DicFace: Dirichlet-Constrained Variational Codebook Learning for Temporally Coherent Video Face Restoration

Yan Chen[1*], Hanlin Shang[1*], Ce Liu[1*], Yuxuan Chen[1], Hui Li[1], Weihao Yuan[2], Hao Zhu[3], Zilong Dong[2], Siyu Zhu[1†]
[1]Fudan University [2]Alibaba Group [3]Nanjing University

Abstract

Video face restoration faces a critical challenge in maintaining temporal consistency while recovering fine facial details from degraded inputs. This paper presents a novel approach that extends Vector-Quantized Variational Autoencoders (VQ-VAEs), pretrained on static high-quality portraits, into a video restoration framework through variational latent space modeling. Our key innovation lies in reformulating discrete codebook representations as Dirichlet-distributed continuous variables, enabling probabilistic transitions between facial features across frames. A spatio-temporal Transformer architecture jointly models inter-frame dependencies and predicts latent distributions, while a Laplacian-constrained reconstruction loss combined with perceptual (LPIPS) regularization enhances both pixel accuracy and visual quality. Comprehensive evaluations on blind face restoration, video inpainting, and facial colorization tasks demonstrate state-of-the-art performance. This work establishes an effective paradigm for adapting intensive image priors, pretrained on high-quality images, to video restoration while addressing the critical challenge of flicker artifacts. The source code has been open-sourced and is available at https://github.com/fudan-generative-vision/DicFace.

1. Introduction

Video face restoration aims to reconstruct high-quality face video sequences from degraded inputs. As a significant research branch in the field of computer vision—particularly within low-level vision and face analysis—it has important applications in digital image enhancement, film and media post-production, identity recognition, and security. The video face restoration problem is a domain-specific video restoration task and remains highly ill-posed, necessitating auxiliary guidance—such as facial codebooks [11, 41], geometric priors [6, 37], and reference priors [20, 21]—to enhance fidelity and improve visual details of reconstructed facial images. Meanwhile, image-based face restoration [30, 35, 41] has benefited substantially from advances in visual generative models, which enable both the mapping from degraded low-quality inputs to high-quality outputs and the supplementation of fine details that are absent in the degraded inputs.

One straightforward approach [30, 35, 41] for video face restoration is to directly apply image-based face restoration techniques to each frame of a low-quality input video, producing a sequence of reconstructed high-quality frames. However, simply extending image-based methods [30, 35, 41] to videos often fails to preserve the temporal consistency of reconstructed facial details, as each frame is typically processed with independently defined facial priors or code predictions. An alternative commonly employed strategy is to adopt general video restoration methods [2, 3, 12, 19, 29] and adapt them to the video face restoration domain [9, 33] through data-driven fine-tuning or domain adaptation. Nonetheless, these general video restoration methods often lack crucial facial priors (e.g., facial codebooks, geometric constraints, and reference priors), resulting in insufficiently detailed facial features and diminished visual fidelity in the restored output.

In this paper, we extend the classical vector-quantized autoencoder (VQ-AE)–based face restoration paradigm [8], which is pretrained on large-scale high-fidelity portrait images to form a high-quality codebook prior, from single-frame image restoration to multi-frame video restoration. Traditional methods [8, 11, 41] rely on a Transformer to independently predict discrete codes for each frame via codebook lookup, yet such per-frame processing disregards temporal continuity. Consequently, the latent codes may fluctuate abruptly across adjacent frames, causing perceptual flicker and compromising visual quality. To address this limitation, we propose a novel variational formulation that bridges discrete and continuous representations. Specifically, we relax the discrete codebook by treating con-

* These authors contribute equally to this work.

