INFINITYFACE: BLIND FACE RESTORATION WITH A VISUAL AUTOREGRESSIVE TEXT-TO-IMAGE PRIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Blind face restoration (BFR) is ill-posed, requiring a balance between perceptual realism and fidelity to the degraded input. We explore adapting Infinity, a bitwise visual autoregressive text-to-image model, to BFR. It is non-trivial as generation synthesizes plausible content from a generative prior, whereas restoration must retain the reliable input evidence. We propose ***InfinityFace***, which reformulates open-ended next-scale generation as input-constrained next-scale restoration through two complementary forms of LQ conditioning. Specifically, a scale-dependent suffix controller combines the restored prefix with the corrected unresolved LQ features, preserving reliable LQ evidence while suppressing fine-scale feature mismatch. Meanwhile, scale-agnostic Text–LQ conditioning constructs an LQ visual memory and jointly injects it with text through multimodal cross-attention. Together, they anchor restoration to the observation while leveraging generative priors for realistic detail refinement. Experiments on synthetic and real-world BFR benchmarks show strong no-reference perceptual quality and competitive restoration performance.

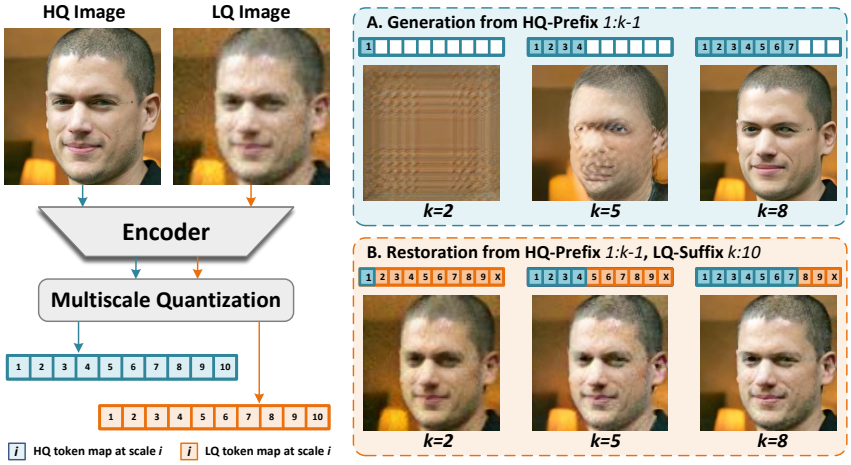


Figure 1: Our key motivation: restoration cannot be simply treated as LQ-conditioned generation, since the benefits of LQ evidence are inherently scale-dependent.

1 INTRODUCTION

Blind face restoration (BFR) aims to recover a high-quality (HQ) face image from a degraded observation affected by blur, noise, compression artifacts, and low effective resolution. The task is inherently ill-posed: many plausible HQ faces may correspond to the same LQ input, especially when identity details and local textures are severely corrupted. To constrain this ambiguous solution space, existing BFR methods commonly rely on generative priors, including GAN priors Wang et al. (2021a), codebook-based priors Gu et al. (2022); Zhou et al. (2022), and diffusion/consistency-based priors Yue & Loy (2022); Wang et al. (2025). These methods have substantially improved perceptual realism, but also expose a core tension in BFR: *the model must use the prior to synthesize missing details while still respecting the fidelity cues that remain in the LQ input.*

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.

Infinity Han et al. (2025) is a text-to-image generation model based on visual autoregressive (VAR) modeling and binary spherical quantization (BSQ). It represents an image as binary token maps whose resolution increases from coarse to fine, providing strong generation capacity and detail modeling. As shown in Figure 1, an HQ image and its LQ counterpart are encoded into multiscale token maps. The right side compares the input state used to predict scale k. In the oracle illustration, Figure 1(A) provides only the HQ prefix from scales 1 to $k-1$, corresponding to the generation-style input, whereas Figure 1(B) augments the same prefix with the raw LQ suffix from scales k to 10. This construction isolates the effect of retaining unresolved LQ information. We make two observations:

1. **From generation to restoration.** The comparisons at $k = 2, 5$ motivate transferring Infinity from next-scale generation to restoration. Even with HQ prefix, generation remains open-ended, whereas raw LQ suffixes constrain facial position and coarse structure. This shifts the objective from predicting what the next token map *could be* to what it *should become*.
2. **Scale-dependent LQ conditioning.** The comparison at $k = 8$ shows that the raw LQ suffix is not uniformly beneficial. Once the coarse facial structure has been largely determined by earlier scales, high-scale LQ token maps may carry degradation artifacts and become incompatible with high-quality detail. This motivates scale-dependent conditioning: retaining LQ guidance at coarse scales while progressively attenuating it at later scales, allowing the generative prior to synthesize clean details.

Based on these insights, we propose ***InfinityFace***, a text-guided blind face restoration framework that adapts Infinity from next-scale generation to next-scale restoration. During inference, each next-scale input is constructed from the restored prefix and a controlled unresolved LQ suffix, with the predicted residual progressively replacing the corresponding LQ residual. In parallel, the original LQ image is projected into the transformer representation space and injected together with text through multimodal cross-attention. This design anchors the restoration process to the observed face while retaining the generative prior for detail refinement.

Our contributions are summarized as follows:

- We explore transferring Infinity's next-scale generation paradigm to LQ-anchored autoregressive restoration for blind face restoration.
- We introduce two complementary mechanisms for precisely leveraging LQ evidence in autoregressive face restoration: scale-dependent next-scale restoration with an LQ suffix controller and scale-agnostic Text–LQ conditioning with LQ visual memory.
- We evaluate the proposed framework on synthetic and real-world BFR benchmarks, achieving favorable visual results and competitive quantitative performance.

2 Related Work

Blind Face Restoration. Blind face restoration recovers high-quality faces from inputs with unknown blur, noise, compression, and resolution loss. Because the task is highly ill-posed, existing methods rely on strong priors to narrow the solution space. GAN-prior methods use pretrained face generators to provide a realistic facial manifold: GFP-GAN Wang et al. (2021a) and GPEN Yang et al. (2021) are representative examples that embed GAN priors into the restoration pipeline. These methods are effective for perceptual realism, but the generator prior may dominate when the LQ input provides weak evidence, motivating restoration spaces with more explicit fidelity control. Codebook and dictionary-based methods constrain restoration with learned high-quality facial entries: VQFR Gu et al. (2022) builds a vector-quantized dictionary, while CodeFormer Zhou et al. (2022) predicts discrete codes with a transformer and exposes a fidelity-quality tradeoff. Recent work GeoMAR (Gan et al., 2026) incorporates geometric conditioning into masked autoregressive code prediction. Such discrete priors improve stability and controllability, but their finite entries can limit fine-grained expression and motivate higher-capacity visual tokenization. Diffusion and consistency-based methods exploit powerful generative denoisers for restoration: DifFace Yue & Loy (2022) formulates BFR with a diffusion prior, while OSDFace Wang et al. (2025) and InterLCM Li et al. (2025) seek faster restoration through one-step diffusion or latent consistency modeling. These approaches provide strong perceptual priors, but they still require careful conditioning to avoid hallucinating beyond the trustworthy information in the degraded input.

