Blind Face Video Restoration with Temporal Consistent Generative Prior and Degradation-Aware Prompt

Jingfan Tan*
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
tjfky2001@gmail.com

Hyunhee Park
the Department of Camera Innovation Group, Samsung Electronics
Suwon, Republic of Korea
inextg.park@samsung.com

Ying Zhang
Samsung Research China - Beijing
Beijing, China
ying09.zhang@samsung.com

Tao Wang
Nanjing University
Nanjing, China
taowangzj@gmail.com

Kaihao Zhang
Harbin Institute of Technology (Shenzhen)
Shenzhen, China
super.khzhang@gmail.com

Xiangyu Kong
Samsung Research China - Beijing
Beijing, China
xiangyu.kong@samsung.com

Pengwen Dai
Shenzhen Campus of Sun Yat-sen University
Shenzhen, China
daipw@mail.sysu.edu.cn

Zikun Liu
Samsung Research China - Beijing
Beijing, China
zikun.liu@samsung.com

Wenhan Luo†
Hong Kong University of Science and Technology
Hong Kong, China
whluo.china@gmail.com

Abstract

Within the domain of blind face restoration (BFR), approaches lacking facial priors frequently result in excessively smoothed visual outputs. Exiting BFR methods predominantly utilize generative facial priors to achieve realistic and authentic details. However, these methods, primarily designed for images, encounter challenges in maintaining temporal consistency when applied to face video restoration. To tackle this issue, we introduce StableBFVR, an innovative Blind Face Video Restoration method based on Stable Diffusion that incorporates temporal information into the generative prior. This is achieved through the introduction of temporal layers in the diffusion process. These temporal layers consider both long-term and short-term information aggregation. Moreover, to improve generalizability, BFR methods employ complex, large-scale degradation during training, but it often sacrifices accuracy. Addressing this, StableBFVR features a novel mixed-degradation-aware prompt module, capable of encoding specific degradation information to dynamically steer the restoration process. Comprehensive experiments demonstrate that our proposed StableBFVR outperforms state-of-the-art methods.

*This work is done during the internship of Tan in Samsung.
†Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680917

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Blind face restoration, diffusion model, facial generative prior, video restoration

ACM Reference Format:
Jingfan Tan, Hyunhee Park, Ying Zhang, Tao Wang, Kaihao Zhang, Xiangyu Kong, Pengwen Dai, Zikun Liu, and Wenhan Luo. 2024. Blind Face Video Restoration with Temporal Consistent Generative Prior and Degradation-Aware Prompt. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680917

1 Introduction

In real-world scenarios, both face images and videos may suffer from unknown and varied types of degradation, such as downsampling, noise, blur, and compression. Blind Face Restoration (BFR) is a challenging task that aims at restoring low-quality faces suffering from unknown degradation. Existing BFR methods usually use facial priors such as reference prior, geometry prior, and generative prior in the network structure. Among various priors, the generative prior from pre-trained generators, due to its ability to bring more realistic texture and details, has been popularly leveraged by recent BFR methods to restore faces. For example, GFP-GAN [55] incorporates the pre-trained StyleGAN [22] as a decoder into an encoder-decoder architecture. DiffBIR [35] first utilizes a restoration model for preliminary restoration, then introduces Stable Diffusion [45] as generative prior to further refine facial details. Although these existing BFR methods work well in the blind face image restoration (BFIR) problem, they do not fully consider blind face videos. When these BFIR methods are applied to face videos, they usually restore the face cropped from each frame of the video

Figure 1: Four consecutive frames restored by different methods. Blind face image restoration method CodeFormer results in inconsistent visual effects between the face and other regions and temporal inconsistency. Video restoration method BasicVSR++ results in an over-smoothing effect. Our method strikes a good balance between generating texture and temporal consistency.

and paste it back into the original frame, and the background is restored using other restoration models such as RealESRGAN [56]. As shown in Fig. 1, this strategy typically leads to two problems: 1) the visual effects of the face and the background are inconsistent; 2) the generated texture is unstable, and the face attributes (*e.g.,* hairstyle, eyes, mouth) may change between frames.