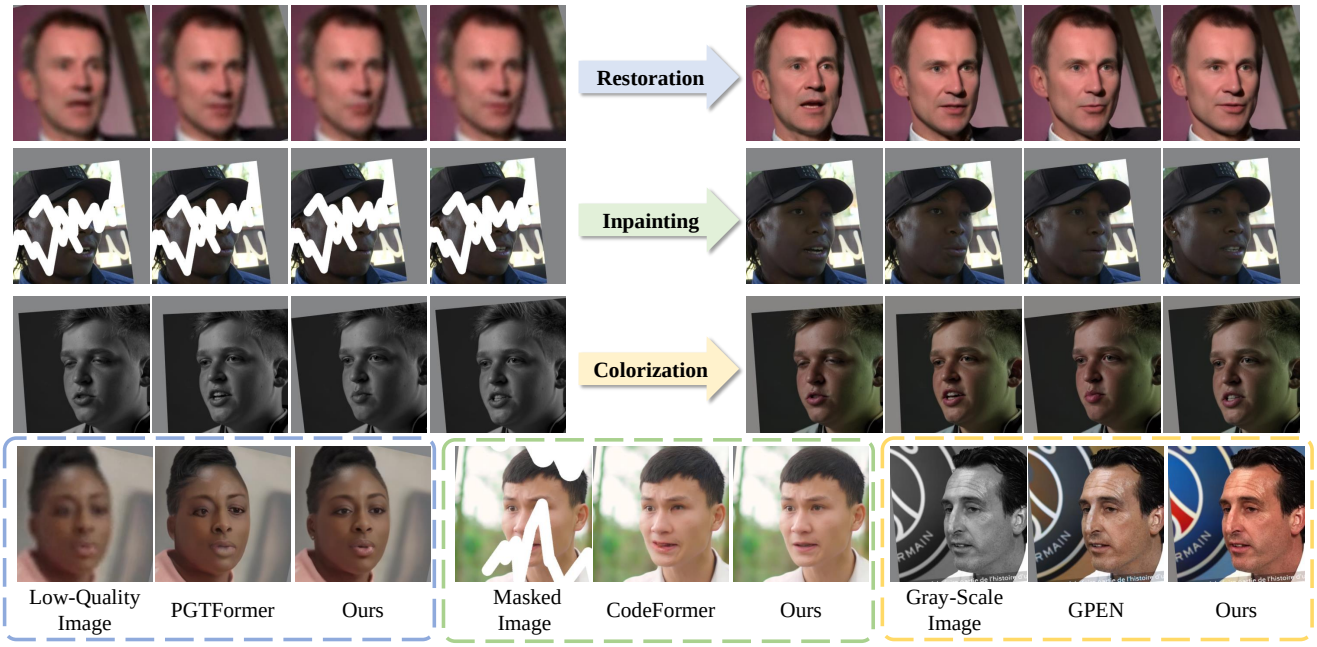


Figure 1. Our innovative facial restoration model exhibits markedly superior performance compared to existing state-of-the-art single-task methods (e.g. PGTFormer [33] for restoration ,CodeFormer [41] for inpainting and GPEN [35] for colorization). We have successfully developed a more precise and natural restoration model that effectively maintains the facial structure and detailed characteristics of subjects. Furthermore, our approach enhances color consistency and ensures temporal continuity. This advancement represents a significant breakthrough in the domain of high-quality facial video and portrait restoration.

vex combinations of its items as Dirichlet-distributed latent variables. A Transformer then learns spatial-temporal dependencies over consecutive frames, while a Laplacian assumption on the reconstruction error together with LPIPS loss [39] promotes perceptually compelling results. By embedding discrete codebook representations into this continuous framework, the proposed strategy effectively mitigates temporal flicker in multi-frame outputs and achieves high-quality reconstructions through a principled variational training scheme.

Extensive experiments on the VFHQ [32] benchmark demonstrate our method's effectiveness: quantitative evaluations on blind video face restoration task reveal that, compared to baseline methods, our approach achieves 1.27dB PSNR and 8.2% LPIPS [39] improvements for image quality. Additionally, in terms of temporal stability, our method shows 5.6% improvement in TLME metric. While qualitative assessments reveal enhanced temporal consistency under challenging conditions like occlusions and rapid motions. Furthermore, ablation studies validate our design choices, particularly the importance of Dirichlet-based variational modeling for temporal coherence. These advancements establish new state-of-the-art performance while providing a promising framework for integrating discrete facial priors with continuous video dynamics.

2. Related Works

Image Face Restoration. The objective of face image restoration is to recover high-quality facial images from low-quality inputs undergoing complex degradations. Existing methods primarily rely on three types of facial priors to mitigate dependence on degraded inputs. Geometric priors [6, 37] capture morphological constraints using landmarks [16, 17, 40], parsing maps [5, 26], and component heatmaps [1]. Reference priors [20, 21], drawing on high-quality examples, facilitate fine-detail restoration and identity preservation (e.g., DFDNet [23] builds a facial component dictionary from VGGFace [25]). Generative priors further enhance restoration by identifying optimal latent codes (PULSE [24]) or by incorporating pretrained StyleGAN [14, 15] into encoder-decoder frameworks (GPEN [35], GFP-GAN [30]). Pretrained vector-quantization codebooks are also adopted to refine fine-grained details, as demonstrated by VQFR [11] and CodeFormer [41]. More recently, DR2 [31] and DifFace [38] leverage diffusion models to remove degradations with strong fidelity and robustness. However, directly extending image-based restoration methods to the video domain often fails to ensure temporal consistency of reconstructed facial details. Concurrently, there is a strong motivation to exploit pretrained prior distributions from single-image restoration—such as high-quality facial codebooks—in video scenarios. This work is driven by the objective of effectively