Visual Autoregressive Modeling. Visual Autoregressive modeling Tian et al. (2024) redefines image autoregression as coarse-to-fine next-scale prediction rather than raster-order next-token prediction. This scale-wise formulation better matches the hierarchical structure of images and has inspired restoration-oriented extensions such as VARSR Qu et al. (2025) and AlignVAR Liu et al. (2026). Unlike VAR-based models built on finite codebook tokenization, Infinity Han et al. (2025) is a text-to-image generation model that extends visual autoregressive modeling under a bitwise token-prediction framework. Its multiscale BSQ tokenization and bitwise prediction improve generation capacity and detail modeling beyond fixed-size discrete codebooks, making the trained model a useful high-capacity prior for BFR. Adapting this prior requires two changes: the next-scale process must be transformed from open-ended prediction of plausible token maps into restoration of the corresponding HQ token maps, and LQ evidence must be incorporated without allowing degraded details to dominate the generative prior. InfinityFace addresses these requirements through controlled LQ suffixes for autoregressive input construction and multimodal text-LQ cross-attention.

3 PRELIMINARIES: INFINITY

Multiscale BSQ tokenization. Infinity uses Binary Spherical Quantization (BSQ) Zhao et al. (2024) to tokenize images without a learned finite VQ codebook. An image I is encoded by the VAE encoder $\mathcal{E}$ into a continuous feature map $F = \mathcal{E}(I) \in \mathbf{R}^{h\times w\times d}$. The visual tokenizer decomposes F into K multiscale residual token maps $(R_1, R_2, \ldots, R_K)$, where R_k has spatial size $h_k \times w_k$ and the resolution grows from coarse to fine. Given the token maps up to scale k, the cumulative feature is $F_k = \sum_{i=1}^{k} \mathrm{up}(R_i, (h, w))$, where $\mathrm{up}(\cdot)$ denotes bilinear upsampling and $F_0 = 0$. At scale k, the residual feature before quantization is obtained by downsampling the remaining residual:

$$z_k = \mathrm{down}(F - F_{k-1}, (h_k, w_k)), \tag{1}$$

where $\mathrm{down}(\cdot)$ denotes downsampling. Each d-dimensional residual vector is quantized into a binary output q_k by BSQ:

$$q_k = \mathrm{BSQ}(z_k) = \sqrt{\frac{1}{d}}\,\mathrm{sign}\left(\frac{z_k}{\|z_k\|_2}\right), \tag{2}$$

Stacking these binary outputs over spatial positions gives the token map R_k, whose bit labels can be deterministically written as $\mathrm{bit}(R_k) \in \{0, 1\}^{h_k\times w_k\times d}$. After all token maps are obtained, the VAE decoder $\mathcal{D}$ maps the accumulated feature F_K back to the image space.

Unlike finite-codebook tokenizers, BSQ performs lookup-free bitwise quantization: each d-dimensional binary token yields an effective vocabulary size $V_d = 2^d$ without requiring a learned V_d-entry embedding table, alleviating the codebook bottleneck and supporting richer visual variations and finer reconstruction and generation details.

Bitwise Next-scale Generation. Infinity predicts multiscale token maps autoregressively. Given a text prompt t and its text embeddings $\Psi(t)$, the original next-scale likelihood is formulated as

$$p(R_1, \ldots, R_K \mid \Psi(t)) = \prod_{k=1}^{K} p(R_k \mid R_1, \ldots, R_{k-1}, \Psi(t)). \tag{3}$$

When predicting R_k, the context is formed by previous token maps and the text embeddings. The transformer input state for the k-th scale is the downsampled cumulative feature from previous scales, $X_k^{\mathrm{gen}} = \mathrm{down}(F_{k-1}, (h_k, w_k))$. For the first scale, X_1^{gen} is replaced by a start token derived from the text condition.

4 OUR METHOD

Given an LQ image I^{LQ} and its caption t, the VAE encoder and text processor produce visual features and text conditions. The LQ features then enter two complementary branches: the LQ suffix processor applies scale-specific corrections in the VAE feature space, while the LQ adapter maps them to the transformer space for cross-attention with text. At each scale, the transformer predicts a restored token map, which is accumulated for subsequent predictions. The accumulated multiscale feature is finally decoded into the restored face. The method framework is illustrated in Figure 2.

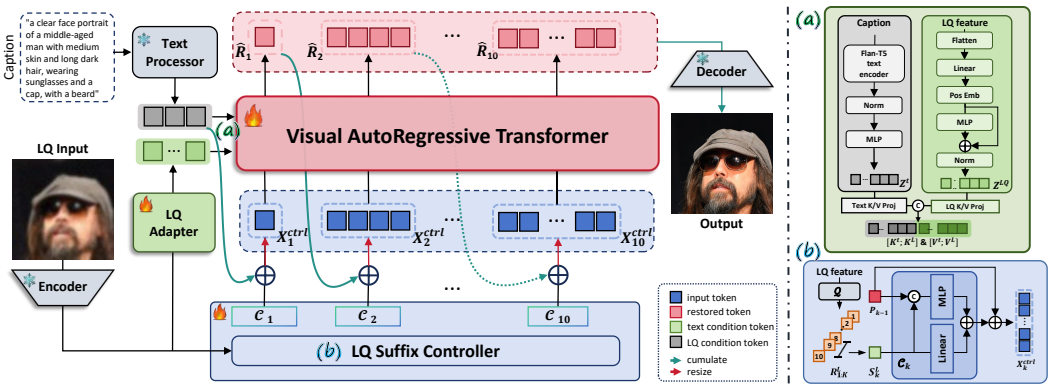


Figure 2: **Overview of InfinityFace.** Its key technical novelty lies in two complementary forms of LQ conditioning: (a) scale-agnostic Text–LQ conditioning in the transformer space and (b) scale-dependent LQ suffix control in the VAE feature space.