On the other hand, existing video restoration methods achieve temporal consistency by fusing information across frames. For example, BasicVSR++ [7] uses second-order grid propagation and flow-guided deformable alignment to effectively exploit information from the entire input video. VRT [32] adopts transformer architecture for attaining long-range receptive fields. As shown in Fig 1, due to the lack of facial prior, when these methods are used for blind face video restoration, they usually produce over-smooth results that are very inconsistent with human perception. Thus, existing video restoration methods are not applicable for restoring blind face videos. Besides, to the best of our knowledge, there is still no specialized method for restoring blind face videos.

To tackle these challenges, we present a Stable Blind Face Video Restoration (StableBFVR). It uses the pre-trained latent diffusion model (LDM) Stable Diffusion as facial prior. At the same time, to maintain temporal consistency and use multiple frame information to improve the restoration performance, we introduce temporal layers to Stable Diffusion, which comprehensively consider both long-term and short-term information in the video. Specifically, we present Shift-ResBlock using the proposed forward temporal shift block and backward temporal shift block alternatively to achieve bidirectional aggregation. The temporal shift blocks first shift input features in the temporal dimension, followed by fusion using convolution blocks. By using Shift-ResBlock repeatedly, the aggregation of long-term information is achieved. For short-term information aggregation, we introduce a Nearby-Frame Attention (NFA). By seeking complementary sharp information existing in neighboring frames, NFA can refine restoration details.

BFR methods usually utilize a wide range of degradation when synthesizing training data. This enhances the generalization ability of the restoration model but also results in a decrease in accuracy. To further improve the performance, we propose a Degradation-Aware Prompt Module (DAPM). DAPM first extracts degradation-aware features from the input frames to predict prompt weights about different types of degradation. Then DAPM utilizes these weights to adjust the corresponding prompt corresponding to different types of degeneration and fuses these prompts to obtain degradation-aware prompts which encode discriminative information about various types of degradation. By interacting with degradation-aware prompts, the StableBFVR can make adaptive responses to various unknown degradations to effectively restore input faces.

Our main contributions are as follows: (1) We propose StableBFVR which uses generative facial prior to address the blind face video restoration task for the first time. To maintain temporal consistency and improve the restoration performance, we convert pre-trained Stable Diffusion into video restoration models by inserting temporal layers. (2) We present a Degradation-Aware Prompt Module (DAPM) to generate prompts that contain degradation-specific information for dynamically guiding the restoration network. In this way, we can improve the restoration performance and enable the restoration network to adapt to diverse, unknown degradations. (3) Extensive experimental studies demonstrate StableBFVR achieves SOTA performance on both the public synthetic dataset and real-world low-quality face video dataset we collected from the Internet.

2 Related Work

2.1 Video Restoration

Video restoration aims to restore high-quality videos from low-quality ones. Most existing methods can be divided into two categories according to the way they propagate information.

The first [54, 67] usually uses sliding window to aggregate information from adjacent frames to restore the middle single frame. During the alignment stage, they often align all frames in the sliding window towards the middle frame. Earlier methods [5, 62] estimate the optical flow between low-quality neighbouring frames and then perform spatial warping for alignment. Recent approaches employ implicit alignment. For example, some methods [49, 54] align different frames at the feature level with the deformable convolution. Some methods [21, 73] leverage dynamic filters to achieve motion compensation. Some methods [32, 33] mainly use transformer to fuse useful features from adjacent frames. However, multi-frame

input leads to higher computational complexity and it is hard to use larger window sizes to aggregate more distant frames.

The second [6, 7, 33] typically utilizes the recurrent-based method to propagate information from one frame to the next frame, which is accumulated to restore the subsequent frames. These methods usually focus on designing efficient propagation methods for utilizing longer distance frames. For example, RSDN [19] proposes a novel unidirectional propagation with a hidden state adaptation module to enhance robustness to appearance change and error accumulation. Some methods [6, 7] employ bidirectional propagation to better exploit temporal features.

2.2 Generative Prior for Blind Face Image Restoration

Early blind face image restoration (BFIR) methods usually employ geometric [4, 8, 65, 68] and reference priors [29–31, 48] to improve the restoration performance. Reference priors use the facial component dictionary obtained from additional high-quality face images to guide the face restoration process. Geometric priors use the unique geometric shape and spatial distribution information of faces like facial landmarks, facial heatmaps, and facial parsing maps to help restore high-quality faces. However, geometric prior and reference prior are unable to provide rich facial details.