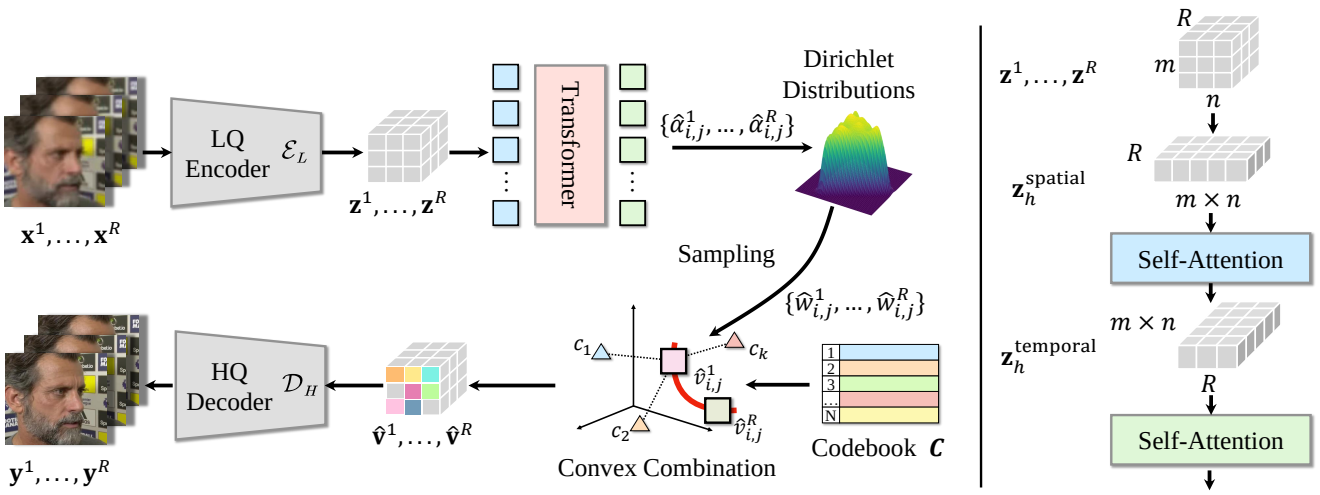


Figure 2. Overview of our framework. The framework processes a sequence of low-quality frames using three core components: (1) an encoder network that extracts spatial features; (2) a spatio-temporal Transformer that models inter-frame dependencies and predicts Dirichlet parameters for latent code distributions; and (3) a decoder that reconstructs high-quality frames from convex combinations of learnable codebook items. Crucially, the latent representation at each spatial location is formulated as a probabilistic mixture of codebook entries, enabling smooth transitions via variational inference over the Dirichlet manifold. This continuous relaxation is regularized by an ELBO objective, balancing reconstruction fidelity with temporal coherence.

harnessing these priors for consistent and robust video face restoration.

General Video Restoration. Video restoration aims to enhance degraded video content to near-lossless quality. FSTRN [19] employs fast spatio-temporal networks with 3D convolutions for feature alignment and motion extraction. EDVR [29] leverages space-time deformable convolutions to aggregate temporal information across adjacent frames. RSDN [12] splits inputs into structural and detail components, passing them through recurrent two-stream modules to refine texture details. BasicVSR [2] and BasicVSR++ [3] utilize optical flow for temporal alignment and integrate bidirectional hidden states from past and future frames, achieving state-of-the-art performance through efficient sequence-wide information fusion. However, when these general video restoration methods are adapted (e.g., via data-driven fine-tuning or domain adaptation) to video face restoration, the lack of face-specific priors—such as facial codebooks, geometric constraints, and reference priors—often results in suboptimal detail recovery and diminished visual fidelity in facial regions.

Codebook Based Learning. Vector-quantized (VQ) codebooks were first introduced in VQ-VAE [28], where the codebook is learned during training rather than predefined. VQ-GAN [8] subsequently incorporates perceptual and adversarial losses, enabling a more compact codebook while preserving or improving expressiveness and visual fidelity. Recent studies [18, 36] further optimize codebook usage through L2 normalization and edge-prior regularization. Learned sparse representations have shown significant benefits in image restoration tasks (e.g., super-resolution [10, 27] and denoising [7]), offering higher efficiency and flexibility compared to traditional manually crafted dictionaries [13, 21]. In the face restoration domain, a high-quality codebook pretrained on extensive high-resolution facial data provides a powerful face-specific prior, with quantization helping to reduce representational ambiguity. To address the inherent limitations of discrete quantization, we relax the latent space constraint, learning and predicting a continuous latent distribution for enhanced reconstruction.

3. Methodology

This section presents our methodology for video face restoration through five components: Section 3.1 establishes fundamental concepts in vector quantization and Dirichlet distributions. Section 3.2 details our continuous latent space formulation using variational Dirichlet inference. Section 3.3 describes the hybrid loss function combining evidence lower bound optimization with perceptual metrics. Section 3.4 outlines the spatial-temporal transformer architecture, while Section 3.5 specifies training protocols and inference strategies for temporal coherence. The framework overview is shown in Figure 2.

3.1. Preliminary

Vector-Quantized Autoencoder. The vector-quantized autoencoder framework aims to reconstruct a high-quality image $\mathbf{y} \in \mathbb{R}^{H\times W\times 3}$ from its degraded counterpart $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$. The architecture comprises four components: an encoder $\mathcal{E}_L$, a decoder $\mathcal{D}_H$, a Transformer network $\mathcal{H}$, and a learnable codebook $\mathbf{c} = [c_k]_{k=1}^N$ with $c_k \in \mathbb{R}^d$.

Specifically, the encoder $\mathcal{E}_L$ first maps $\mathbf{x}$ to a latent feature map $\mathbf{z} \in \mathbb{R}^{m \times n \times d}$. The Transformer $\mathcal{H}$ then processes $\mathbf{z}$ to infer a probability distribution $\mathbf{s} \in \mathbb{R}^{m \times n \times N}$ over the codebook, where each spatial location (i, j) satisfies:

$$\sum_{k=1}^{N} s_{i,j,k} = 1 \quad \text{and} \quad s_{i,j,k} \geq 0. \tag{1}$$

The quantized feature map $\hat{\mathbf{z}}$ is obtained by replacing each $\mathbf{z}_{i,j}$ with the codebook entry c_k corresponding to $\arg\max_k s_{i,j,k}$. Finally, the decoder $\mathcal{D}_H$ reconstructs $\mathbf{y}$ from $\hat{\mathbf{z}}$.

This discrete quantization reduces representational ambiguity by constraining the latent space to the codebook entries. However, in video processing, hard quantization introduces temporal inconsistencies that manifest as flicker artifacts across reconstructed frames.

