# Efficient Video Face Enhancement with Enhanced Spatial-Temporal Consistency

Yutong Wang[1] Jiajie Teng[2] Jiajiong Cao Yuming Li Chenguang Ma Hongteng Xu[3]
Dixin Luo[1*]
[1]Beijing Institute of Technology [2]Zhejiang University [3]Renmin University of China

## Abstract

*As a very common type of video, face videos often appear in movies, talk shows, live broadcasts, and other scenes. Real-world online videos are often plagued by degradations such as blurring and quantization noise, due to the high compression ratio caused by high communication costs and limited transmission bandwidth. These degradations have a particularly serious impact on face videos because the human visual system is highly sensitive to facial details. Despite the significant advancement in video face enhancement, current methods still suffer from $i$) long processing time and $ii$) inconsistent spatial-temporal visual effects (e.g., flickering). This study proposes a novel and efficient blind video face enhancement method to overcome the above two challenges, restoring high-quality videos from their compressed low-quality versions with an effective de-flickering mechanism. In particular, the proposed method develops upon a 3D-VQGAN backbone associated with spatial-temporal codebooks recording high-quality portrait features and residual-based temporal information. We develop a two-stage learning framework for the model. In Stage I, we learn the model with a regularizer mitigating the codebook collapse problem. In Stage II, we learn two transformers to look up code from the codebooks and further update the encoder of low-quality videos. Experiments conducted on the VFHQ-Test dataset demonstrate that our method surpasses the current state-of-the-art blind face video restoration and de-flickering methods on both efficiency and effectiveness. Code is available at* https://github.com/Dixin-Lab/BFVR-STC.

## 1. Introduction

Real-world videos usually suffer from varying degradations due to shooting methods, transmission compression, etc., which leads to blurring, noise, brightness flickering, color jitter, and other perturbations. With the development of video generation technologies such as Sora [32], AI-generated videos have few problems with content quality. Still, the inherent randomness of the generative model often leads to inconsistency between frames thus causing AI-based (pixel) flickering, resulting in its limited practical applications. In recent years, deep learning techniques have greatly contributed to the development of the computer vision field, giving rise to a series of works dedicated to achieving video enhancement, including video restoration [28, 44], video super-resolution [19, 38], video deflickering [13, 26], video colorization [51, 55], and so on.

The face usually serves as a common focal point or region of interest in media such as images and videos, and a huge amount of research has been devoted to realizing face restoration. Blind face restoration (BFR) aims to reconstruct high-quality faces from low-quality faces suffering from multiple degradations [27]. Numerous studies have addressed blind face image restoration (BFIR) by leveraging face geometry priors [6, 7, 15], generative priors [3, 20, 21, 50], and codebook priors [14, 47, 58]. However, applying BFIR methods directly to degraded videos frame-by-frame will lead to inconsistencies in background and portrait features between frames due to the lack of temporal constraints. Consequently, blind face video restoration (BFVR) approaches typically adopt carefully crafted strategies to enhance the inter-frame correlation [12, 41, 49]. Despite these advancements, these works suffer from inefficiency, e.g., long inference times, and an inadequate temporal perceptive field. KEEP [12] needs an additional restoration method RealESRGAN [46] and face detection model to process background and faces separately, while StableBFVR [41] requires BasicVSR++ [4] for preliminary restoration. Additionally, PGTFormer [49] only considers two neighboring frames in its perceptive field, making it difficult to achieve global consistency across the video.

To improve the above BFVR paradigms, in this study, we propose a novel 3D-VQGAN framework to enable efficient video-level compression and quantization. To conquer the instability and artifacts caused by video-level compression, we leverage a pre-trained feature network, such as DINOv2, as a more powerful discriminator, followed by multiple multi-scale discriminator heads. In terms of quan-

*Corresponding author: dixin.luo@bit.edu.cn.

tization, the intrinsic properties of videos motivated us to design spatial-temporal codebooks, where the spatial codebook is used to record high-quality portrait features and the temporal codebook to record residual-based temporal information. We also propose a marginal prior regularization to alleviate the codebook collapse problem caused by nearest neighbor retrieval as well as multiple codebooks.

The overall video face enhancement framework comprises two stages. The Stage I trains the spatial-temporal codebooks and high-quality (HQ) auto-encoders by reconstruction. The Stage II predicts the code sequences of spatial-temporal compressed tokens in two codebooks by two code lookup transformers respectively. The training of Stage II fixes the codebooks and decoder learned in Stage I and trains the code lookup modules and low-quality (LQ) encoder by HQ-LQ video pairs. The proposed enhancement framework can be used to remove degradations such as low resolution, noise, and blurring in BFR tasks, and can also be used to remove brightness flickering in real-world videos and flickering due to intra-frame inconsistencies in AI-generated videos.

In summary, the contribution of our work includes:

- We propose a novel video face enhancement framework that efficiently solves the BFVR and de-flickering tasks.
- We propose a 3D-VQGAN backbone and a powerful discriminator that can support effective video compression.
- We propose spatial-temporal codebooks for video quantization and marginal prior regularization to mitigate the codebook collapse problem.
- We conduct comprehensive experiments on VFHQ-Test on BFVR and de-flickering tasks and compare ours to state-of-the-art methods. Experimental results demonstrate the effectiveness of our method.

## 2. Related Works

### 2.1. Blind Face Restoration

Early blind face image restoration (BFIR) methods typically rely on geometry priors-based methods [6, 7, 15]. However, for highly degraded images, the geometry information such as face landmarks and face heatmaps often deviates drastically from those of the real images, limiting their effectiveness. Subsequently, researchers attempt to introduce the power of generative models as the generative prior for better visual effects and model robustness. The works in [3, 21, 50] adopt the pre-trained StyleGAN [20] as the generator, while the works in [29] leverage pre-trained Stable Diffusion as prior information to achieve realistic restoration. The works in [14, 47, 58] propose to train a VQGAN with a codebook to learn discrete high-quality portrait features, which can be seen as a special case of generative prior. On the basis of BFIR methods, blind face video restoration (BFVR) methods focus more on the temporal consistency between frames. For example, PGTFormer [49] proposes to derive face parsing context cues from input video to reduce artifacts and jitters. StableBFVR [41] inserts temporal layers in the diffusion process to incorporate temporal information into the generative prior. KEEP [12] employs Kalman filtering principles to guide the restoration of the current frame using previously restored frames.