For better visual effects, the generative facial priors from pre-trained generators have been explored for BFIR recently. Some works [55, 63, 75] incorporate the pre-trained StyleGAN [22] as a decoder into an encoder-decoder architecture. Other works [13, 57, 72] first train VQGAN [11] on high-quality faces with a reconstruction objective, then fine-tune the decoder to adapt to BFIR. Recently, some works [9, 35] leverage the pre-trained Stable Diffusion (SD) [45] as generative prior which provides more prior knowledge compared with existing GAN prior and achieves realistic face restoration. Inspired by these works, our approach, for the first time, applies generative priors to blind face video restoration. Moreover, we develop effective techniques to maintain the temporal consistency among continuous frames when restoring facial details.

2.3 Diffusion Model

Recently, due to the more stable generation ability than GAN, the diffusion model has been popular in image restoration. Some methods [14, 46, 58] train a diffusion model conditioned on low-quality images and perform restoration through a stochastic denoising process. Some methods [35, 51] fine-tune directly on the pre-trained stable diffusion model to achieve impressive performance. Although the diffusion model has shown promise in image restoration, it is still under-explored in video restoration.

With notable advancements in image generation diffusion model, a number of methods [1, 12, 15, 44] use off-the-shelf image diffusion models with additional temporal layers to achieve video generation. Some methods [16, 17, 47] extend image diffusion models by training them on extensive video pairs. Some methods [24, 36, 53] employ temporal attention mechanisms to generate videos. Some methods [10, 41] introduce optical flow warping in diffusion process. Inspired by video generation works [15, 44, 61] that employ off-the-shelf image diffusion models, our video restoration method exploits pre-trained stable diffusion as a generative prior and proposes a novel temporal strategy, resulting in temporal consistency.

2.4 Prompt Learning

With the extensive application of prompt learning in the field of NLP [3, 37] and high-level vision tasks [18, 20], prompt learning has recently also been widely used in image restoration to better utilize the degradation context, such as the all-in-one restoration tasks [38, 43, 52]. Although prompt learning performs well in all-in-one restoration tasks, the degraded image contains only a single type of degradation. Our approach for the first time explores the application of prompt learning in dealing with mixed degradation tasks like the case of blind face restoration.

3 Methodology

BFR methods can use pre-trained generation models to restore high-quality images with clear facial details. However, if we directly use generative prior for video restoration, the inherent stochastic nature of the generation model leads to temporal inconsistencies in the restored video. Especially for face videos, in addition to the flickering artifacts, it also causes the face attributes (*e.g.,* hairstyle, eyes, mouth) in the restored video to be inconsistent.

By training on a large amount of data, Stable Diffusion has powerful prior knowledge about faces. We aim to harness the knowledge from Stable Diffusion for blind face video restoration. As shown in Fig. 2, we introduce temporal layers in the Stable Diffusion to preserve temporal consistency. First, we propose Shift-Resblock which implicitly captures global information for long-term aggregation. Second, we further improve restoration performance and temporal consistency by introducing Nearby-Frame Attention to aggregate short-term information. Moreover, to enable adaptive responses to complex and large-range blind degradation, we propose a degradation-aware prompt module to encode degradation-specific information as prompts to guide the restoration.

3.1 Preliminary: Latent Diffusion Models

Pre-trained Stable Diffusion Stable Diffusion employs the LDM framework. It utilizes the encoder of Variational Autoencoders (VAE) to map the image x into the latent z to perform the diffusion and denoising processes, then reconstruct it with the decoder of VAE. In the diffusion process, the diffused latent z_t can be directly generated by applying Gaussian noise with variance $\beta_t \in (0, 1)$ to the latent z at any time step t. This process can be described as:

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\epsilon \sim \mathcal{N}(0, I)$ is a random Gaussian noise. In the reverse process, the U-Net denoiser ϵ_θ parameterized by θ is trained to predict the noise ϵ. The optimization objective of the latent diffusion model can be defined as follows:

$$\mathcal{L} = \mathbb{E}_{z_t, t, c, \epsilon}[||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2], \tag{2}$$

where t is a randomly selected time-step, c is an optional condition (*e.g.,* text, images, and representations), and ϵ is sampled from the standard Gaussian distribution.