Dirichlet Distribution. A Dirichlet distribution is a continuous multivariate distribution defined on the $(N-1)$-dimensional simplex. For a parameter vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]$ with $\alpha_k > 0$, its probability density function for $\mathbf{w} = [w_1, \ldots, w_N]$ is given by

$$\mathrm{Dir}(\mathbf{w} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{N} \alpha_k\right)}{\prod_{k=1}^{N} \Gamma(\alpha_k)} \prod_{k=1}^{N} w_k^{\alpha_k - 1}, \tag{2}$$

$$\text{subject to } \sum_{k=1}^{N} w_k = 1 \text{ and } w_k \geq 0. \tag{3}$$

Here, $\Gamma(\cdot)$ denotes the Gamma function. As the conjugate prior for categorical distributions, the Dirichlet provides a probabilistic framework for modeling uncertainty in discrete classifications. In our approach, it enables continuous relaxation of latent code assignments by representing each spatial location's latent code as a convex combination of codebook entries. This probabilistic formulation permits smooth transitions between frames through variational inference over the Dirichlet manifold, effectively mitigating temporal artifacts such as flicker in video reconstructions.

3.2. Continuous Latent Space

To overcome the limitations of discrete quantization, we relax the latent space constraint by allowing it to span the convex hull of the codebook $\{c_k\}_{k=1}^{N}$. Specifically, each latent code $\hat{\mathbf{v}} = [\hat{v}_{i,j}]_{m \times n}$ is modeled as a convex combination of the code items, weighted by $\hat{\mathbf{w}} = [\hat{w}_{i,j}]_{m \times n}$:

$$\hat{v}_{i,j} = \hat{w}_{i,j}^{T} \mathbf{c}, \tag{4}$$

where $\hat{w}_{i,j} \in \mathbb{R}^N$, $\hat{w}_{i,j} \geq \mathbf{0}$, and $\hat{w}_{i,j}^{T} \mathbf{1} = 1$. This formulation ensures that each $\hat{v}_{i,j}$ resides within the convex hull of the learned code items, thereby providing a continuum of feasible latent representations.

We adopt a variational approach to learn and predict the distribution of $\hat{\mathbf{w}}$ via neural networks. Specifically, we approximate the true posterior distribution $p(\hat{\mathbf{w}} \mid \mathbf{x})$ by $q_\theta(\hat{\mathbf{w}} \mid \mathbf{x})$, where θ denotes the network parameters. We further assume:

$$q_\theta(\hat{\mathbf{w}} \mid \mathbf{x}) = \texttt{Dir}(\hat{\alpha} \mid \mathbf{x}), \tag{5}$$

where $\texttt{Dir}$ denotes the Dirichlet distribution and $\hat{\alpha} = [\hat{\alpha}_{i,j}]_{m \times n}$ are the Dirichlet hyper-parameter satisfying $\hat{\alpha}_{i,j} \in \mathbb{R}^N$ and $\hat{\alpha}_{i,j} \geq 0$.

To learn θ, we maximize the evidence lower bound (ELBO):

$$\log p_\theta(\mathbf{y} \mid \mathbf{x}) \geq \mathcal{L}_{\texttt{ELBO}}(\theta; \mathbf{x}, \mathbf{y}), \tag{6}$$

where

$$\begin{aligned} \mathcal{L}_{\texttt{ELBO}}(\theta; \mathbf{x}, \mathbf{y}) = &- \texttt{KL}\Big(q_\theta(\hat{\mathbf{w}} \mid \mathbf{x}) \Big\| p_\theta(\hat{\mathbf{w}})\Big) \\ &+ \mathbb{E}_{q_\theta(\hat{\mathbf{w}} \mid \mathbf{x})}\big[\log p_\theta(\mathbf{y} \mid \mathbf{x}, \hat{\mathbf{w}})\big]. \end{aligned} \tag{7}$$

The first term is the Kullback–Leibler (KL) divergence between the approximate posterior $q_\theta(\hat{\mathbf{w}} \mid \mathbf{x})$ and the prior $p_\theta(\hat{\mathbf{w}})$. Assuming $p_\theta(\hat{\mathbf{w}}) = \texttt{Dir}(\alpha)$, each KL term can be derived in closed form:

$$\begin{aligned} \texttt{KL}(\cdot \| \cdot) = \texttt{C} &+ \log \Gamma\Big(\sum_k \hat{\alpha}_{i,j,k}\Big) - \sum_k \log \Gamma\big(\hat{\alpha}_{i,j,k}\big) \\ &+ \sum_k \Big(\hat{\alpha}_{i,j,k} - \alpha_{i,j,k}\Big)\Big[\psi\big(\hat{\alpha}_{i,j,k}\big) - \psi\Big(\sum_k \alpha_{i,j,k}\Big)\Big], \end{aligned} \tag{8}$$

where $\texttt{C}$ is a constant and $\psi(\cdot)$ denotes the digamma function. When all the items in $\hat{\alpha}_{i,j}$ are close to zero, the sampled weight vector $\hat{w}_{i,j}$ tends to be a one-hot vector. In such cases, only the nearest code items in the codebook would have impact on the latent code $\hat{v}_{i,j}$. Conversely, if all the items in $\hat{\alpha}_{i,j}$ are larger, the sampled weight vector tends to be uniform, and the latent code $\hat{v}_{i,j}$ would be an uniform average of the code items in the codebook.

The second term is the expected reconstruction error, and would encourage the sampled weight vector $\hat{w}_{i,j}$ to produce a latent code that can recover the high-quality image $\mathbf{y}$ as close as possible by the decoder network $\mathcal{D}_H$.