### 2.2. Video De-flickering

Video flickering is common in real-world and AI-generated videos, resulting in poor visual viewing. The reasoning for flickering in two types of video is different. In real-world videos, flickering usually comes from the shooting equipment and environment. For example, the unstable brightness flickering in old films is usually due to outdated cameras that can't set a fixed exposure time [8]. Prior works require additional guidance to remove the flicker, such as flickering frequency [18] and counterpart unprocessed video [2]. In contrast, pixels from inconsistent texture and details between frames are responsible for flickering in AI-generated videos. FastBlend [9, 10] adopts intra-frame fuzzy interpolation to solve flickering in AI-generated videos, but the generated results suffer from detail loss and blurring. The work in [26] utilizes the neural atlas with a neural filtering strategy to learn consistent features and provide temporal guidance. Our proposed video face enhancement framework can effectively and efficiently conquer both the brightness flickering in real-world face videos and pixel flickering in AI-generated face videos.

### 2.3. Codebook Collapse

Vector quantization (VQ) maps continuous feature vectors into a discrete space by projecting them to the closest items in a codebook, which has been widely used in various vision and language tasks [34, 35, 57]. However, optimizing the code items in vector quantization models, such as VQ-VAE [43] and VQ-GAN [11], is not toilless. The nearest neighbor retrieval method [53] and the stop-gradient operation caused by quantization [16] lead to the codebook collapse problem, which means only a small subset of items receive gradients for optimization, while the majority remain unused and unupdated. The strenuous effort has been made to alleviate codebook collapse, such as Exponential Moving Averages (EMA) [43], codebook reset [24, 52], probabilistic sampling [1, 40], and anchor updating [56]. The work in [53] proposed to regularize the tokens used for quantization with a pre-defined uniform distribution. However, it relies on retrieved indexes to count the frequency of each code item, which is still unfair to inactive code items. To address this limitation, we propose marginal prior regularization, which counts the frequency of code items by accumulating similarity scores, instead of one-hot encodings, further improving the utilization of codebooks.

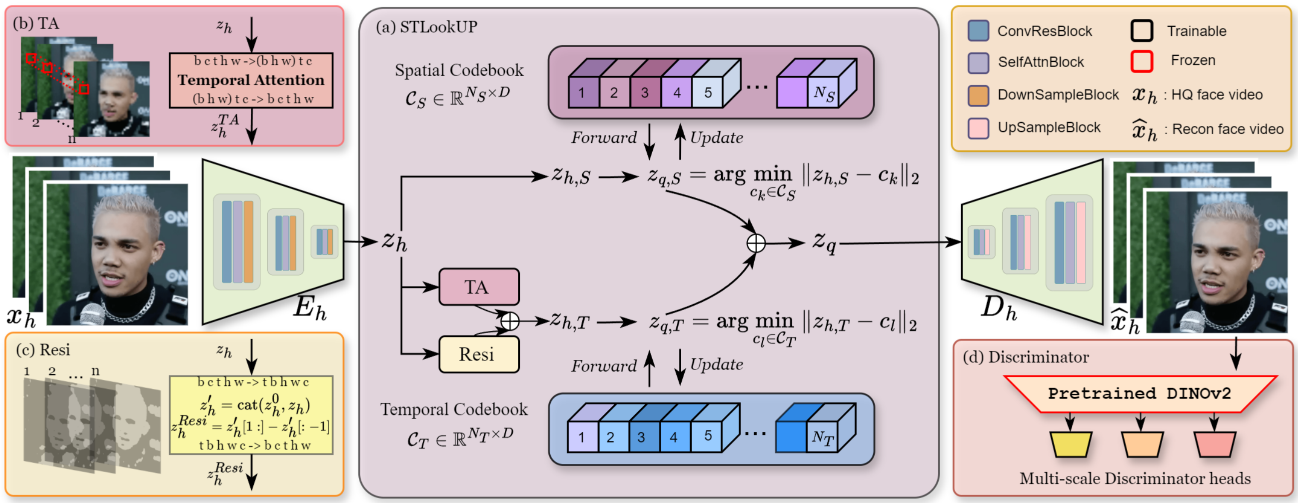



Figure 1. **Network architecture of Stage I.** Stage I uses HQ face videos to train **HQ 3D-VQGAN** ($E_h$ and $D_h$) and **spatial and temporal codebooks** ($\mathcal{C}_S$ and $\mathcal{C}_T$). **(a)** illustrates the quantization operation STLookUp through two codebooks in our proposed framework. **(b)** and **(c)** display the computation process of temporal attention and motion residual, respectively. **(d)** We leverage a pre-trained feature network DINOv2 and trainable multi-scale discriminator heads to construct a more powerful discriminator for stable training.

# 3. Methodology

## 3.1. Overview

The primary objective of video face enhancement is to reconstruct high-quality (HQ) face videos from heavily-degraded low-quality (LQ) face videos, which may suffer from downsampling, blurring, noise, flickering, and color jitters. The limitations of existing video face enhancement methods are the inefficiencies arising from long processing chains and the limited temporal perceptive field that cannot guarantee global consistency. We extend the VQGAN paradigm to the video domain to more effectively capture spatial-temporal information for a more direct and comprehensive video face enhancement. The proposed framework comprises two stages, as illustrated in Fig 1 and Fig 2. In Stage I, we utilize HQ videos to train temporal and spatial codebooks under a 3D-VQGAN backbone, through which the discrete HQ facial representations and temporal shifts can be stored in the codebooks in a self-supervised manner. In Stage II, using HQ-LQ video pairs, we incorporate two transformers to predict temporal and spatial code indices of patches from LQ inputs.

## 3.2. Codebook Learning (Stage I)

In Stage I, we pre-train a 3D-VQGAN model and spatio-temporal codebooks with HQ face videos, where the spatial codebook records HQ facial representations and the temporal one records motion dynamics. We design a more powerful discriminator to stabilize model training and reduce artifacts. Besides, we introduce a marginal prior regularization to mitigate the issue of codebook collapse.

