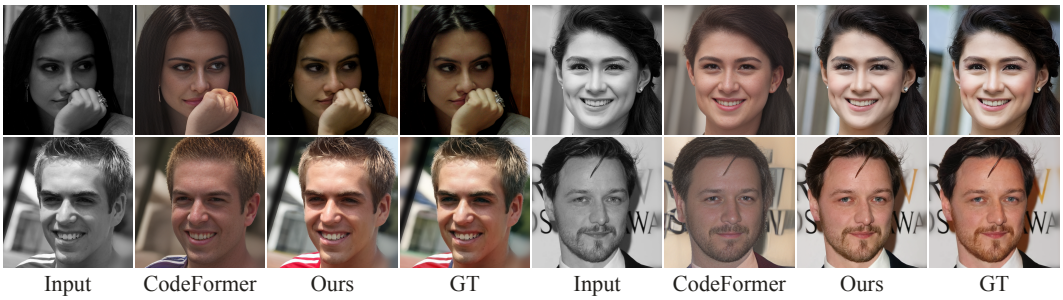


Figure 11: Visualization of Colorization Task on CelebA-Test.

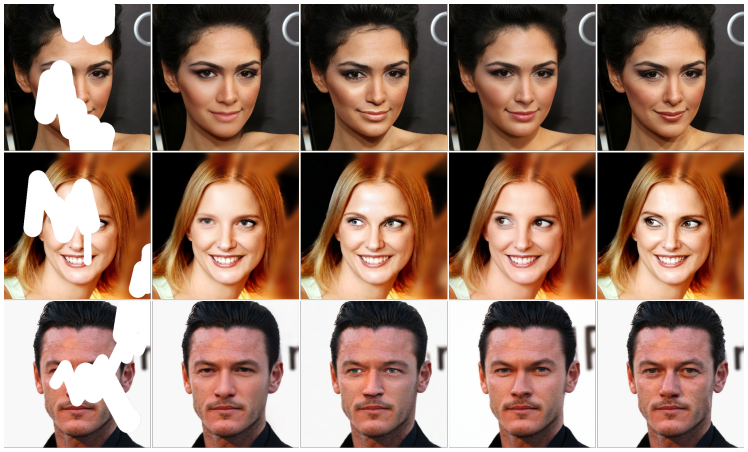


Figure 12: Visualization of Inpainting Task on CelebA-Test

A.2.2 Face Inpainting

Training pairs. We apply random masks to HQ faces and use the original images as reconstruction targets. Free-form brush masks and rectangular masks are sampled with probabilities of 0.6 and 0.4, respectively. Brush widths range from 24 to 80 pixels, and rectangular masks cover 10–40% of the image area. Masked regions are filled with random noise before encoding. The DDT is conditioned on the masked-image latents to predict the complete HQ latents.

Evaluation. We evaluate on CelebA-Test using fixed brush masks covering 20–30% of each image, with the original unmasked images serving as ground truth. Qualitative comparisons are shown in Fig. 12. When key facial regions are occluded, RAEFace better predicts the underlying facial semantics from the visible context, whereas both CodeFormer and PixelHacker tend to produce semantically implausible results in the masked regions.

A.3 Correlation analysis

A.3.1 Experiment Settings

We use 13 pretrained autoencoders: eight representation autoencoders (RAE), built on DINOv3-L, DINOv3-B, DINOv2-B, SigLIP 2-L, EUPE-B, MoCo v3-B, I-JEPA-H and MAE-L, each decoder trained on FFHQ for 40 epochs, and five VAEs: VA-VAE, E2E-VA-VAE, SD-VAE3.5, FLUX-VAEv1 and SD-VAE-ema. All experiments run on CelebA. For each model we sweep the extrapolation strength from 0 to 1, evaluating 64 images at each of the 101 settings and recording PSNR together with four no-reference quality metrics: MANIQA, TOPIQ-Face, CLIPIQA and MUSIQ.