3.3. Loss Function

Given the prediction $\mathbf{y}^{\texttt{pred}}$ and the ground truth $\mathbf{y}$, the overall loss function combines the evidence lower bound (ELBO) and the learned perceptual image patch similarity (LPIPS) loss as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\texttt{ELBO}} + \lambda_2 \mathcal{L}_{\texttt{LPIPS}}, \tag{9}$$

where λ_1 and λ_2 balance the two terms. During training, we set $\lambda_1 = -1.0$ and $\lambda_2 = 1.0$ to prioritize reconstruction fidelity while maintaining perceptual quality.

ELBO Loss. We compute the ELBO following Eq. (7). Specifically, the KL divergence term is computed following Eq. (8) to encourage the predicted $\hat{\alpha}$ to be close to the prior α. For the expected reconstruction error term, we approximate the expectation by the Monte Carlo method. Moreover, we assume the $\log p_\theta(\mathbf{y} \mid \mathbf{x}, \hat{\mathbf{w}})$ follows the Laplacian distribution whose location parameters are $\mathbf{y}^{\texttt{pred}}$ and the scale parameters are 1. Accordingly, the expected reconstruction error term is computed as follows:

$$\begin{aligned}\mathbb{E}_{q_\theta(\hat{\mathbf{w}}|\mathbf{x})}\left[\log p_\theta(\mathbf{y} \mid \mathbf{x}, \hat{\mathbf{w}})\right] &\approx \sum_{\hat{\mathbf{w}}_l \sim q_\theta(\hat{\mathbf{w}}|\mathbf{x})} \log p_\theta(\mathbf{y} \mid \mathbf{x}, \hat{\mathbf{w}}_l) \\ &= \sum_{\hat{\mathbf{w}}_l \sim q_\theta(\hat{\mathbf{w}}|\mathbf{x})} |\mathbf{y} - \mathbf{y}_l^{\texttt{pred}}| + \texttt{C} \end{aligned} \tag{10}$$

where $\hat{\mathbf{w}}_l \sim q_\theta(\hat{\mathbf{w}} \mid \mathbf{x})$ means differentiably sampling $\hat{\mathbf{w}}_l$ from the distribution $q_\theta(\hat{\mathbf{w}} \mid \mathbf{x})$, $\mathbf{y}_l^{\texttt{pred}}$ is the predicted high-quality image when the sampled weight is $\hat{\mathbf{w}}_l$, and $\texttt{C}$ is a constant.

Perceptual Loss. We incorporate LPIPS [39] to enhance visual quality, computed as:

$$\mathcal{L}_{\texttt{LPIPS}} = \|\phi(\mathbf{y}) - \phi(\mathbf{y}^{\texttt{pred}})\|_2^2, \tag{11}$$

where ϕ denotes deep features from a pretrained VGG network.

3.4. Network Architecture

The proposed architecture processes a sequence of R low-quality frames $\{\mathbf{x}^r\}_{r=1}^R$ to restore their high-quality counterparts $\{\mathbf{y}^r\}_{r=1}^R$, as illustrated in Figure 2. The framework comprises three key components:

Encoder. Each input frame $\mathbf{x}^r$ is first encoded into a latent feature map $\mathbf{z}^r \in \mathbb{R}^{m \times n \times d}$ via a convolutional encoder $\mathcal{E}_L$. The encoder reduces spatial resolution through five strided convolutions, yielding $m = H/32$ and $n = W/32$ for input resolution $H \times W$.

Spatio-Temporal Transformer. The Transformer $\mathcal{H}$ with $2H$ alternating attention blocks then process the sequence $\{\mathbf{z}^1, \ldots, \mathbf{z}^R\}$. For odd-indexed blocks $h \in \{1, 3, ..., 2H-1\}$, spatial self-attention is applied by reshaping features into $\mathbf{Z}_h^{\texttt{spatial}} \in \mathbb{R}^{R \times (mn) \times d}$ and computing:

$$\mathbf{Z}_{h+1}^{\texttt{spatial}} = \texttt{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d}}\right) V_h + \mathbf{Z}_h^{\texttt{spatial}}, \tag{12}$$

where $Q_h, K_h, V_h \in \mathbb{R}^{(mn) \times d}$ are query, key, and value projections. For even-indexed blocks, temporal attention operates on $\mathbf{Z}_h^{\texttt{temporal}} \in \mathbb{R}^{(mn) \times R \times d}$ via analogous computations across the temporal dimension. Sinusoidal positional embeddings augment queries and keys to encode spatial/temporal positions.

Code Prediction & Decoder. The Transformer output $\{\mathbf{z}^r\}_{r=1}^R$ is linearly projected to predict Dirichlet parameters $\{\hat{\boldsymbol{\alpha}}^r\}_{r=1}^R$. For each frame, latent codes $\hat{\mathbf{v}}^r$ are sampled by $\hat{\mathbf{w}}^r \sim \texttt{Dir}(\hat{\boldsymbol{\alpha}}^r)$ and computed as $\hat{\mathbf{v}}^r = \sum_{k=1}^N \hat{w}_{i,j,k}^r c_k$ per spatial location. These codes are decoded to $\mathbf{y}^r$ through a mirrored decoder $\mathcal{D}_H$ with five transposed convolutions.

Implementation Details. The encoder/decoder each contain 12 residual blocks and 5 resolution-scaling layers. The Transformer employs $H = 4$ alternating spatial-temporal blocks (8 total) with 8 attention heads. We evaluate codebook sizes $N \in \{256, 512, 1024\}$, with $d = 256$ codes. During training, we progressively unfreeze components: first the encoder/decoder, then the Transformer, and finally the codebook.