**3D-VQGAN.** The input consists of a sequence of HQ face video frames denoted as $\boldsymbol{x}_h \in \mathbb{R}^{T\times H\times W\times 3}$. These frames are fed into the 3D encoder $E_h$, which produces spatial-temporal compressed latent representations $\boldsymbol{z}_h \in \mathbb{R}^{t\times h\times w\times D}$. With meticulously designed spatial-temporal codebooks $\boldsymbol{\mathcal{C}}_S$ and $\boldsymbol{\mathcal{C}}_T$, we derive the quantized latent representations $\boldsymbol{z}_q \in \mathbb{R}^{t\times h\times w\times D}$ through nearest neighbor (NN) codebook retrieval. Finally, the 3D decoder $D_h$ reconstructs the quantized representations into the output $\widehat{\boldsymbol{x}}_h \in \mathbb{R}^{T\times H\times W\times 3}$.

$$\begin{aligned} \boldsymbol{z}_h &= E_h(\boldsymbol{x}_h;\theta_{E_h}), \\ \boldsymbol{z}_q &= \mathrm{STLookUp}(\boldsymbol{z}_h;\boldsymbol{\mathcal{C}}_S,\boldsymbol{\mathcal{C}}_T), \\ \widehat{\boldsymbol{x}}_h &= D_h(\boldsymbol{z}_q;\theta_{D_h}). \end{aligned} \tag{1}$$

The encoder and decoder are implemented as purely convolutional structures to support faster processing efficiency while accommodating inputs of various resolutions. The autoencoder is constructed by multiple blocks, each containing a residual block, a down/up-sampling block, and a convolution-based self-attention block. Spatial-temporal compression is achieved through downsampling layers, where the degree of spatial compression is larger than that of temporal compression so that both spatial-only and spatial-temporal versions of the sampling layer will exist. Guided by the design strategy in prior work [23], we didn't consider the temporal-only sampling layer.

Due to the instability and artifact-prone nature of training video-level VQGANs, we propose to leverage a partially initialized and more powerful discriminator. Specifically, the discriminator consists of a frozen pre-trained feature network $\mathcal{F}$, such as DINOv2 [33, 36, 37], paired with a set of trainable lightweight MLP-based discriminator heads $\mathcal{D}_{\phi,k}$. For a reconstructed video $\widehat{\boldsymbol{x}}_h(\theta)$, the adversarial loss of the model incorporates the outputs from multiple discriminator heads, such as

$$D_{\phi}\left(\widehat{\boldsymbol{x}}_h(\theta)\right) = -\mathbb{E}_{\boldsymbol{x}_h}\left(\sum_k \mathcal{D}_{\phi,k}\left(\mathcal{F}\left(\widehat{\boldsymbol{x}}_h(\theta)\right)\right)\right), \tag{2}$$
$$\mathcal{L}_{adv}\left(\boldsymbol{x}_h, \theta, \phi\right) = [\log D_{\phi}(\boldsymbol{x}_h) + \log(1 - D_{\phi}(\widehat{\boldsymbol{x}}_h(\theta)))].$$

**Spatial-temporal codebooks.** Compared to image face enhancement, the difficulty of video face enhancement lies in maintaining the consistency of facial features and background details between frames. The discretization nature of a codebook dictates that applying it directly to video tasks often results in poor temporal consistency, leading to the common flickering problem. In addition, conventional codebooks are limited to capturing spatial features and fail to account for the motion information embedded in the videos. To tackle the above challenges, we propose learnable motion-aware spatial-temporal codebooks, where the spatial codebook $\boldsymbol{\mathcal{C}}_S = \{\boldsymbol{c}_k \in \mathbb{R}^D\}_{k=0}^{N_S}$ records portrait features and the temporal codebook $\boldsymbol{\mathcal{C}}_T = \{\boldsymbol{c}_l \in \mathbb{R}^D\}_{l=0}^{N_T}$ stores motion residuals between frames.

Given the compressed latent representation $\boldsymbol{z}_h \in \mathbb{R}^{t\times h\times w\times D}$ output by $E_h$, we first calculate the spatial and temporal latents, respectively. The spatial latents $\boldsymbol{z}_{h,S}$ are directly derived from $\boldsymbol{z}_h$, whereas the temporal latents $\boldsymbol{z}_{h,T}$ incorporate inter-frame temporal-attention (TA) information and motion residuals, defined as follows:

$$\boldsymbol{z}_{h,S} = \boldsymbol{z}_h; \quad \boldsymbol{z}_{h,T} = \text{TA}(\boldsymbol{z}_h) + \text{Residual}(\boldsymbol{z}_h) \tag{3}$$

where the motion residuals are defined as the difference between latents of two frames separated by a time window, as illustrated in Fig 1 (b) and (c). Subsequently, we can calculate the code (indices) sequences and quantized latents corresponding to spatial and temporal latents respectively based on nearest neighbor retrieval:

$$\begin{aligned}
I_S^{(i,j)} &= \arg\min_k \|\boldsymbol{z}_{h,S}^{(i,j)} - \boldsymbol{c}_k\|_2 \in \{0,\dots,N_S-1\},\\
\boldsymbol{z}_{q,S}^{(i,j)} &= \arg\min_{\boldsymbol{c}_k\in\boldsymbol{\mathcal{C}}_S} \|\boldsymbol{z}_{h,S}^{(i,j)} - \boldsymbol{c}_k\|_2 \in \mathbb{R}^D;\\
I_T^{(i,j)} &= \arg\min_l \|\boldsymbol{z}_{h,T}^{(i,j)} - \boldsymbol{c}_l\|_2 \in \{0,\dots,N_T-1\},\\
\boldsymbol{z}_{q,T}^{(i,j)} &= \arg\min_{\boldsymbol{c}_l\in\boldsymbol{\mathcal{C}}_T} \|\boldsymbol{z}_{h,T}^{(i,j)} - \boldsymbol{c}_l\|_2 \in \mathbb{R}^D.
\end{aligned} \tag{4}$$

The quantized latents can be obtained by considering two types of latents. The Eqn. 4 and Eqn. 5 correspond to the `STLookUp` operation in Eqn. 1.

$$\boldsymbol{z}_q = \boldsymbol{z}_{q,S} \oplus \boldsymbol{z}_{q,T} \tag{5}$$

where the fusion operator $\oplus$ is set as element-wise addition by default.