In this work, we start with the pre-trained Stable Diffusion and create a new video diffusion model for blind face video restoration. By adopting temporal strategies within the LDM framework, our

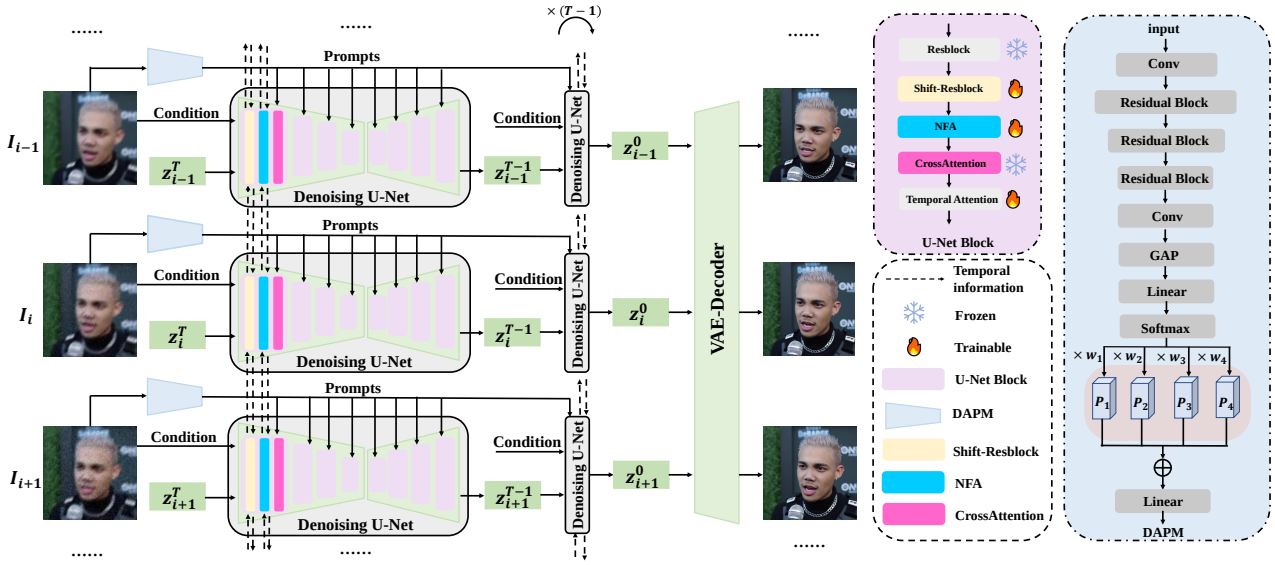

Figure 2: The architecture of the proposed StableBFVR. We turn Stable Diffusion into a video restoration method by adding temporal layers Shift-Resblock and Nearby-Frame Attention (NFA) into the U-Net block. To further improve performance, we adopt a Degradation-Aware Prompt Module (DAPM) that dynamically guides the diffusion process.

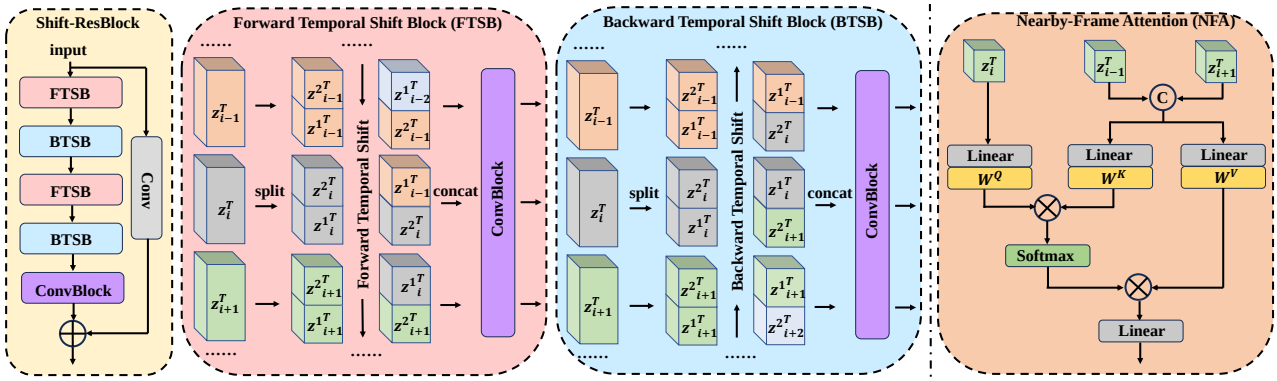

Figure 3: The structure of the proposed Shift-Resblock and Nearby-Frame Attention (NFA).

method can achieve temporal consistency while leveraging the prior knowledge from Stable Diffusion.