4.1 Scale-Dependent Next-Scale Restoration

Reformulation on Restoration. Unlike Infinity, which predicts each scale solely from the cumulative generated prefix, our restoration framework constructs this feature from the *restored prefix* and the unresolved *LQ suffix*. This design keeps each prediction anchored to the observation while progressively replacing LQ information with restored content. Given an LQ image I^{LQ}, we first encode it with the VAE encoder and quantize the resulting feature map into K residual token maps with the multiscale BSQ tokenizer:

$$R^{LQ}_{1:K} = \mathcal{Q}(\mathcal{E}(I^{LQ})), \tag{4}$$

where $\mathcal{Q}$ denotes multiscale BSQ quantization. We define the restored prefix feature as $P_{k-1} = \sum_{i<k} \mathrm{up}(\hat{R}_i, (h, w))$ and the unresolved LQ suffix feature as $S^{LQ}_k = \sum_{i\geq k} \mathrm{up}(R^{LQ}_i, (h, w))$. The raw restoration input state for predicting the k-th scale is then

$$X^{\mathrm{raw}}_k = \mathrm{down}\left(P_{k-1} + S^{LQ}_k, (h_k, w_k)\right). \tag{5}$$

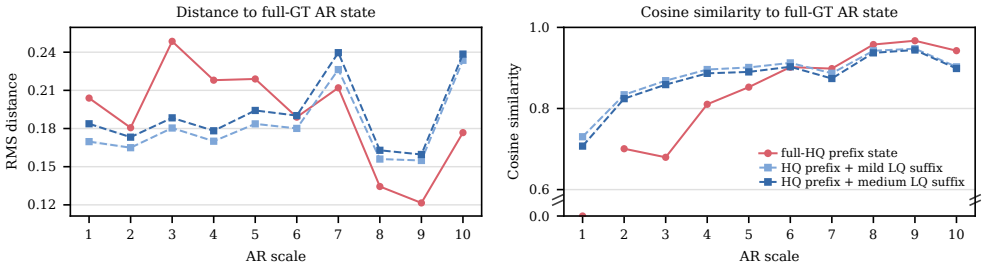


Figure 3: HQ prefix oracle analysis of the input visual state for predicting the k-th scale.

HQ Prefix Oracle Analysis. The model predicts $\hat{R}_k$ and uses it to expand the restored prefix for subsequent scales. The raw suffix provides useful input evidence, but its reliability is not uniform across scales. To isolate this effect from autoregressive prediction errors, we perform an HQ-prefix oracle analysis in Figure 3. At each scale, we compare three input states against a full-HQ reference constructed from the complete HQ token sequence: the HQ prefix alone, the HQ prefix integrated with either a mildly or heavily degraded LQ suffix. The mildly degraded LQ suffix generally reduces the RMS distance and increases the cosine similarity to the full-HQ reference at scale 1-6, whereas the heavily degraded suffix is less reliable and can already be worse at scale 6. More importantly, from scale 7 onward, both suffixes increasingly deviate from the reference, indicating that LQ suffixes contain useful coarse-scale evidence but introduce a scale-dependent feature mismatch at finer scales.

Scale-Dependent LQ Suffix Control. To preserve useful LQ evidence while mitigating scale-dependent mismatch, we propose a prefix-conditioned LQ suffix controller, illustrated in Figure 2(a). At scale k, P_{k-1} and S_k^{LQ} are reshaped into spatial tokens. Their normalized concatenation is projected to condition a residual MLP, whose output is then added to a scale-specific linear transform of the LQ suffix. The resulting LQ Suffix Controller for correction is formulated as:

$$\begin{aligned} \tilde{S}_k^{LQ} &= \mathcal{C}_k(P_{k-1}, S_k^{LQ}), \\ X_k^{\text{ctrl}} &= \text{down}\left(P_{k-1} + \tilde{S}_k^{LQ}, (h_k, w_k)\right), \end{aligned} \tag{6}$$

where $\mathcal{C}_k$ has independent parameters at each scale and operates in the VAE feature space. This forward process adjusts the LQ suffix according to the current restored prefix.

The restoration likelihood is written as

$$\begin{aligned} &p_\theta(\hat{R}_{1:K} \mid I^{LQ}, \Phi(t, I^{LQ})) \\ &\quad = \textstyle\prod_{k=1}^{K} p_\theta(\hat{R}_k \mid \hat{R}_{<k}, \tilde{S}_k^{LQ}, \Phi(t, I^{LQ})), \end{aligned} \tag{7}$$

where $\hat{R}_{<k}$ denotes the restored prefix available before scale k with $\hat{R}_{<1} = 0$. $\Phi(t, I^{LQ})$ denotes the text-LQ condition injected through multimodal cross-attention as detailed in Sec 4.2. This formulation keeps the unresolved scales anchored to the LQ observation after scale-dependent correction, while the restored prefix progressively overwrites the input state.

4.2 Scale-Agnostic Text-LQ Conditioning

As shown in Figure 3, LQ suffixes provide useful coarse-scale evidence but increasingly mismatch fine-scale representations. We therefore exploit LQ information in two complementary spaces: scale-dependent LQ suffix controller in the VAE feature space (Sec. 4.1) and scale-agnostic Text–LQ conditioning in the transformer space (Sec. 4.2). The latter jointly injects text and LQ image features through shared cross-attention at every scale, combining semantic guidance with observation-grounded visual cues.

LQ Visual Memory Construction. To inject LQ evidence into the higher-level representation space of the transformer, we introduce a lightweight LQ adapter that projects the VAE-encoded LQ features into its hidden space (see Figure 2). We denote the resulting LQ condition tokens by

$$\mathbf{M}^{\text{LQ}} = \mathcal{A}_{\text{LQ}}\left(\mathcal{E}(I^{LQ})\right), \tag{8}$$

where $\mathcal{A}_{\text{LQ}}$ denotes the LQ adapter. The resulting sequence $\mathbf{M}^{\text{LQ}}$ serves as the LQ-side condition for multimodal cross-attention and provides global information about the observed face in the transformer's semantic representation space. It acts as an external visual memory that can be selectively read at each scale, and its contribution complements the text condition by allowing cross-attention to balance image-derived evidence with the generative semantic prior.

Text Semantic Prior Construction. Following the original Infinity, the caption t is mapped to semantic condition tokens by

$$\mathbf{M}^{\text{t}} = \Psi(t), \tag{9}$$

where Ψ denotes the text encoder and subsequent process. The resulting text tokens provide a semantic prior for facial appearance and coarse attributes, guiding the transformer toward plausible high-quality facial content.

Shared-Query Text-LQ Conditioning. As illustrated in Figure 2(b), at scale k, the visual hidden sequence H_k produces a shared query for both conditions. The LQ tokens $\mathbf{M}^{\text{LQ}}$ and text tokens $\mathbf{M}^{\text{t}}$ are mapped through separate key–value projections to accommodate their heterogeneous representation spaces, yielding LQ keys and values (K^{LQ}, V^{LQ}) and text keys and values (K^t, V^t), respectively. The same projection and attention modules are shared across all scales, making the conditioning scale-agnostic. The multimodal Text-LQ cross-attention is computed as

$$Y_k = \text{Attn}\left(Q_k,\, [K^t; K^{LQ}],\, [V^t; V^{LQ}]\right), \tag{10}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes concatenation along the condition-token dimension. The shared query aligns both conditions with the current autoregressive state, while the concatenated keys and values enable their contributions to be selected dynamically rather than fused with fixed weights.