**Marginal prior regularization.** Given the flattened spatial latents $\boldsymbol{z}_{h,S}^f \in \mathbb{R}^{(t\cdot h\cdot w)\times D}$ and spatial codebook $\boldsymbol{\mathcal{C}}_S \in \mathbb{R}^{N_S\times D}$, we can calculate the Euclidean distance matrix $\boldsymbol{D}_E \in \mathbb{R}^{(t\cdot h\cdot w)\times N_S}$ between them. From this, we derive the similarity scores $\boldsymbol{s} = \sigma_r(1/\boldsymbol{D}_E) = [s_{i,j}] \in \mathbb{R}^{(t\cdot h\cdot w)\times N_S}$, where $\sigma_r(\cdot)$ means applying normalization operation to each row of the matrix. The posterior distribution is approximated by the marginal distribution of the similarity score matrix $\boldsymbol{s}$, e.g., $P_{post} = \sum_i^{thw} s_{i,j} = [p_1, p_2, \cdots, p_{N_S}] \in \mathbb{R}^{N_S}$. The uniform prior distribution is defined as $P_{prior} = [1/N_S, 1/N_S, \cdots, 1/N_S] \in \mathbb{R}^{N_S}$. We leverage KL-divergence to implement the marginal prior regularization:

$$\mathcal{L}_{KL}^S = \text{KL}\left(P_{post}, P_{prior}\right) = -\sum_n^{N_S} p_n \log\left(\frac{1/N_S}{p_n}\right). \tag{6}$$

Similarly, we can calculate $\mathcal{L}_{KL}^T$ between flattened temporal latents $\boldsymbol{z}_{h,T}^f$ and temporal codebook $\boldsymbol{\mathcal{C}}_T$.

**Training objectives.** In stage I, we adopt the loss functions commonly used in VQGAN training, including reconstruction loss $\mathcal{L}_1$, perceptual loss [17, 54] $\mathcal{L}_{per}$, adversarial loss [11] $\mathcal{L}_{adv}$ and code-level loss [11, 43] $\mathcal{L}_f$:

$$\begin{aligned}
&\mathcal{L}_1 = \|\boldsymbol{x}_h - \widehat{\boldsymbol{x}}_h\|_1, \quad \mathcal{L}_{per} = \|\Phi(\boldsymbol{x}_h) - \Phi(\widehat{\boldsymbol{x}}_h)\|_2^2,\\
&\mathcal{L}_f = \|\text{sg}(\boldsymbol{z}_h) - \boldsymbol{z}_q\|_2^2 + \beta\|\boldsymbol{z}_h - \text{sg}(\boldsymbol{z}_q)\|_2^2,\\
&\mathcal{L}_I = \mathcal{L}_1 + \mathcal{L}_{per} + \mathcal{L}_f + (\mathcal{L}_{KL}^S + \mathcal{L}_{KL}^T) + \lambda_{adv}\cdot\mathcal{L}_{adv}
\end{aligned} \tag{7}$$

where $\Phi$ denotes the feature extractor of VGG19 [39], $\text{sg}(\cdot)$ is the stop-gradient operator. We hope to jointly optimize the encoder $E_h$ and codebooks $\boldsymbol{\mathcal{C}}_S$ and $\boldsymbol{\mathcal{C}}_T$ via the code-level loss $\mathcal{L}_f$, and $\beta$ controls the relative update rates of the encoder and codebooks. We also adopt straight-through gradient estimator [11, 43] to solve the non-differentiable problem of quantization operation.

### 3.3. Lookup Transformer Learning (Stage II)

The heavy degradations in real-world videos not only destroy the structure and details of the original face but also disrupt the continuity of temporal motions. In order to support LQ video input $\boldsymbol{x}_l$, we introduce two additional Transformer-based code lookup modules designed to predict the code sequence indices of the compressed spatial latents $\boldsymbol{z}_{l,S}$ and temporal latents $\boldsymbol{z}_{l,T}$ within the codebooks, respectively. The lookup transformer consists of multiple transformer layers based on self-attention. In stage II, we fix the HQ decoder $D_h$ and spatial-temporal codebooks trained in stage I, and train the LQ encoder $E_l$ and code lookup modules via HQ-LQ video pairs.

Given the compressed latents $\boldsymbol{z}_l \in \mathbb{R}^{t\times h\times w\times D}$ output by $E_l$, the spatial latents $\boldsymbol{z}_{l,S}$ and temporal latents $\boldsymbol{z}_{l,T}$ can

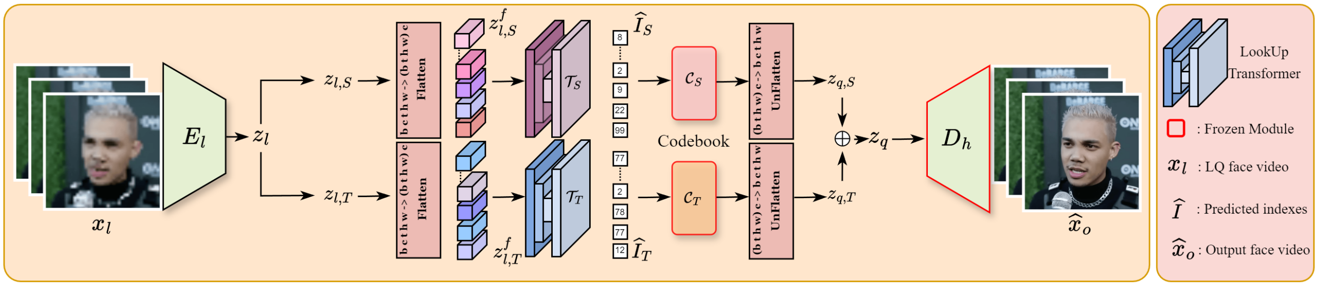



Figure 2. **Network architecture of Stage II.** Stage II uses HQ-LQ face video pairs to train LQ encoder $E_l$ and LookUp Transformers ($\mathcal{T}_S$ and $\mathcal{T}_T$). The weights of $D_h$ are pre-trained in Stage I and fixed in Stage II.

be calculated analogously to Eqn. ( 3). The spatial latents $\boldsymbol{z}_{l,S} \in \mathbb{R}^{t\times h\times w\times D}$ are initially flattened into sequence-shaped latents $\boldsymbol{z}^f_{l,S} \in \mathbb{R}^{(t\cdot h\cdot w)\times D}$ and then processed by a cascaded Transformer module $\mathcal{T}_S$. Learnable position embeddings are inserted into the lookup module to enhance the model's awareness of spatial relative positions. The lookup module outputs classification probabilities for each sequence item, which are used to generate code sequences $\widehat{I}_S \in \{0,\ldots,N_S-1\}^{t\cdot h\cdot w}$. Similarly, code sequences $\widehat{I}_T \in \{0,\ldots,N_T-1\}^{t\cdot h\cdot w}$ for temporal latents $\boldsymbol{z}_{l,T}$ can be obtained using the code lookup module $\mathcal{T}_T$. The proposed 3D-VQGAN achieves a substantial compression ratio (*i.e.*, $\frac{H}{h}=\frac{W}{w}=8$ and $\frac{T}{t}=2$), significantly alleviating the computational burden on the Transformers.