3.5. Training and Inference

Training. The model parameters were initialized using a combination of pre-trained weights and random initialization. To investigate the impact of codebook size on model performance, we retrained CodeFormer [22] with codebook sizes of 256 and 512 on the FFHQ dataset [14], which consists of aligned and resized facial images at a resolution of 512x512. During the model parameter initialization phase, we loaded the pre-trained weights of CodeFormer. To ensure temporal stability in the task of video face restoration, we incorporated a 9-layer Multi-Head Attention (MHA) Temporal Transformer, whose parameters were randomly initialized based on a Gaussian distribution. Throughout the training process, the codebook parameters remained frozen. The newly added Temporal Transformer parameters were consistently trained, and we also experimented with freezing the encoder and decoder to assess their impact on performance metrics and results. The experiments demonstrated that allowing the encoder and decoder parameters to participate in training yielded superior image quality and temporal stability.

Inference. During inference, we process input videos using a sliding window approach with a stride of 1 frame. Each window consists of 5 consecutive frames, padded at the sequence boundaries by replicating the initial and final frames. The network predicts the restored central frame (third position) for each window. Final video reconstructions are obtained by aggregating these center frame predictions, ensuring temporal coherence through overlapping window processing. This strategy effectively balances computational efficiency with temporal consistency preservation.

4. Experiments

4.1. Experimental Settings

Datasets. The proposed method is trained utilizing the VFHQ [32] dataset, comprising 16000 video clips with na-

Figure 3. Qualitative comparison with other state-of-the-art methods on the VFHQ-Test dataset for blind face restoration. Our model demonstrates superior performance in recovering finer details and achieving significantly better temporal consistency, particularly under challenging conditions such as large facial angles, body occlusions, and substantial facial movements.

Method	PSNR↑	SSIM↑	LPIPS↓	IDS↑	AKD↓	FVD↓	TLME↓
GPEN	26.509	0.739	0.341	0.856	2.920	405.926	1.641
GFPGAN	27.221	0.775	0.311	0.861	2.998	359.197	1.223
CodeFormer	26.064	0.740	0.320	0.781	3.479	510.034	1.530
RealBasicVSR	26.030	0.715	0.407	0.811	3.181	635.216	1.777
BasicVSR++	27.001	0.775	0.409	0.826	3.513	823.908	1.598
PGTFormer	27.829	0.786	0.292	0.879	2.566	**332.340**	1.333
KEEP	27.810	0.797	0.268	0.863	2.466	378.72	1.156
Ours	**29.099**	**0.831**	**0.246**	**0.908**	**2.093**	336.015	**1.091**

Table 1. Quantitative comparison on the VFHQ-Test dataset for blind video face restoration.

tive 512×512 resolution.

For blind face restoration (BFR) task, we adopt the degradation pipeline from VFHQ [32]. We put the details in Appendix. For the colorization task, we converted the videos to grayscale, and for the inpainting task, we applied brush stroke masks [35] to randomly draw irregular polyline masks for generating masked faces. During evaluation, facial alignment preprocessing is applied to test samples to ensure compatibility with baseline methods requiring geometrically normalized inputs.

Settings and Metrics. We introduce the settings and metrics in Appendix.

4.2. Comparisons with State-of-the-Art Methods

Restoration For the video face restoration task, we perform comparisons with state-of-the-art methods, including KEEP [9] and PGTFormer [33]. We also present the results of BasicVSR++ [3] and RealBasicVSR [4], which are designed for general-purpose video restoration. The results of single-image face restoration frameworks including CodeFormer [22], GFPGAN [30] and GPEN [35] are also presented for comparison. The quantitative results are in Table 1. We achieve the best performance under all the metrics except for the FVD. Especially, our framework reduced the TLME from 1.156 to 1.091, showing better temporal consistency. The visualization of the predictions is shown

Methods	PSNR↑		SSIM↑		LPIPS↓		IDS↑		AKD↓		FVD↓		TLME↓	
	inp.	col.	inp.	col.	inp.	col.	inp.	col.	inp.	col.	inp.	col.	inp.	col.
GPEN	28.170/24.364		0.919/0.945		0.156/0.204		0.902/0.999		2.742/1.350		649.8/212.1		4.779/1.181	
CodeFormer	31.463/18.505		0.949/0.682		0.144/0.457		0.927/0.711		2.645/9.082		224.3/826.5		1.744/3.264	
PGDiff	23.146/21.735		0.853/0.862		0.236/0.388		0.859/0.973		5.929/2.111		590.7/470.2		1.831/1.667	
Ours	**31.630/26.885**		**0.953/0.962**		**0.069/0.141**		**0.945/0.999**		**1.379/0.907**		**147.7/155.7**		**1.290/0.964**	

Table 2. Quantitative comparison on the VFHQ-Test for video face inpainting and colorization. "inp.": inpainting, "col.": colorization.