3.2 Temporal Layers in StableBFVR

To apply the pre-trained Stable Diffusion to video-related tasks, we propagate long-term and short-term temporal information among different input frames to maintain temporal consistency. Long-term information helps preserve face attribute consistency among long-range frames. Short-term information relieves flickering artifacts of adjacent frames. Considering that the degree of degradation between different frames of the video is different, propagating temporal information also helps improve the restoration performance.

Long-term Information Aggregation. Previous works [10, 41] demonstrate the benefits of optical flow-guided long-term propagation in video diffusion models. Considering that blind degraded video usually contains severe blur or noise, this kind of degradation will affect the accuracy of the optical flow estimation network [28, 66, 74], subsequently leading to poor performance of the optical flow-guided restoration network. This suggests that optical flow is not suitable for our task. Prior works [34, 40] have demonstrated that temporal shift operation can blend information from other frames with the current frame along the temporal dimension and establish the temporal correspondences implicitly. Thus, we introduce the Shift-Resblock in the basic U-Net blocks to effectively establish temporal correspondences and conduct long-term fusion.

As shown in Fig. 3, supposing the i-th frame input feature of Shift-Resblock at time step T is $Z_i^T \in \mathbb{R}^{C\times H\times W}$, Shift-Resblock consists of Forward Temporal Shift Block (FTSB) which fuses the feature of (Z_{i-1}, Z_i) and a Backward Temporal Shift Block (BTSB) which fuses the feature of (Z_i, Z_{i+1}). By stacking FTSB and BTSB alternatively, Shift-Resblock can achieve bi-directional aggregation.

Although a single Shift-Resblock can only fuse adjacent frame information, we can achieve long-term aggregation by using Shift-Resblock in our framework repeatedly. In the temporal shift, each ${Z_i^T}$ is split into two parts ${Z^1}_i^T \in \mathbb{R}^{C^1 \times H \times W}$ and ${Z^2}_i^T \in \mathbb{R}^{C^2 \times H \times W}$ along the channel dimension. In the forward temporal shift, we shift the feature ${Z^1}_{i-1}^T$ from the $(i-1)$-th frame to the i-th frame, then feature ${Z^1}_{i-1}^T$ and feature ${Z^2}_i^T$ are merged as feature ${Z^f}_i^T$ of the i-th frame. The output of the forward temporal shift for the i-th frame can be defined as:

$$ {Z^f}_i^T = Concat({Z^1}_{i-1}^T, {Z^2}_i^T), 0 < i \leq F, \tag{3} $$

where F is the number of input frames. In particular, in the forward temporal shift, the first frame remains unchanged. In the backward temporal shift, we shift the feature ${Z^2}_{i+1}^T$ from the $(i+1)$-th frame to the i-th frame, then feature ${Z^2}_{i+1}^T$ and feature ${Z^1}_i^T$ are merged as feature ${Z^b}_i^T$ of the i-th frame. The output of the backward temporal shift for the i-th frame can be defined as

$$ {Z^b}_i^T = Concat({Z^1}_i^T, {Z^2}_{i+1}^T), 0 \leq i < F. \tag{4} $$

Analogously, in the backward temporal shift, the last frame remains unchanged. After the temporal shift, we utilize a simple convolution block independently on each frame to capture and aggregate both the spatial and temporal information.

Moreover, following [1, 59], we also introduce temporal attention to the U-Net blocks. The temporal attention performs self-attention along the temporal dimension for temporal modeling.

Short-term Information Aggregation. The original U-Net block has a spatial self-attention. When the input is multiple frames, it acts only on each frame alone. Considering that the short-term adjacent frames are usually highly similar to the current frame, they can provide sufficient information for the restoration of the current frame. Thus, to further enhance the restoration performance, we present a Nearby Frame Attention (NFA) mechanism that extends the spatial self-attention to the temporal domain. By seeking complementary sharp information in neighboring frames, NFA can capture spatio-temporal consistency. The structure of NFA is shown in Fig. 3. Specifically, given the i-th frame feature maps $Z_i \in \mathbb{R}^{C \times H \times W}$ as input, our NFA takes current frame input features Z_i as query $Q = W_q(Z_i)$, while the key $K = W_k(Concat(Z_{i-1}, Z_{i+1}))$ and value $V = W_v(Concat(Z_{i-1}, Z_{i+1}))$ are generated from the concatenation of the former frame and the latter frame, where W_q, W_k, W_v are projection matrices shared across space and time. Finally, we adopt the self-attention mechanism to conduct short-term information aggregation. The output is a weighted sum of the value, weighted by the similarity between the query and key features. Note that, in the training process, we only update the parameters of the query process, and the other parts of the parameters are frozen.