**Training objectives.** Following the work [58], we consider only cross-entropy prediction loss and code-level loss $\mathcal{L}'_f$:

$$\mathcal{L}^S_{CE}=\sum_k^{thw}-I^k_S\log\left(\widehat{I}^k_S\right),\quad \mathcal{L}^T_{CE}=\sum_l^{thw}-I^l_T\log\left(\widehat{I}^l_T\right),$$
$$\mathcal{L}'_f=\|\boldsymbol{z}_l-\mathrm{sg}(\boldsymbol{z}_q)\|_2^2,\;\; \mathcal{L}_{II}=\mathcal{L}'_f+\lambda_{CE}\cdot(\mathcal{L}^S_{CE}+\mathcal{L}^T_{CE}) \tag{8}$$

where $\widehat{I}_S$ and $\widehat{I}_T$ are the code sequences predicted by lookup modules, while $I_S$ and $I_T$ are the ground-truth sequences generated by HQ encoder and frozen spatial-temporal codebooks. We use code-level loss $\mathcal{L}'_f$ to minimize the distance between compressed latents $\boldsymbol{z}_l$ and quantized latents $\boldsymbol{z}_q$, which simplifies the training of Stage II.

# 4. Experiments

## 4.1. Implementation Details

**Datasets.** We train the model on the **VFHQ** [48] and our own collected live broadcast datasets. The VFHQ dataset contains 16,000 video sequences in total. We filtered VFHQ based on face proportion, face orientation, and the presence of text, and finally used approximately 3,200 videos, the data processing details can be found in the Data Manipulation subsection. The collected live broadcast dataset contains about 9,000 videos of people talking. We evaluate our method and other baselines on **VFHQ-Test** dataset. VFHQ-Test is composed of 50 high-quality videos, we choose the first 24 frames (1s) of each sequence as our test set.

For the blind face video restoration task, our data degradation method is consistent with other competitors [12, 41, 49]. Specifically, the degradation model is as follows [14]:

$$y=\left\{\left[(x\circledast k_\sigma)_{\downarrow_r}+n_\delta\right]_{\mathrm{FFMPEG}_{crf}}\right\}_{\uparrow_r} \tag{9}$$

where $y$ is a degraded frame, $x$ is a HQ frame. $x$ is first convolved with a gaussian kernel $k_\sigma$ to add blurring, then downsample to scale $r$. Then, gaussian noise $n_\delta$ and video coding degradation FFMPEG with a constant rate factor are applied into the frames. Finally, the output is resized to $512^2$. All frames in one video have a consistent $\sigma$, $r$, and $\delta$.

For the de-flickering task, we consider brightness flickering and pixel flickering. Each frame in the input video has a 30% probability of being degraded. The brightness degradation is implemented through `cv2.convertScaleAbs` to simulate the brightness flickering problem in old movies and damaged videos. The pixel flickering is implemented by re-rendering the frame through the Stable Diffusion model to simulate the inconsistency of details and textures between frames.

**Data Manipulation.** We preprocess the training videos in three aspects, including **face proportion** ($\mathbb{A}$), **face orientation** ($\mathbb{B}$), and **the presence of text** ($\mathbb{C}$). Previous methods typically employ a face detection module to separate faces for separate processing, resulting in inefficient inference. We improve the quality of the training data by cropping videos based on face proportions, thus simplifying the data processing chains and eliminating the need for face background separation. Specifically, we select the region with the largest face proportion in the video as a reference to crop that video to maximize the face region. Secondly, the work in [58] proposes that frontal faces are more conducive to the model learning high-quality portrait features. We filtered out side-face-dominated videos by assessing the ratio of the horizontal distances from the eyes to the nose

| Task | Method | Quality and Fidelity | | | Pose Consistency | | | Temporal Consistency | | Efficiency |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | LPIPS↓ | AKD↓ | Face-Cons↑ | IDS↑ | FVD↓ | Flow-Score↓ | Runtime(s)↓ |
| BFIR | VQFR [14] | 25.94 | 0.7852 | 0.2467 | 5.978 | 0.9947 | 0.6659 | 388.2 | 1.451 | 15.60 |
| | GFPGAN [45] | 27.15 | 0.8207 | 0.2279 | 4.134 | 0.9950 | 0.9206 | 246.9 | 1.316 | 14.44 |
| | CodeFormer [58] | 26.77 | 0.8102 | 0.2373 | 4.543 | 0.9947 | 0.8596 | 261.8 | 2.672 | 28.18 |
| VSR | BasicVSR++ [4] | 27.22 | 0.8218 | 0.2742 | 5.129 | **0.9965** | 0.9234 | 392.7 | 1.286 | 72.21 |
| | Real-BasicVSR [5] | 27.45 | 0.7929 | 0.2968 | 4.780 | 0.9936 | 0.8785 | 305.7 | 1.404 | 12.20 |
| BFVR | PGTFormer [49] | **28.68** | <u>0.8426</u> | **0.1752** | **3.519** | 0.9942 | <u>0.9296</u> | <u>107.6</u> | <u>1.154</u> | <u>7.085</u> |
| | KEEP [12] | 27.04 | 0.8223 | 0.2370 | 3.979 | 0.9953 | 0.8783 | 264.9 | 1.302 | 19.01 |
| | Ours | <u>27.47</u> | **0.8641** | <u>0.1829</u> | <u>3.858</u> | <u>0.9954</u> | **0.9312** | **105.1** | **1.150** | **2.995** |

Table 1. **Quantitative comparison on the VFHQ-test dataset for blind face video restoration.** The inference efficiency of our method is much higher than that of other baselines. We bold the **best** and underline the <u>second-best</u> performances.