4.3 Training and Inference

Training. Training follows a teacher-forcing scheme. Since the restored prefix $\hat{R}_{<k}$ is unavailable when all scales are processed in parallel, we replace it with the HQ prefix while retaining the unresolved LQ suffix. The caption is also generated from the HQ image by VLM, avoiding expensive online generation. Let $P_{k-1}^{HQ} = \sum_{i<k} \mathrm{up}(R_i^{HQ}, (h, w))$. The training input for scale k is

$$X_k^{\mathrm{train}} = \mathrm{down}\left(P_{k-1}^{HQ} + \mathcal{C}_k(P_{k-1}^{HQ}, S_k^{LQ}), (h_k, w_k)\right). \tag{11}$$

Let ℓ_k denote the resulting bit logits. We optimize the weighted multiscale binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{bit}} = \sum_{k=1}^{K} w_k \,\mathrm{BCE}(\ell_k, \mathrm{bit}(R_k^{HQ})), \tag{12}$$

where $w_k = \frac{1}{\sqrt{h_k \cdot w_k}}$, balancing the loss between each scale.

Inference. At inference, the teacher-forced HQ prefix is replaced by the autoregressively restored prefix, yielding the next-scale restoration process in Sec. 4.1. The caption t is generated from the input LQ image by VLM. This practical caption source may differ from an HQ-derived oracle caption in training; its effect is examined by the caption-source ablation in Table 4.

5 Experiments

5.1 Experimental Settings

Datasets. We train on FFHQ Karras et al. (2019) dataset which contains 70,000 high-quality face images. All images are resized to 512×512. We follow Real-ESRGAN Wang et al. (2021b) to synthesize the corresponding LQ inputs. We evaluate our method on three widely used benchmarks for BFR, including one synthetic dataset and two real-world datasets. The synthetic CelebA-Test contains 3,000 images derived from CelebA-HQ Karras et al. (2018), degraded using the same pipeline as training. The two real-world datasets include LFW-Test Huang et al. (2008) (1,711 images) and Wider-Test Yang et al. (2015) (970 images). Captions for all images are generated with Qwen3-8B **?**, please refer to Appendix A.1.1 for the details of caption generation.

Evaluation Metrics. We adopt reference-based and no-reference quality metrics for the comprehensive evaluation. Specifically, we use no-reference metrics FID Heusel et al. (2017), CLIPIQA Wang et al. (2022), MUSIQ Ke et al. (2021), MANIQA Yang et al. (2022), and TOPIQ Chen et al. (2024). FID against the FFHQ measures distribution-level realism, while MUSIQ focuses on multiscale quality, MANIQA captures local distortions and textures, CLIPIQA evaluates CLIP-aligned naturalness, and TOPIQ measures semantic-to-distortion quality. For the synthetic CelebA-Test set, we also use reference-based metrics, including PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018).

Implementation Details. We build InfinityFace based on Infinity-2B and use the VAE with $d_{\mathrm{vae}} = 32$, introducing approximately 283.4M additional parameters over the original backbone. The proposed method is trained for 20 epochs using eight Nvidia GeForce RTX 4090 GPUs with batch size of 5 per GPU. We optimize the model parameters using AdamW, with the learning rate initialized at 5×10^{-5} and cosine-decayed to 2×10^{-6}.

Comparison Methods. We compare InfinityFace with representative blind face restoration methods including CodeFormer Zhou et al. (2022), DAEFR Tsai et al. (2024), GeoMAR Gan et al. (2026), DifFace Yue & Loy (2022), OSDFace Wang et al. (2025), and VARSR Qu et al. (2025), covering codebook-based, diffusion-based and VAR-based restoration. We also include the LQ input as a baseline for reference. In comparison, VARSR is adapted to BFR by fine-tuning its VQ-VAE on FFHQ and all methods are evaluated using their official implementations.

5.2 Main Results

Evaluation on Real-world Datasets. As shown in Table 1, InfinityFace obtains the best MANIQA and TOPIQ scores on both LFW-Test and Wider-Test, while also achieving the second-best FID,



Figure 4: Qualitative comparison on real-world LFW-Test (Rows 1-2) and Wider-Test (Rows 3-4).

Table 1: Quantitative comparison on real-world datasets LFW-Test and Wider-Test. Bold and underline indicate the best and second-best results.

	LFW-Test					Wider-Test				
Method	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
LQ input	138.88	0.283	26.87	0.199	0.296	201.31	0.247	14.22	0.175	0.141
CodeFormer	51.81	0.697	75.47	0.523	0.849	39.23	0.698	73.40	0.496	0.835
DAEFR	47.70	0.697	**75.84**	0.542	0.857	36.81	0.698	74.15	0.521	0.833
GeoMAR	47.03	0.692	75.60	0.540	0.853	37.44	0.680	73.15	0.509	0.826
DifFace	46.67	0.611	69.86	0.461	0.779	37.70	0.594	64.89	0.431	0.744
OSDFace	**44.97**	**0.720**	75.35	<u>0.549</u>	0.852	**34.80**	**0.728**	74.60	<u>0.523</u>	0.839
VARSR	50.61	0.631	75.20	0.508	<u>0.869</u>	40.92	0.650	**75.35**	0.516	<u>0.872</u>
InfinityFace	<u>46.65</u>	<u>0.709</u>	<u>75.83</u>	**0.554**	**0.893**	<u>36.79</u>	<u>0.724</u>	<u>75.13</u>	**0.552**	**0.880**

CLIPIQA, and MUSIQ scores on both datasets. These results indicate robust perceptual quality under both mild and heavy real-world degradations. The qualitative results in Figure 4 further illustrate the behavior. The two upper rows from LFW-Test show that InfinityFace restores natural facial details without introducing excessive smoothing or overly sharpened textures. In the severely degraded third row, it recovers a coherent facial contour, expression, and local contrast while avoiding conspicuous color shifts and artificial textures. In the fourth row, despite the large pose variation and sunglasses, it preserves the side-facing geometry, hair silhouette, and accessory structure.

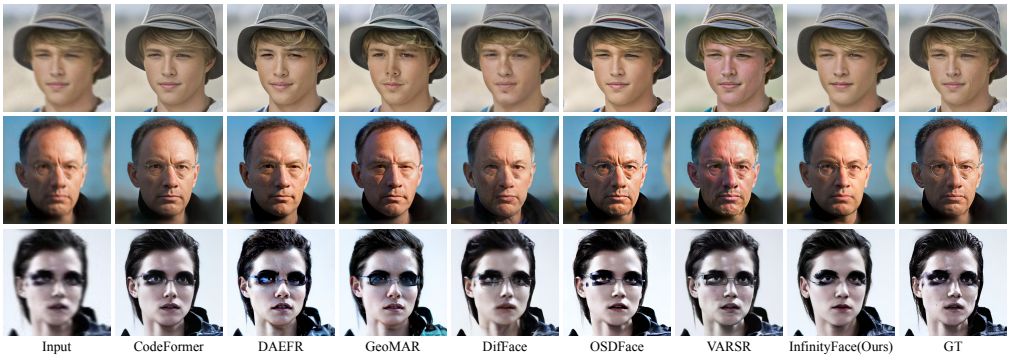


Figure 5: Qualitative comparison on synthetic CelebA-Test.