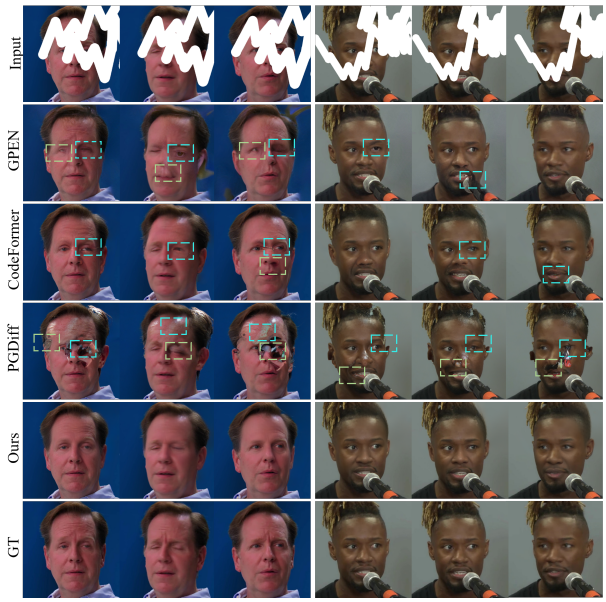


Figure 4. Visual comparison with advanced inpainting methods on the VFHQ-Test dataset for challenging inpainting cases. CodeFormer and other methods struggle to restore details like eyes and lips. In contrast, our method is able to recover these details reasonably well, achieving a more realistic result.

in Figure 3. Our framework produces higher-quality images with richer details.

We also evaluate the temporal consistency in Figure 6 and 7. In Figure 6, our predictions show lower errors at facial landmarks. In Figure 7, our framework ensures temporal consistency and mitigates jitters across frames.

In addition, we evaluate the methods on out-of-domain data. The results and analysis are in Appendix.

Inpainting In Table 2, we compare with state-of-the-art frameworks for face inpainting, including GPEN [35], CodeFormer [22], and PGDiff [34]. We found our framework achieves much lower LPIPS, AKD and FVD, meaning that our prediction is closer to the ground truth under these metrics. In Figure 4 we present the predictions of different methods. Our framework produces correct eyes and mouses. We put more comparisons about the temporal variations in Appendix.

Figure 5. Visual comparison of colorization results on the VFHQ-Test dataset. Other methods yield unrealistic colorization, which include issues such as unnatural skin tones. In contrast, our method generates more realistic results, accurately capturing the natural hues and achieving a more lifelike appearance.

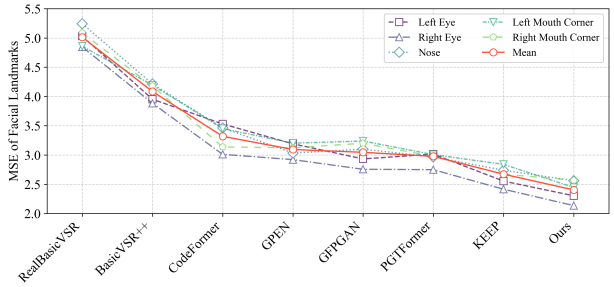


Figure 6. Comparison of temporal stability with other state-of-the-art methods on the VFHQ-Test dataset. In the context of blind face restoration, we calculate the discrepancies between the five facial landmarks of the output results and the ground truth for each method. These results indicate that our method achieves the lowest mean squared error (MSE) at each landmark, demonstrating the best temporal stability among all evaluated approaches.

Colorization We evaluate the colorization ability of our framework and compare it with GPEN [35], CodeFormer [22], and PGDiff [34]. The results in Table 2 show our framework achieves best performance under all the metrics. Especially, we reduce FVD from 212.075 to 155.727, and reduce TLME from 1.181 to 0.964, meaning that our

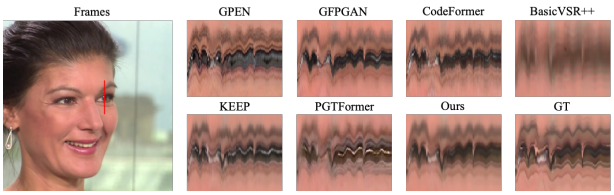


Figure 7. Comparison of temporal stability with other state-of-the-art methods for face restoration on the VFHQ-Test dataset. We selected a column (red line) along the subject's eye and plotted its temporal variations over time. Our method exhibits significantly mitigated temporal jitter, enhancing spatial consistency across restored frames and preserving temporal continuity over time.

predictions are closer to the ground truth and have better temporal consistency. In Figure 5, our framework produces more realistic colors.

4.3. Ablation Studies

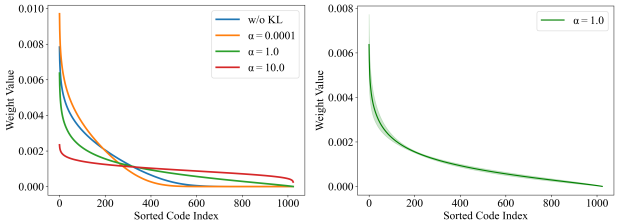
We perform ablations to demonstrate the effectiveness of our design. All the experiments are performed on VFHQ-Test dataset for the task of blind video face restoration. More experiments about the temporal Transformer, codebook size, and number of samples are in Appendix.

Dirichlet Prior. We investigate the effects of Dirichlet prior hyper-parameters α for blind video face restoration on VFHQ-Test dataset. The results are shown in Table 3. We found that the prior constraint by KL loss is beneficial for the face restoration task. Specifically, we achieve the lowest FVD when the α takes 1.0. In Figure 8, we visualize the weight vector $\hat{\mathbf{w}}$ predicted on an example image. In Figure 8 (a), for each "pixel" we sort the weight for code items in descending order, and compute the average weight vector over all the pixels. We found when α takes smaller value, the sampled weights focus on the first fewer items. When α takes larger value, the sampled weights are more uniform. We also show the variance of the weight values when $\alpha = 1$ in Figure 8 (b). The weight values are stable over different pixels.