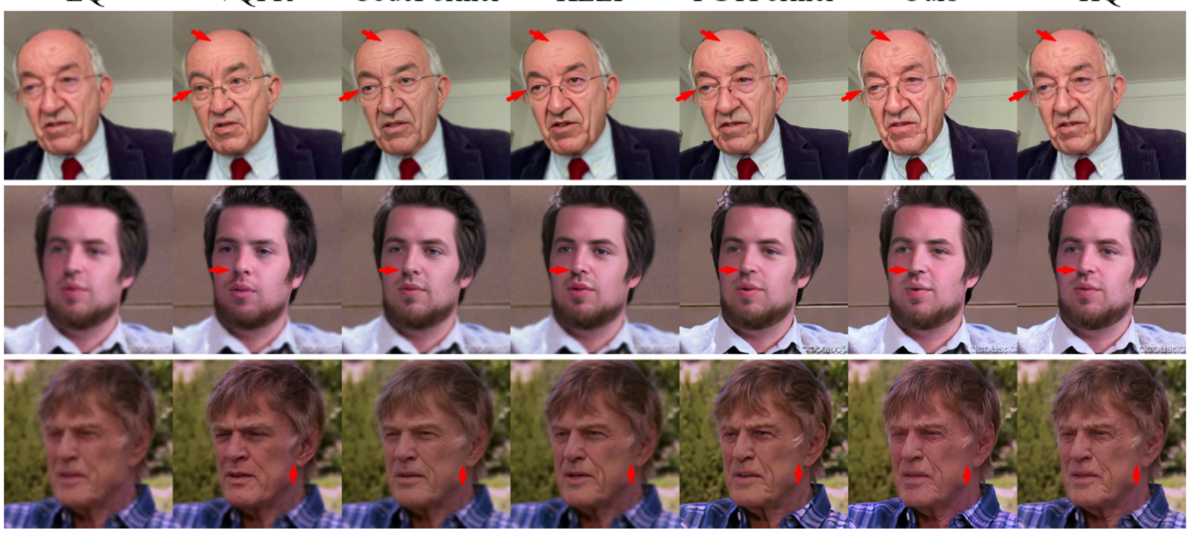



Figure 3. **Qualitative comparison on the VFHQ-Test for BFVR task.** Our method has better fidelity and fewer hallucination cases compared to other methods, such as wrinkles (1st row), eye orientation (1st row), nose shape (2nd row), and hairstyle (3rd row).

to determine face orientation. Thirdly, we found that some text obscures the face in the training data, so we removed videos containing text using PaddleOCR.

**Evaluation Metrics.** For the evaluation of the BFVR task, we adopt **PSNR**, **SSIM**, and **LPIPS** [54] to evaluate the quality and fidelity of restoration videos. Following video evaluation benchmark EvalCrafter [30], we utilize two metrics to gauge temporal consistency, including **FVD** and Average Flow (**Flow-Score**). Flow-Score assesses the average dense flows extracted by RAFT [42] from the generated videos. Following the work in [12], we apply Face Consistency (**Face-Cons**), Identity Preservation Scores (**IDS**), and Average Keypoint Distance (**AKD**) to evaluate identity preservation and pose consistency. AKD is calculated by the average distance between detected face landmarks in the generated and ground-truth video frames. Face-Cons calculates the average cosine similarity between the embedding of the first frame $emb(x_1)$ and subsequent frame embeddings $\{emb(x_t)\}_{t=2}^{T}$, focusing on the consistency of human identity in generated videos. Additionally, we measure the efficiency of different methods by considering the time required to generate a 24-frame (1-second) video. For the evaluation of de-flickering tasks, we adopt FVD and Flow-Score to assess temporal consistency.

**Model Configuration.** In stage I, the training videos are resized to $256^2$ and only the first 24 frames are used for training. The latent dimension $D$ is set to 256. Both the size of spatial and temporal codebooks are 1024. The training iteration is 250,000. The used pre-trained feature network $\mathcal{F}$ is `dinov2_vits14`. In stage II, the training videos are resized to $512^2$. The training iteration is 50,000. The degradation degree $\sigma \in [2, 5]$, $r \in [2, 4]$, $\delta \in [0, 5]$. The training

| Method | GT | FVD↓ | Flow-Score↓ | Runtime(s)↓ |
|---|---|---|---|---|
| Input | - | 494.3 | 1.572 | - |
| FastBlend [9] | ✓ | 34.58 | 1.574 | **18.44** |
| DVP [25] | ✓ | **14.53** | **1.492** | 410.2 |
| NeuralAtlas [26] | ✗ | 111.0 | 1.468 | 851.3 |
| Ours | ✗ | **100.7** | **1.100** | **2.934** |

Table 2. **Quantitative comparison for brightness de-flickering.**

| Method | GT | FVD↓ | Flow-Score↓ | Runtime(s)↓ |
|---|---|---|---|---|
| Input | - | 65.32 | 1.702 | - |
| FastBlend [9] | ✓ | 45.80 | 1.566 | **18.60** |
| DVP [25] | ✓ | **15.09** | **1.505** | 413.9 |
| NeuralAtlas [26] | ✗ | 110.5 | 1.471 | 854.1 |
| Ours | ✗ | **86.88** | **1.063** | **2.967** |

Table 3. **Quantitative comparison pixel de-flickering.**

is conducted on 4 NVIDIA A100 GPUs, with batch size 4. Both the two stages use Adam optimizer [22]. We apply `stable-diffusion-x4-upscaler` to generate pixel (AI-based) flickering.

### 4.2. Quantitative and Qualitative Comparisons

**Blind face video restoration.** Our comparison for BFVR task primarily focuses on existing BFIR, VSR, and BFVR methods, including VQFR [14], GFPGAN [45], CodeFormer [58], BasicVSR++ [4], Real-BasicVSR [5], PGTFormer [49], and KEEP [12]. We present the quantitative comparison on the VFHQ-Test in Table 1. Our method outperforms both BFIR and VSR methods across almost all metrics. Compared to image-based methods, our generated results exhibit higher quality, fidelity, and improved temporal consistency. Among the BFVR methods, our results are comparable to those of state-of-the-art methods, with enhanced inference efficiency. This demonstrates that the proposed 3D-VQGAN and spatial-temporal codebooks enable effective video compression and quantization, offering promising restoration quality coupled with reduced inference costs. Figure 3 provides some representative restoration results from different methods. Our method successfully captures portrait features in low-quality inputs and achieves higher fidelity. In contrast, due to a limited temporal perceptive field, other methods inaccurately alter key features, such as wrinkles, eye orientation (1st row), nose shape (2nd row), and hairstyle (3rd row).