Evaluation on Synthetic Dataset. On CelebA-Test, InfinityFace achieves the best MUSIQ, MANIQA, and TOPIQ scores and the second-best CLIPIQA, while its FID and LPIPS remain competitive. Its lower PSNR and SSIM reflect the trade-off between perceptual enhancement and

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

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
LQ input	25.33	0.706	0.574	145.23	0.296	17.00	0.160	0.198
CodeFormer	**25.15**	**0.685**	**0.227**	53.10	0.671	75.55	0.538	0.864
DAEFR	22.59	0.627	0.243	40.64	0.668	75.28	0.544	0.854
GeoMAR	23.94	0.597	0.271	38.59	0.701	74.82	0.539	0.847
DifFace	24.80	0.685	0.272	**38.32**	0.571	68.96	0.468	0.778
OSDFace	23.38	0.658	0.240	44.18	**0.695**	75.46	0.550	0.853
VARSR	23.77	0.681	0.261	47.92	0.605	75.04	0.512	0.869
InfinityFace	22.59	0.644	0.266	48.56	0.681	**75.57**	**0.572**	**0.891**

strict pixel fidelity. Qualitatively, Figure 5 shows that InfinityFace restores natural facial details while preserving the visible global structure, hairstyle, expression, and accessories. It avoids excessive smoothing, over-sharpening, and color or texture shifts visible in other restorations, producing results that remain visually close to the HQ references while restoring plausible facial details.

5.3 Ablation Studies

Ablation Protocol. We ablate the two LQ-conditioning components introduced in the Method: the feature-space LQ suffix controller and the hidden-space LQ visual memory. All variants use the same backbone, restored prefix, training degradation, text condition and evaluation protocol. As shown in Table 3, without suffix control, the raw LQ suffix constructs X^{raw}; with suffix control, it is corrected by $\mathcal{C}_k$ to construct X^{ctrl}. Independently, LQ visual memory is either excluded from or injected into the shared-query Text-LQ cross-attention. Figure 6 also includes LQ and pure T2I diagnostic references. T2I generates a plausible but unrelated face. In the first row, incomplete variants show visible shifts in eye and eyebrow positions and produce less natural mouth and teeth textures. In the second row with heavier degradation, they struggle with hairstyle, sunglasses, and overall facial texture. The full model better preserves these identity-relevant structures, showing the necessity of combining feature-space input-state correction with hidden-space visual evidence.

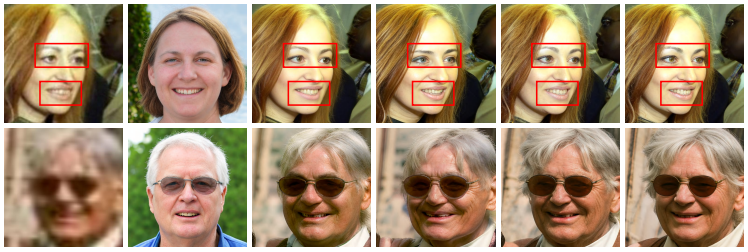


Figure 6: Qualitative ablation of the two LQ-conditioning components. (a) Baseline; (b) LQ suffix control only; (c) LQ visual memory injection only; and (d) full model.

Effect of LQ Suffix Controller. LQ suffix control operates before transformer prediction by correcting the unresolved LQ suffix in the VAE feature space. Comparing (a) with (b), and (c) with (d) isolates its effect without and with LQ visual memory, respectively. The controlled suffix improves all metrics in both matched comparisons, supporting the role of scale-specific suffix control in constructing a more reliable input state.

Table 3: Ablation of our two key designs.

Setting	Suffix Control	LQ Memory	Wider-Test		
			CLIPIQA↑	MUSIQ↑	FID↓
(a)	×	×	0.689	74.02	44.15
(b)	√	×	0.704	73.54	**36.56**
(c)	×	√	0.717	74.69	37.09
(d)	√	√	**0.724**	**75.13**	36.79

Effect of LQ Visual Memory Construction. To isolate the hidden-space LQ visual memory, we compare (a) with (c) under the raw suffix and (b) with (d) under the controlled suffix. With the raw suffix, adding LQ visual memory improves all metrics. With the controlled suffix, LQ visual memory improves CLIPIQA and MUSIQ while slightly reduce FID. The component therefore provides

Table 4: Ablation of text condition on CelebA-Test by varying the caption source at inference.

Caption Source	LPIPS↓	FID↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
HQ (oracle)	**0.264**	**47.05**	**0.683**	**75.62**	0.571
LQ (default)	0.266	48.56	0.681	75.57	**0.572**
Fixed prompt	0.277	49.08	0.665	75.15	0.563

broad perceptual gains. Taken together, the results indicating that suffix correction and hidden-space evidence optimize complementary aspects of perceptual quality.

Effect of Text Condition. Beyond the two LQ-conditioning components, InfinityFace also uses text condition as a semantic prior. We therefore ablate the caption source in Table 4 while keeping the trained model and visual inputs fixed. LQ captions correspond to the default inference setting, the fixed prompt uses "a clear face portrait", and HQ captions provide an oracle reference. The result shows that LQ captions consistently outperform the fixed prompt and remain close to HQ oracle captions, showing that practical captions provide useful semantic guidance without requiring oracle text. Please refer to Appendix A.1 for more details of caption influence and Appendix A.3 for discussion on failure cases when VLM generates inaccurate captions.

5.4 User Study

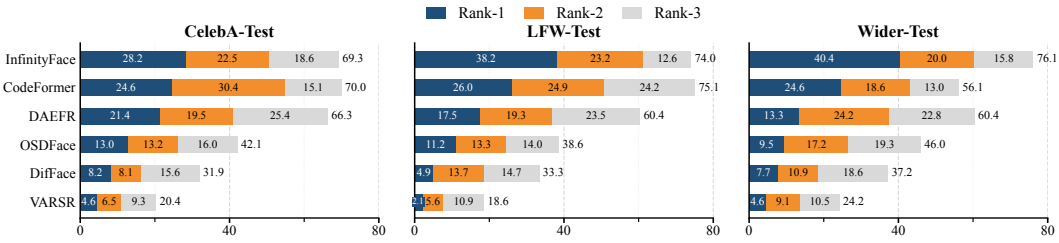


Figure 7: User-study ranking rates for Rank-1, Rank-2, Rank-3 on each dataset.

We further conduct a user study to assess perceptual preference beyond automatic metrics. The study includes 57 participants and 20 questions, consisting of 10 CelebA-Test, 5 LFW-Test, and 5 Wider-Test examples. For each question, participants were instructed to jointly consider perceptual quality and the fidelity of facial attributes supported by the input (and GT image in CelebA-Test) when comparing the restored images. They then ranked their three preferred results as Rank-1, Rank-2, and Rank-3. As shown in Figure 7, InfinityFace achieves the highest Rank-1 rate on all datasets and its Rank-1 advantage over CodeFormer is more pronounced on the real-world datasets. On CelebA-Test, InfinityFace obtains a Top-3 rate of 69.30%, close to CodeFormer's 70.00%. This pattern suggests that InfinityFace tends to produce perceptually natural details that remain plausible with respect to the observable input evidence, making its advantage more pronounced when no clean reference is available, while retaining strong competitiveness under GT-referenced evaluation.