3.3 Degradation-Aware Prompt Module

Previous work [70] has proven that when using complex large-scale degradation to train blind face restoration methods, it will enhance the generalization ability, but at the cost of decreasing accuracy. To address this problem, we propose a degradation-aware prompt module (DAPM) to generate prompts that can dynamically adjust the prediction of the degree of degradation of the input frames. It can help the restoration network make adaptive responses to various unknown degradations.

The structure of DAPM is shown in Fig 2. Considering that blind degradation usually consists of four different degradations: blur, noise, compress, and downsample, we establish a degradation set $P = \{P_1, P_2, P_3, P_4 \mid P_N \in \mathbb{R}^{L \times C}\}$ to encapsulate information for different degradation. Serving as learnable parameters, P can interact with the dynamical weights that are predicted from the input degraded frame. Thus it can function as prompts aware of degradations.

Specifically, to predict dynamical weights, DAPM first extracts features $F_i^0 \in \mathbb{R}^{C \times H \times W}$ from a given degraded input frame I_i by applying a convolution operation. Then the feature is sent to a three-level encoder, with each level of the encoder employs several residual blocks. The feature will be transformed into the compact feature $F_i^1 \in \mathbb{R}^{4C \times \frac{H}{4} \times \frac{W}{4}}$ which is rich in degradation-aware information. Then we apply global average pooling (GAP) across the spatial dimension to generate a feature vector $V_i \in \mathbb{R}^C$. Next, we use a linear layer and softmax operation to obtain prompt weights w_1, w_2, w_3, w_4 about the four kinds of degradation. Finally, considering that different degradation is not independent, degradation will also affect each other. After we use these weights to make adjustments in the degradation set P, we use a linear layer to fuse them. The process of generating degradation-aware prompts is

$$ w_i = Softmax(Linear(GAP(F_i^1))), \tag{5} $$

$$ \hat{P} = Linear(\sum_{i=1}^{4} w_i P_i). \tag{6} $$

The generated degradation-aware prompt will be fed into CrossAttention in the denoising U-Net block to dynamically guide the restoration network.

4 Experiment

4.1 Datasets and Implementation

Training Datasets. We train our method on 2,000 clips randomly chosen from VFHQ [60]. VFHQ is a high-quality video face dataset, which contains over 16,000 high-fidelity clips of human faces. We resize all the frames to 512×512 during training. We train our method on synthetic data that approximate to the real low-quality images. Similar to the common practice in blind face restoration [55, 72], the degradation model is as follows:

$$ y = [(x \circledast k_\sigma) \downarrow_r + n_\delta]_{\text{FFMPEG}_{crf}}, \tag{7} $$

where y is the synthetic low-quality frame, x is the high-quality frame, k_σ is Gaussian blur kernel, r represents the down-sample factor, and n_δ is white Gaussian noise. We incorporate the $FFMPEG_{crf}$ compress into the degradation model, where crf is the constant rate factor that decides how many bits will be used for each frame. Compared with the $JPEG$ compress used in BFIR, $FFMPEG_{crf}$ can implicitly consider the inter-dependencies between frames, providing temporal and spatial degradations. For each training pair, we randomly sample σ, r, δ, and crf from $[0.1, 10]$, $[1, 4]$, $[0, 15]$, $[18, 25]$, respectively.

Testing Datasets. We evaluate our method on synthetic dataset **VFHQ-Test** and real-world dataset **WebVideo-Test**. They both

Table 1: Quantitative comparison on VFHQ-Test and WebVideo-Test for blind face video restoration. Red and Blue indicate the best and the second-best performance.