**De-flickering.** We report the quantitative results for brightness de-flickering and pixel de-flickering in Table 2 and Table 3, respectively. For both tasks, we consider three baseline methods, including FastBlend [9], DVP [25], and NeuralAtlas [26]. It should be noted that FastBlend and DVP achieve long-term temporal consistency by integrating or training on a temporally consistent unprocessed video. However, such ground-truth videos are typically unavailable in real-world scenarios. In contrast, NeuralAtlas and our approach require only a single flickering video without additional guidance. Furthermore, the training-at-test-time strategy used by DVP and NeuralAtlas significantly reduces their inference efficiency. Our method achieves strong temporal consistency while maintaining high inference efficiency, as indicated by the quantitative results.

Figure 4 presents the temporal profiles of different methods for brightness de-flickering and pixel de-flickering. We visualize temporal variation by stacking pixels of a particular column over time. In Figure 4(a), the LQ input video contains some frames with brightness perturbations, and the generated video of NeuralAtlas is affected by the input resulting in subtle color shifts. In contrast, our method, utilizing the 3D-VQGAN backbone and spatial-temporal codebooks, produces de-flickered videos with consistent colors. In Figure 4(b), our method better maintains the consistency of the static text and texture in AI-generated videos, effectively addressing common flickering problems.

### 4.3. Ablation Study

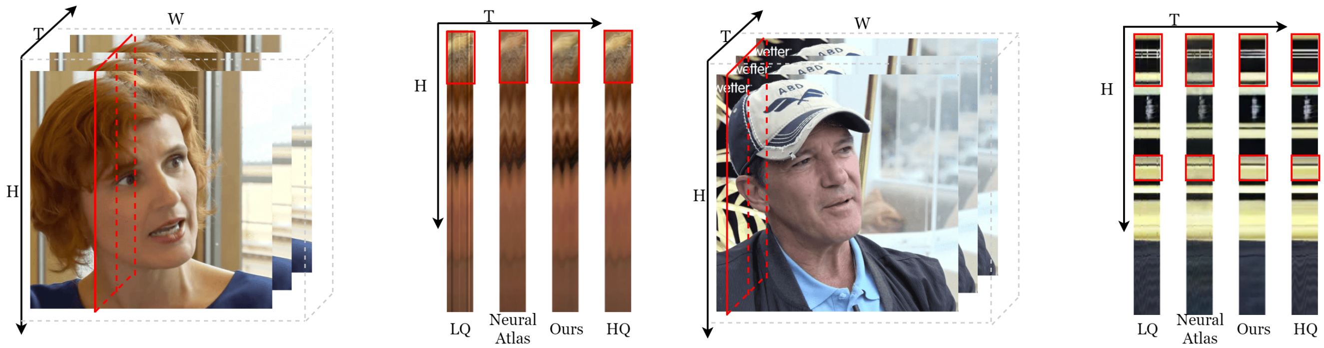
We conducted an ablation study to evaluate several design choices, which we discuss in detail below.

**Discriminator feature networks.** (Table 4(a)). The work in [36] indicates that applying pre-trained feature networks, such as CLIP and DINO, as the discriminator can boost the performance of diffusion distillation. We adopt a pre-trained network to serve as a more powerful discriminator to reduce artifacts and stabilize training. The results show that DINO's objective is more suitable than CLIP's and the scale of ViT has little effect on the performance.

**Fusion operation.** (Table 4(b)). For the fusion operation $\oplus$ between quantized spatial latent $z_{q,S}$ and temporal latent $z_{q,T}$, we have tried various fusion methods, such as direct addition, convolution projection, and 3DFFT fusion [31]. Convolution projection means concatenating two latents along the channel dimension and using 3D convolution to reduce it to normal size. 3DFFT represents the fusion of latents by processing two variables separately by 3DFFT with a pre-defined pass threshold. Among these methods, direct addition yields the best performance.

**Loss terms.** (Table 4(c)). The results in the table show the contribution of the discriminator and proposed marginal prior regularization in the training of Stage I.

**Codebook regularization.** (Table 4(d)). We report the reconstruction performance in Stage I under different regularization. PDR and MPR contribute to the learning of codebooks and backbones compared with the vanilla version. CVQ failed when jointly optimizing multiple codebooks, leading to model collapse. Figure 5 display the comparison of codebook utilization when reconstructing the videos in




(a) **Brightness de-flickering.** (b) **Pixel de-flickering.**

Figure 4. Comparison of temporal profile for brightness and pixel de-flickering. We select a column to observe the changes across time.

| Arch | Objective | SSIM↑ | FVD↓ |
|---|---|---|---|
| ViT-S | DINOv2 | **0.9054** | 49.11 |
| ViT-B | DINOv2 | 0.9050 | **49.08** |
| ViT-B | CLIP | 0.8935 | 66.86 |

(a) **Discriminator feature networks.**

| Fusion $\oplus$ | SSIM↑ | FVD↓ |
|---|---|---|
| Add | **0.9054** | **49.11** |
| Conv | 0.8799 | 97.67 |
| 3DFFT [31] | 0.8741 | 118.7 |

(b) **Fusion operation** $\oplus$ between quantized spatial and temporal latents.

| $\mathcal{L}_{base}$ | $\mathcal{L}_{adv}$ | $\mathcal{L}_{KL}$ | SSIM↑ | FVD↓ |
|---|---|---|---|---|
| ✓ | ✗ | ✗ | 0.8887 | 66.82 |
| ✓ | ✓ | ✗ | 0.8914 | 65.90 |
| ✓ | ✓ | ✓ | **0.9054** | **49.11** |

(c) **Loss terms** for Stage I.

| Regularization | SSIM↑ | FVD↓ |
|---|---|---|
| None | 0.8914 | 65.90 |
| PDR [53] | 0.9015 | 56.46 |
| CVQ [56] | 0.4667 | 940.1 |
| MPR(Proposed) | **0.9054** | **49.11** |

(d) **Codebook regularization.**

| $N_S$ | $N_T$ | $D$ | SSIM↑ | FVD↓ |
|---|---|---|---|---|
| 512 | 512 | 256 | 0.8919 | 69.28 |
| 1024 | 1024 | 256 | **0.9054** | **49.11** |
| 2048 | 2048 | 256 | 0.8934 | 66.29 |
| 1024 | 1024 | 64 | 0.8871 | 71.31 |

(e) **Codebook size and dimension.**

| $\mathbb{A}$ | $\mathbb{B}$ | $\mathbb{C}$ | #Video | SSIM↑ | FVD↓ |
|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | 15967 | 0.8891 | 71.67 |
| ✓ | ✗ | ✗ | 15967 | 0.8918 | 67.62 |
| ✓ | ✓ | ✗ | 12216 | 0.8995 | 61.81 |
| ✓ | ✓ | ✓ | 11057 | **0.9054** | **49.11** |

(f) **Data manipulation.**

Table 4. **Ablation study.** We report SSIM and FVD results of different variants for reconstruction on the VFHQ-test dataset. The training length is 250,000 iterations. Default settings are marked as gray.