6 Conclusion

We presented InfinityFace, which unlocks bitwise visual autoregressive text-to-image models, exemplified by Infinity, for BFR. Its key technical novelty lies in two complementary forms of LQ conditioning: scale-dependent next-scale restoration with LQ suffix controller, and scale-agnostic Text–LQ conditioning involving LQ visual memory construction. Together, they achieve strong perceptual restoration across synthetic and real-world benchmarks, with competitive results on no-reference quality metrics and in a crowdsourced user study. Further improving input fidelity remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for implement methods and generate captions for images used in training and inference. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for polishing paper writing and finding related works. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide dataset construction, model configuration, evaluation protocols, and ablation settings in the main paper Sec. 5 and supplementary materials. These descriptions are intended to support reproducibility of the reported restoration and comparison results.

REFERENCES

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. doi: 10.1109/TIP.2024.3378466.

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.

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, 2022.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30, 2017.

Gary B. Huang, Marwan A. Mattar, Tamara L. Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying 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, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer, 2021.

Senmao Li, Kai Wang, Joost van de Weijer, Fahad Shahbaz Khan, Chun-Le Guo, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. Interlcm: Low-quality images as intermediate states of latent consistency models for effective blind face restoration, 2025.

Cencen Liu, Dongyang Zhang, Wen Yin, Jielei Wang, Tianyu Li, Ji Guo, Wenbo Jiang, Guoqing Wang, and Guoming Lu. Alignvar: Towards globally consistent visual autoregression for image super-resolution, 2026.

Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024.

Yu-Ju Tsai, Yu-Lun Liu, Lu Qi, Kelvin C. K. Chan, and Ming-Hsuan Yang. Dual associated encoder for face restoration, 2024.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022.

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, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior, 2021a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021b.

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.

Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark, 2015.

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, 2022.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild, 2021.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization, 2024.

Shangchen Zhou, Kelvin C. K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022.

A APPENDIX

This appendix provides additional analyses and results to complement the main paper. It is organized into four parts:

- **Text Condition Analysis (Sec. A.1):** We describe the design principles of the sparse captions, compare captions generated from LQ and HQ images, and analyze the effect of text attribute intervention.
- **User Study Details (Sec. A.2):** We report more user study details and the calculation of Rank-1, Rank-2, Rank-3,.
- **Failure Cases and Limitations (Sec. A.3):** We discuss representative cases where insufficient visual evidence, ambiguous observations, or an overly influential generative prior can reduce restoration fidelity.
- **Additional Qualitative Results (Sec. A.4):** We provide further visual comparisons across synthetic and real-world restoration settings, including different degradation levels, poses, and face-adjacent accessories.

A.1 TEXT CONDITION ANALYSIS

This section discusses the principles of VLM caption-prompt design and provides additional visual analyses of text-condition usage, including comparisons of different caption sources and text attribute intervention. These analyses clarify the role of text as a semantic prior in InfinityFace.

A.1.1 CAPTION DESIGN PRINCIPLES

The captions used by InfinityFace are generated by a VLM, and our caption design is implemented as the instruction prompt used by the VLM. We ask the VLM to produce sparse weak captions, not exhaustive image descriptions. Their role is to expose stable visible semantics while leaving identity, pose, facial geometry, and fine details to the LQ image condition. Therefore, we use a short natural sentence rather than comma-only tags or a dense HQ-style description.

Under this prompt, the captioner describes only coarse attributes that are reliably visible under degradation. These include broad age and gender category, coarse skin tone and hair, expression or mouth state, facial hair, visible eyewear or head/face-adjacent accessories, major occlusion or face contact, and salient face-level makeup, markings, or lighting. We also encourage an attribute order that binds stable attributes to the subject phrase: age, gender, skin tone, and hair are placed close to the person noun, while expression, gaze, accessories, occlusion, and lighting are appended later.

Equally important, the caption prompt constrains the VLM to avoid identity-level or HQ-only details. It does not describe view angle or head pose, fine face geometry, eye/nose/lip shape, face shape, skin texture, wrinkles, small marks, detailed hairstyle, background, or image-quality and photographic-style terms. If an attribute is ambiguous due to blur, occlusion, lighting, or low effective face resolution, the VLM is instructed to skip it rather than guess. This constraint is also intended to reduce VLM hallucination: the caption should not complete invisible facial details from a generic face prior before restoration. This makes the caption neither as generic as a fixed prompt nor as dense as an oracle description: *it tells the model what kind of visible face semantics are present, but not what the restored identity should look like.*

A.1.2 CAPTION SOURCE COMPARISON

Figure 8 compares three caption sources under the same visual input and restoration model: a fixed generic prompt, the default LQ-derived caption, and an HQ-derived oracle caption. The LQ and HQ captions are generated with the same BFR-safe VLM prompt, but from different visual sources: the degraded LQ image for the practical inference setting and the HQ image for the oracle setting. Following the sparse design in Sec. A.1.1, the HQ caption is not allowed to become a dense description of identity geometry or fine HQ details. Consequently, the HQ caption contains only limited additional information, controlling the distribution gap between HQ-derived captions used in training and LQ-derived captions available at inference.



Figure 8: Comparison of caption sources.

The two rows in Figure 8 illustrate how these captions affect restoration. In the first row, the fixed prompt produces a plausible but generic facial appearance, without preserving the clearly visible smile and sunglasses on the head. The LQ caption explicitly supplies these expression and accessory cues, yielding a result closer to the GT, while the HQ caption differs only in its coarse skin-tone description and produces a similar result. In the second row, specifying a middle-aged smiling man with facial hair leads to more appropriate age- and expression-related facial texture and more accurate restoration of the mustache and stubble than the fixed prompt. The HQ caption changes only the coarse facial-hair term from *mustache* to *beard*, again giving an output close to the LQ-caption result. Together, these examples show that LQ-derived captions correct the default semantic prior mainly through visible age, expression, and face-adjacent accessory cues, whereas the constrained HQ captions provide only a modest oracle advantage.

A.1.3 Text Attribute Intervention

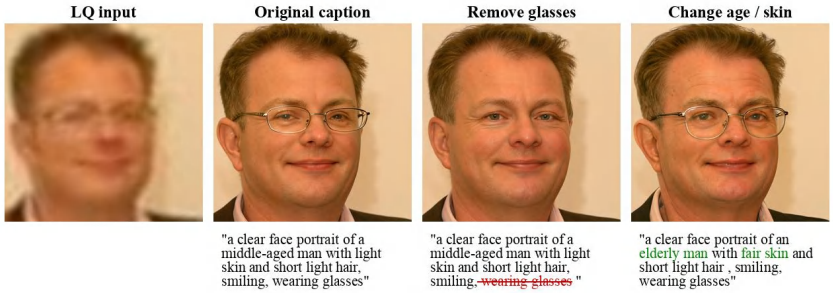


Figure 9: Comparison of text attribute intervention.