	VFHQ-Test							WebVideo-Test		
	LPIPS↓	NIQE↓	MUSIQ↑	CLIP-IQA↑	WE ↓	PSNR↑	SSIM↑	NIQE↓	MUSIQ↑	CLIP-IQA↑
Input	0.4591	10.237	16.35	0.276	13.66	25.84	0.7564	8.361	37.08	0.286
GFPGAN [55]	0.4139	5.587	65.86	0.601	16.10	26.30	0.7624	4.897	72.01	0.615
RestoreFormer [57]	0.4162	5.615	59.86	0.583	16.73	26.17	0.7504	4.842	68.08	0.628
CodeFormer [72]	0.4116	5.603	64.04	0.604	16.24	26.32	0.7588	4.991	70.55	0.640
DiffBIR [35]	0.4354	7.293	53.15	0.512	15.88	26.38	0.7603	6.012	63.78	0.581
BaiscVSR++ [7]	0.3406	9.149	50.15	0.294	6.28	28.05	0.8213	7.803	54.88	0.299
DSTNet [42]	0.3493	9.641	43.66	0.297	6.27	**28.30**	**0.8319**	8.340	53.66	0.350
RVRT [33]	0.3710	9.419	37.81	0.246	**6.21**	27.79	0.8104	8.316	45.57	0.272
Ours	**0.3119**	**5.262**	**75.33**	**0.759**	13.45	26.58	0.7689	**4.512**	**74.20**	**0.693**
GT	0	4.778	72.83	0.645	7.10	∞	1	-	-	-

have no overlap with our training dataset. **VFHQ-Test** is composed of 50 high-quality clips. We choose the first 100 frames of each clip, a total of 5,000 frames as our test set. To synthesize testing pairs, we apply the same degradation model as the training phase. To better evaluate the generalization of blind face video restoration methods in the real world, we propose a real-world test set **WebVideo-Test**. The videos in our WebVideo-Test dataset are collected from video websites. It consists of 10 video clips, each containing 100 frames of diverse and complicated degradation.

Evaluation Metrics. For evaluation of the VFHQ-Test with ground truth, we adopt pixel-wise metrics PSNR and SSIM and perceptual metric LPIPS [71]. We also employ non-reference perceptual metrics NIQE [39], CLIP-IQA [50], and MUSIQ [23]. For the real-world dataset WebVideo-Test without ground truth, we adopt only the three non-reference metrics mentioned above. Compared with BFIR, one major aspect of the BFVR problem is the temporal consistency of the restored videos. In this work, we adopt the average warping error (WE) [26] of the restored videos to quantitatively measure the temporal consistency. It can be calculated as:

$$WE = \frac{1}{N-1}\sum_{i=2}^{N} ||\hat{I}_i - \mathbf{W}(I_{i-1}, S_{i-1->i})||_1, \tag{8}$$

where $\hat{I}_i$ is the predicted frame, $\mathbf{W}$ denotes the spatial warping operation, and $S_{i-1->i}$ is the estimated optical flow from ground-truth video. We use 10^{-3} quantity level when showing this metric.

Implementation Details. We utilize Stable Diffusion V2.1 to initialize the weight of StableBFVR. Then we fix the weight of StableBFVR except for the proposed components and the condition. Regarding the condition, we first employ frozen pre-trained BasicVSR++ for preliminary restoration, then adopt trainable ControlNet [69], initialized with the weight from BFIR method DiffBIR [35], encode the input frame as a condition and inject it into the denoising U-Net. The training is conducted on 4 NVIDIA A100 GPUs, with batch size 4 and the number of input frames 8. The learning rate is set to 1×10^{-4} using the Adam [25] optimizer and we train it for $100K$ iterations. During inference, we divide the low-quality video into multiple sequences. For each sequence, the number of input frames is set to 32 and we run the sampling for 50 steps.

4.2 Results

We compare our StableBFVR with several state-of-the-art methods, including four BFIR models, GFPGAN [55], Restoreformer [57], CodeFormer [72], DiffBIR [35], and three video restoration models, BasicVSR++ [7], DSTNet [42], RVRT [33]. For BFIR models, we adopt their officially released models in the experiments. Following original implementations in video restoration, the BFIR model only restores the face detected in the video frame, while the background is enhanced by RealESRGAN [56]. For video restoration models, to ensure a fair comparison, these methods are re-trained under the same training settings.

Quantitative Comparison. For the synthetic VFHQ-Test, the quantitative results are shown in Tab. 1. The results indicate that our method achieves state-of-the-art performance on all perceptual metrics. Specifically, our StableBFVR achieves the best performance regarding LPIPS, indicating that the perceptual quality of restored face videos is closest to ground truth. Moreover, StableBFVR also obtains the best results of NIQE, MUSIQ, and CLIP-IQA, showing that the outputs better align with human visual and perceptive systems. Note that, like other BFIR methods that use generative prior, our model is also not strong at PSNR and SSIM. Because PSNR and SSIM do not correlate well with the human visual and perceptive systems [2, 27]. In general, over-smoothing images will derive higher PSNR and SSIM values. The methods based on generative priors produce more high-frequency texture details, resulting in lower PSNR and SSIM.