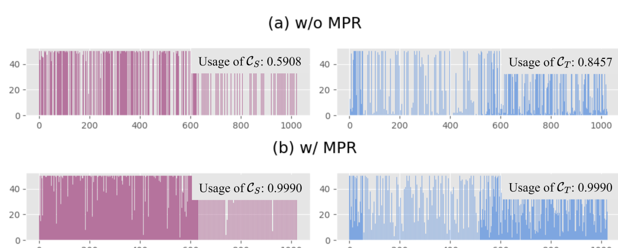



Figure 5. Comparison of spatial and temporal codebooks' utilization when applying different regualizations.

the VFHQ-Test dataset. Applying the proposed MPR can effectively alleviate the codebook collapse problem and improve the utilization of codebooks.

**Codebook size and dimension.** (Table 4(e)). Experiments show that a moderate codebook size and latent space dimension improve performance.

**Data manipulation.** (Table 4(f)). We preprocess training videos in three aspects, comprising face proportion ($\mathbb{A}$), face orientation ($\mathbb{B}$), and the presence of text ($\mathbb{C}$). From the results, we can learn that compared to the quantity, the quality of face videos is more helpful for the learning of 3D-VQGAN backbone and spatial-temporal codebooks.

## 5. Conclusion

In this work, we propose a two-stage video face enhancement framework, which can jointly solve BFVR and de-flickering tasks. This framework develops upon a 3D-VQGAN backbone and a more powerful discriminator for efficient video compression and spatial-temporal codebooks for effective video quantization. We also design a marginal prior regularization to mitigate the codebook collapse problem. Experiments demonstrate that our method achieves promising restoration and de-flickering performance with enhanced inference efficiency.

## 6. Acknowledgements

This work was supported by the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai, P.R. China (AI202409).

## References


[1] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. 2

[2] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015. 2

[3] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 1, 2

[4] 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, 6, 7

[5] 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, 7

[6] 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. 1, 2

[7] 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

[8] Julie Delon and Agnes Desolneux. Stabilization of flicker-like effects in image sequences through local contrast correction. *SIAM Journal on Imaging Sciences*, 3(4):703–734, 2010. 2

[9] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, Jun Huang, and Mingyi Jin. Fastblend: a powerful model-free toolkit making video stylization easier. *arXiv preprint arXiv:2311.09265*, 2023. 2, 7

[10] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, and Jun Huang. Diffutoon: High-resolution editable toon shading via diffusion models. *arXiv preprint arXiv:2401.16224*, 2024. 2

[11] 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. 2, 4

[12] Ruicheng Feng, Chongyi Li, and Chen Change Loy. Kalman-inspired feature propagation for video face super-resolution. *arXiv preprint arXiv:2408.05205*, 2024. 1, 2, 5, 6, 7

[13] Anuj Fulari, Satish Mulleti, and Ajit Rajwade. Unsupervised model-based learning for simultaneous video deflickering and deblotching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4117–4125, 2024. 1

[14] 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, 5, 6, 7

[15] Xiaobin Hu, Wenqi Ren, John LaMaster, Xiaochun Cao, Xiaoming Li, Zechao Li, Bjoern Menze, and Wei Liu. Face super-resolution guided by 3d facial priors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 763–780. Springer, 2020. 1, 2

[16] Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *International Conference on Machine Learning*, pages 14096–14113. PMLR, 2023. 2

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4

[18] Ali Kanj, Hugues Talbot, and Raoul Rodriguez Luparello. Flicker removal and superpixel-based motion tracking for high speed videos. In *2017 IEEE international conference on image processing (ICIP)*, pages 245–249. IEEE, 2017. 2

[19] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE transactions on computational imaging*, 2(2):109–122, 2016. 1

[20] 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. 1, 2

[21] 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. 1, 2

[22] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[23] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 3

[24] 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. 2

[25] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020. 7

[26] Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. Blind video deflickering by neural filtering with a flawed atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10439–10448, 2023. 1, 2, 7

[27] 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

[28] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 1

[29] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 2

[30] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 6

[31] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 7, 8

[32] OpenAI. Sora: Creating video from text., 2024. `https://openai.com/sora/`. 1

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[34] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[36] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 4, 7

[37] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*, 2024. 4

[38] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 1

[39] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015. 4

[40] Yuhta Takida, Takashi Shibuya, WeiHsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547*, 2022. 2

[41] Jingfan Tan, Hyunhee Park, Ying Zhang, Tao Wang, Kaihao Zhang, Xiangyu Kong, Pengwen Dai, Zikun Liu, and Wenhan Luo. Blind face video restoration with temporal consistent generative prior and degradation-aware prompt. In *ACM Multimedia 2024*, 2024. 1, 2, 5

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 4

[44] 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

[45] 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. 6, 7

[46] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1

[47] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17512–17521, 2022. 1, 2

[48] 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. 5

[49] Kepeng Xu, Li Xu, Gang He, Wenxin Yu, and Yunsong Li. Beyond alignment: Blind video face restoration via parsing-guided temporal-coherent transformer. *arXiv preprint arXiv:2404.13640*, 2024. 1, 2, 5, 6, 7

[50] 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

[51] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE transactions on image processing*, 15(5):1120–1129, 2006. 1

[52] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Au-*

*dio, Speech, and Language Processing*, 30:495–507, 2021. 2

[53] Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18467–18476, 2023. 2, 8

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 6

[55] Yuzhi Zhao, Lai-Man Po, Kangcheng Liu, Xuehui Wang, Wing-Yin Yu, Pengfei Xian, Yujia Zhang, and Mengyang Liu. Svcnet: Scribble-based video colorization network with temporal aggregation. *IEEE Transactions on Image Processing*, 2023. 1

[56] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22798–22807, 2023. 2, 8

[57] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425, 2022. 2

[58] 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, 5, 6, 7