We further examine whether the caption is used as an attribute-level semantic prior rather than as a replacement for the LQ image condition. Figure 9 fixes the same LQ input and changes only the text condition shown below each restored image. The original caption describes a middle-aged man with light skin, short light hair, a smiling expression, and glasses. We then construct two interventions: removing the glasses attribute, marked by red strikethrough text, and changing only the coarse age and skin description to an elderly man with fair skin, marked in green. Other attributes, including the smile and the short light hair description, are kept unchanged.

The visual changes follow the edited text attributes while preserving the global image-conditioned structure. With the original caption, the model restores clear eyeglasses and a mild smile that are

consistent with the LQ input. When *wearing glasses* is removed from the caption, the eyewear is largely suppressed, but the overall face layout, hair, skin tone, and expression remain close to the original-caption result. Conversely, when the caption keeps *wearing glasses* but changes the subject to an elderly man with fair skin, the restored image retains the glasses and smile while introducing stronger age-related facial texture, especially around the forehead and eye regions. These results indicate that the text condition provides a controllable semantic bias for ambiguous face attributes such as accessories and age-related appearance, while the LQ image remains the dominant source for identity layout and coarse facial structure.

A.2 User Study Details

To keep the comparison fair, each question presents the six competing restorations produced from the same LQ input under the same display size and asks participants to rank their top three preferred results. All 57 participants completed all 20 questions, yielding $N = 57 \times 20 = 1140$ valid pairs with no invalid pairs.

For a method m, let $C_k(m)$ be the number of valid respondent-question pairs in which m is assigned rank k; if m is not selected in the top three, it contributes zero to all three rank-specific counts. The Rank-k rate is then

$$\text{Rank-}k(m) = \frac{C_k(m)}{N} \times 100\%, \quad k = 1, 2, 3.$$

The Top-3 rate is

$$\text{Top-3}(m) = \sum_{k=1}^{3} \text{Rank-}k(m).$$

Thus Rank-1, Rank-2, and Rank-3 are normalized by the same denominator, and Top-3 measures how often a method appears anywhere in the retained top-three set.

A.3 Failure Cases and Limitations



(a) (b) (c)

Figure 10: Representative failure cases and limitations. The first row shows the LQ inputs and the second row shows the restored results.

Figure 10 summarizes three representative failure modes. In Figure 10(a), the face is restored naturally, but fine text-like patterns on the hat are not faithfully preserved. This shows that small non-facial structures on head accessories remain challenging for our method. In Figure 10(b), the VLM caption describes the expression as smiling, but the restoration model interprets the bright lip reflection in the LQ input as white teeth, producing an incorrect mouth detail. This reflects ambiguity in local image evidence rather than an incorrect text description. In Figure 10(c), rare decorative accessories reveal a long-tail limitation: uncommon face-adjacent structures may be simplified or distorted by the generative prior. These cases suggest that future improvements should focus on uncertainty-aware handling of ambiguous local evidence, long-tail accessories, and fine non-facial patterns.

A.4 Additional Qualitative Results

We provide additional qualitative results on the three evaluation datasets used in our experiments. The comparisons in Figures 11–13 show that instead of producing over-sharpened, waxy, or locally inconsistent facial textures, our method tends to generate clean skin and facial details that remain visually plausible under the input identity layout.

On LFW-Test, which contains real-world portraits with diverse expressions, illumination, accessories, and backgrounds, InfinityFace restores clear facial structures while keeping the results visually balanced. Compared with competing methods that may introduce harsh contrast or unnatural local texture, our results are generally smoother in appearance without becoming overly blurred. On Wider-Test, where the inputs are often more severely degraded and less controlled, our method still produces coherent faces with fewer conspicuous texture artifacts, showing stable behavior under challenging real-world degradations. On CelebA-Test, where GT images are available for synthetic degradations, InfinityFace produces restorations that are close to the GT in overall facial appearance while maintaining natural skin, expression, hair, and accessory details. These qualitative results support the main-paper observation that InfinityFace favors perceptually natural restoration rather than aggressive detail hallucination.

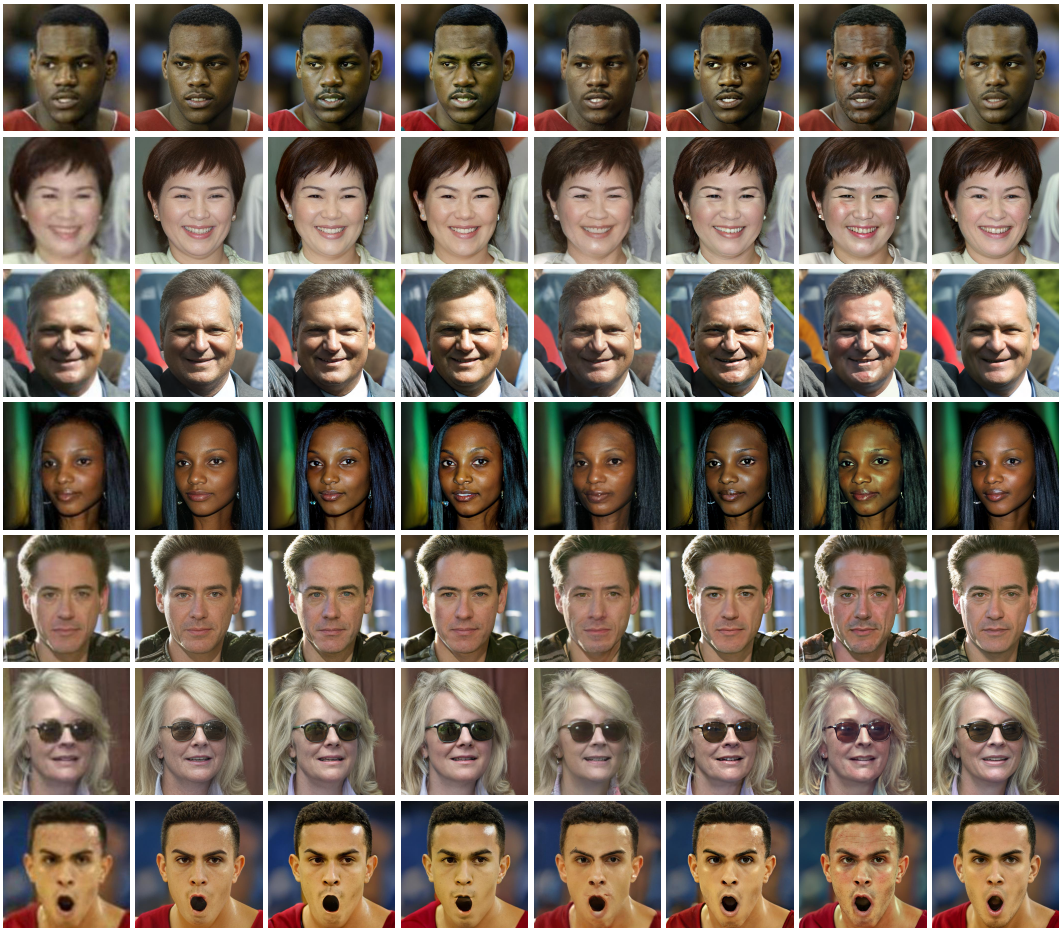


Figure 11: Qualitative comparisons on LFW-Test.