α	PSNR↑	AKD↓	FVD↓	TLME↓
w/o	28.929	**2.088**	336.076	**1.091**
0.0001	29.091	2.093	337.365	1.112
1.0	**29.099**	2.093	**336.015**	1.107
10.0	29.003	2.105	344.922	1.106

Table 3. Ablation for Dirichlet prior hyper-parameter α for Blind Face Restoration on VFHQ-Test dataset. "w/o" means no KL loss.

Code Aggregation Strategy. We evaluate three code aggregation approaches: (1) "Top-K" selection with hard weight assignment during training and soft averaging during inference (K=4, 16), (2) "Average" using learned soft weights without sparsity constraints, and (3) our probabilistic aggregation with Dirichlet prior. Table 4 demonstrates that our method achieves superior performance across all metrics. Compared to top-16, our approach improves PSNR by 0.594 dB and reduces FVD by 4.1%, validating that explicit modeling of code combination probabilities through variational inference better captures facial feature distributions. The 5.7% reduction in AKD versus the Average baseline highlights enhanced structural preservation through our sparsity-inducing Dirichlet prior.



(a) Weight values from different α. (b) Variance of weight items.

Figure 8. Visualization of weight vector $\hat{w}_{i,j}$ which is used for averaging the code items. The weight items have been sorted in descending order. (a) presents predicted weight values trained with different Dirichlet prior hyper-parameters α. "w/o" means no KL loss. In (b) we plot the variance (green shading) of weight values for Dirichlet prior hyper-parameter $\alpha = 1$. The above results are collected during inference on a randomly selected video clip from VFHQ-Test dataset for blind face restoration.

Method	PSNR↑	AKD↓	FVD↓	TLME↓
Top-4	28.178	2.415	361.160	1.217
Top-16	28.505	2.283	349.784	1.140
Average	29.035	2.128	342.141	1.114
Ours	**29.099**	**2.093**	**336.015**	**1.107**

Table 4. Ablation study of code aggregation strategies on VFHQ-Test for blind face restoration. "Top-K": hard top-1 selection during training, top-K averaging during inference. "Average": learned soft weights without constraints.

5. Conclusion

We present a novel video restoration framework that reformulates the latent space of codebook-based VAEs as a probabilistic mixture over codebook entries via Dirichlet distributions. By replacing discrete quantization with continuous convex combinations regularized through variational inference, our approach enables smooth temporal transitions while preserving reconstruction fidelity. The spatio-temporal Transformer architecture effectively captures inter-frame dependencies, predicting Dirichlet parameters that balance the ELBO objective's reconstruction and coherence terms. This work bridges the gap between discrete codebook constraints and continuous video dynamics, offering a principled direction for enhancing generative models in video processing tasks.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 717–732. Springer, 2016. 2

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 1, 3

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 1, 3, 6

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022. 6

[5] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021. 2

[6] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018. 1, 2

[7] Michael Elad and Michal Aharon. Image denoising via learned dictionaries and sparse representation. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, pages 895–900. IEEE, 2006. 3

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 3

[9] Ruicheng Feng, Chongyi Li, and Chen Change Loy. Kalman-inspired feature propagation for video face super-resolution. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 1, 6

[10] Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convolutional sparse coding for image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1823–1831, 2015. 3

[11] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 1, 2

[12] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 645–660. Springer, 2020. 1, 3

[13] Younghyun Jo and Seon Joo Kim. Practical single-image super-resolution using look-up table. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 691–700, 2021. 3

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[16] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 2

[17] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 88–97, 2017. 2

[18] Adrian Łańcucki, Jan Chorowski, Guillaume Sanchez, Ricard Marxer, Nanxin Chen, Hans JGA Dolfing, Sameer Khurana, Tanel Alumäe, and Antoine Laurent. Robust training of vector quantized bottleneck models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 3

[19] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019. 1, 3

[20] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 272–289, 2018. 1, 2

[21] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European conference on computer vision*, pages 399–415. Springer, 2020. 1, 2, 3

[22] Guangming Liu, Xin Zhou, Jianmin Pang, Feng Yue, Wenfu Liu, and Junchao Wang. Codeformer: A gnn-nested transformer model for binary code similarity detection. *Electronics*, 12(7):1722, 2023. 5, 6, 7

[23] Ting Liu, Zhen Lei, Jun Wan, and Stan Z Li. Dfdnet: Discriminant face descriptor network for facial age estimation. In *Biometric Recognition: 10th Chinese Conference, CCBR 2015, Tianjin, China, November 13-15, 2015, Proceedings 10*, pages 649–658. Springer, 2015. 2

[24] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 2

[25] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 2

[26] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8260–8269, 2018. 2

[27] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12*, pages 111–126. Springer, 2015. 3

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[29] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 3

[30] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021. 1, 2, 6

[31] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. 2

[32] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 2, 5, 6

[33] Kepeng Xu, Li Xu, Gang He, Wenxin Yu, and Yunsong Li. Beyond alignment: blind video face restoration via parsing-guided temporal-coherent transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1489–1497, 2024. 1, 2, 6

[34] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *Advances in Neural Information Processing Systems*, 36: 32194–32214, 2023. 7

[35] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 672–681, 2021. 1, 2, 6, 7

[36] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[37] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018. 1, 2

[38] Zongsheng Yue and Chen Change Loy. Diffface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 5

[40] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 94–108. Springer, 2014. 2

[41] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 1, 2