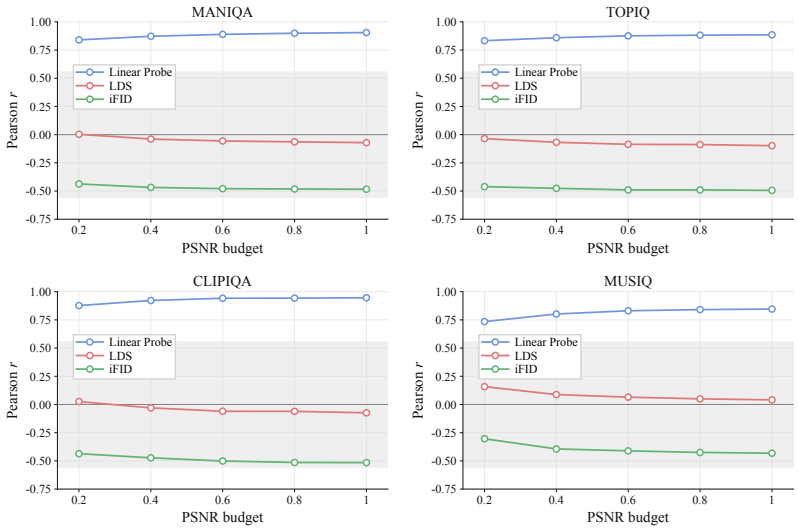


Figure 13: Pearson correlation between each representation property and the oracle perceptual gain, as a function of the matched PSNR-loss budget, for four no-reference metrics. The grey band marks $|r|$ below the significance threshold ($p < 0.05$, $|r| < 0.553$).

A.3.2 Representation Analysis Details

Linear probing. Following RAEv2 (Singh et al., 2026), we assess semantic information using linear probing on 14 facial attributes of FFHQ, annotated by Qwen3(Bai et al., 2025). A linear classifier is trained on spatially averaged HQ latents and evaluated on paired LQ test images:

$$\mathrm{LP}_{\mathrm{HQ}\to\mathrm{LQ}} = \frac{1}{A}\sum_{a=1}^{A} \mathrm{Acc}\big(h_a(\bar{z}_{\mathrm{LQ}}), y_a\big), \tag{9}$$

where $A = 14$, $\bar{z}$ denotes spatially averaged features, and h_a is the HQ-trained classifier for attribute a. HQ-to-HQ and LQ-to-LQ accuracy are measured analogously.

Local distance similarity (LDS). Following iREPA (Singh et al., 2025), we measure spatial organization on HQ images from CelebA-Test:

$$\mathrm{LDS} = \mathbb{E}_{(i,j)\in\mathcal{N}}[\cos(z_i, z_j)] - \mathbb{E}_{(i,j)\in\mathcal{F}}[\cos(z_i, z_j)], \tag{10}$$

where z_i denotes a spatial token, and $\mathcal{N}$ and $\mathcal{F}$ contain adjacent and distant token pairs within an image, respectively. Scores are averaged over images.

Interpolated FID (iFID). Following Xu et al. (2026), we evaluate decoded interpolations between neighboring HQ latents on CelebA-Test:

$$\tilde{z}_i = \mathrm{slerp}(z_i, z_{j(i)}; 0.5), \qquad \mathrm{iFID} = \mathrm{FID}\big(\{D(\tilde{z}_i)\}, \{x_i\}\big), \tag{11}$$

where $j(i)$ denotes a sampled latent-space neighbor and D is the corresponding decoder. We use 3,000 HQ images for both interpolation and the reference distribution.

A.3.3 Additional Correlation Results

Fig. 13 reports the Pearson correlation between each representation property and the oracle perceptual gain, computed across the 13 pretrained autoencoders at matched PSNR losses. Linear-probe accuracy correlates strongly and positively with the gain throughout, consistently so across all four no-reference metrics and all five PSNR-loss levels, and remains above the significance threshold in every panel.

A.4 More Qualitative Results

We provide more visual comparisons between our method and other face restoration methods on Real-World Datasets in Fig. 14.



Figure 14: More Visual Comparisons on Real-World Datasets