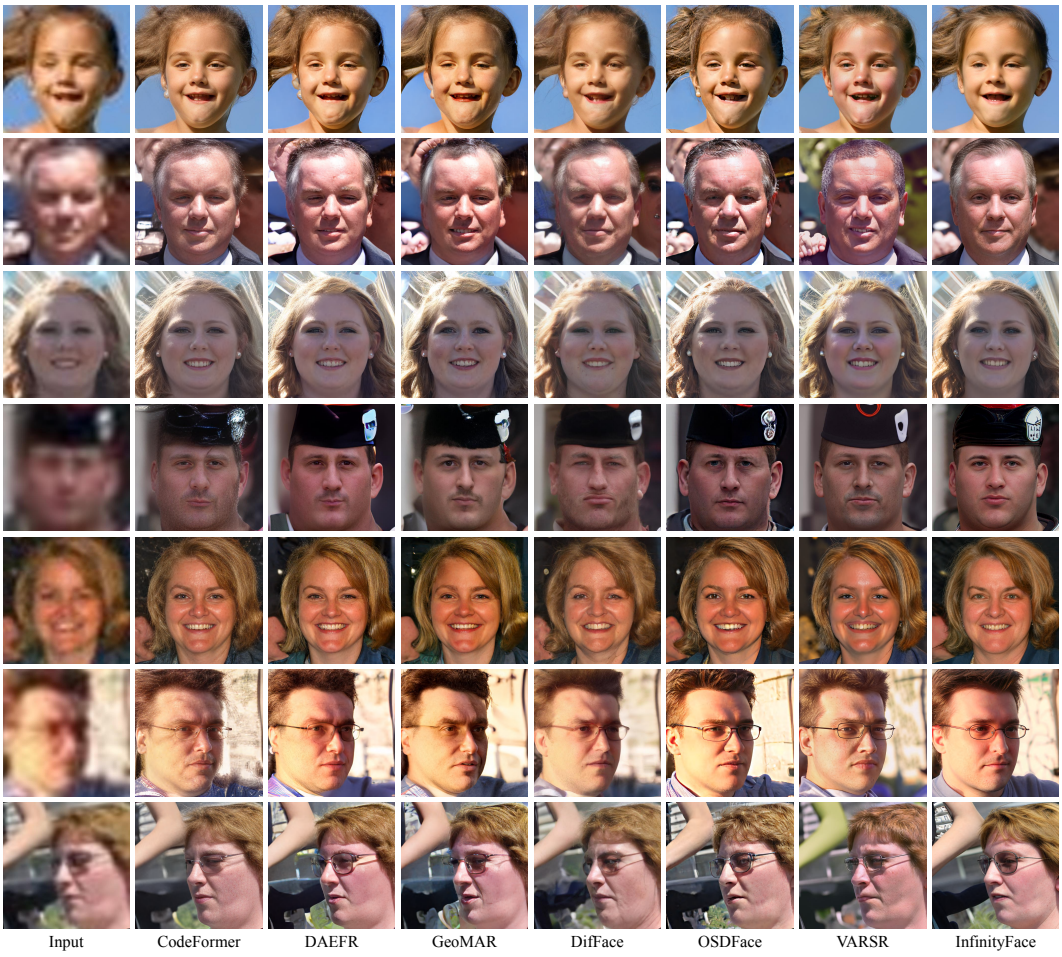


Figure 12: Qualitative comparisons on Wider-Test.

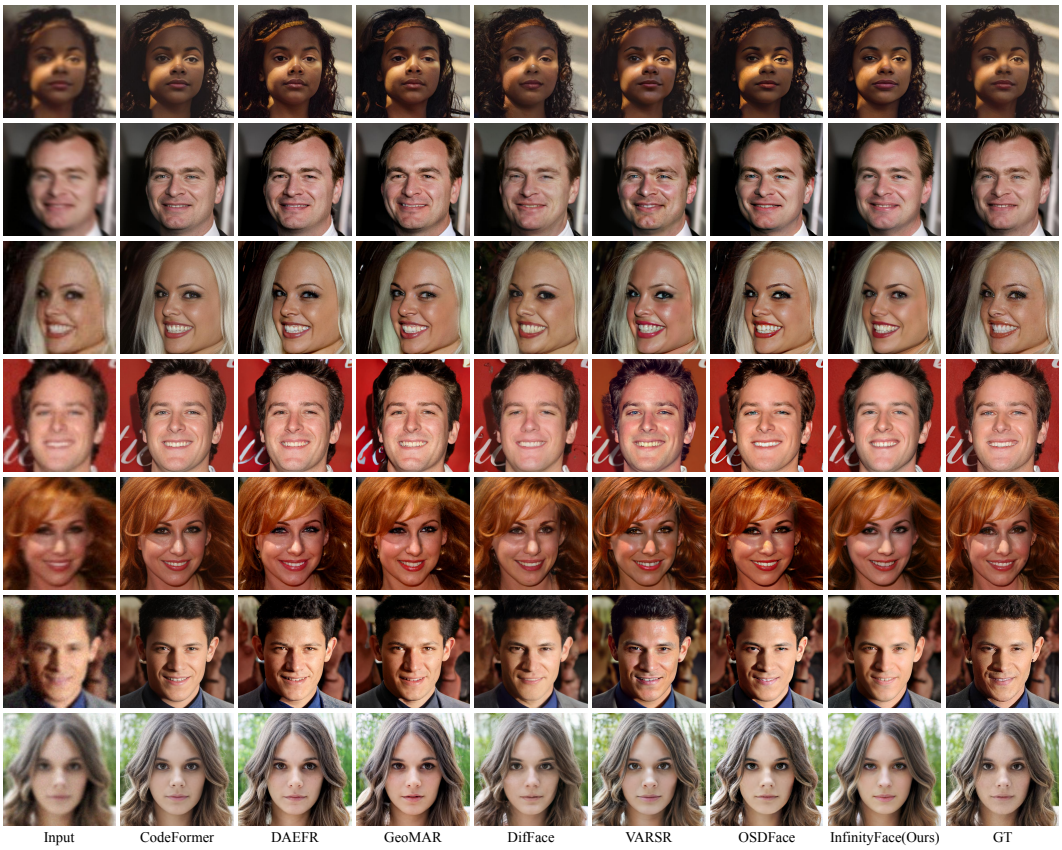


Figure 13: Qualitative comparisons on CelebA-Test.