To assess the generalization ability, we extend the evaluation of our model to the real-world dataset WebVideo-Test. The quantitative results are presented in Tab. 1. StableBFVR exhibits superior performance across all three metrics NIQE, MUSIQ, and CLIP-IQA, showing its remarkable generalization capability. Furthermore, compared with video restoration methods, BFIR methods also show satisfactory performance, suggesting the importance of generative prior in the scenery of unknown degradations in the real world.

Qualitative Comparison. For the synthetic VFHQ-Test, the qualitative results are illustrated in Fig. 4. Compared with video restoration methods, thanks to the powerful generative facial prior, our method recovers faithful details in the eyes, mouth, beard *etc.* On the contrary, face videos restored by video restoration methods

Figure 4: Visual comparison results of different methods on the VFHQ-Test. Our StableBFVR produces more faithful details. Zoom in for best view.

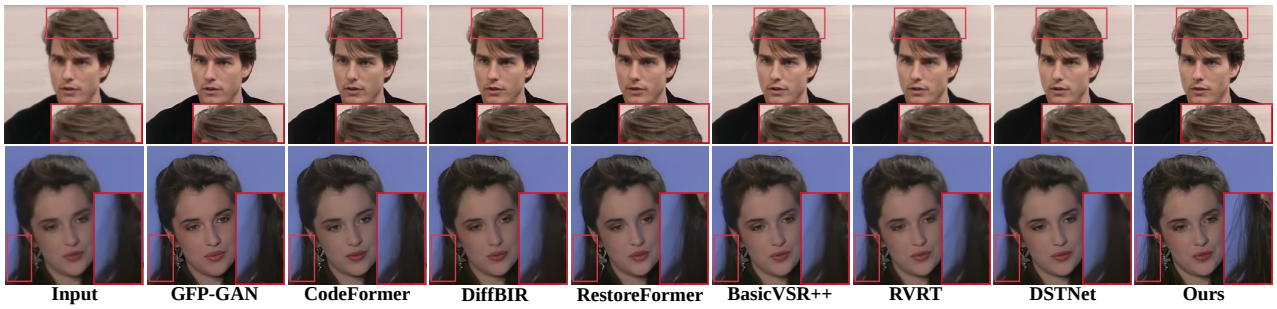

Figure 5: Visual comparison results of different methods on the real-world dataset WebVideo-Test. Our StableBFVR produces more faithful details. Zoom in for best view.

are over-smooth and lose facial texture details. BFIR methods only restore the face detected in the video frame, with the background restored by RealESRGAN. Consequently, face videos restored by BFIR exhibit inconsistent visual effects between the face region and background region, even showing obvious boundaries (3rd row, Fig. 4). In contrast, our method treats the input as a whole in restoration and performs well in all regions. In addition, compared with BFIR methods, our method can aggregate information from other frames to improve performance. As a result, in scenarios where BFIR methods exhibit poor performance, our method can still generate superior facial details (4th row, Fig. 4).

We show qualitative results of WebVideo-Test in Fig. 5. Our method produces realistic facial textures in the case of complicated real-world degradation. As shown in the last column of Fig. 5, previous methods fail to restore the hair textures on the image boundary, while ours is successful. Compared with video restoration methods, our method produces significantly more texture detail.

Temporal Consistency. The quantitative assessment of temporal consistency is presented in Tab. 1. It is worth mentioning that, the metric WE may not be able to faithfully reflect the human perception of the temporal consistency [64]. For example, over-smoothing sequences usually have much higher WE scores despite unpleasant perceptual quality. As shown in Tab. 1, the scores of the video restoration methods are even higher than the ground truth. Given that our StableBFVR tends to generate more details and textures, which adversely impact the WE value, it exhibits a less favorable performance in this regard. Nonetheless, StableBFVR still outperforms others driven by the generative facial priors.

To thoroughly verify our method, we visualize the consecutive frames generated by different methods in Fig. 6. It is observed that, although sequences restored by BFIR methods exhibit realistic texture, there are noticeable differences between the textures of continuous frames. Conversely, sequences restored by video restoration methods demonstrate commendable temporal consistency but

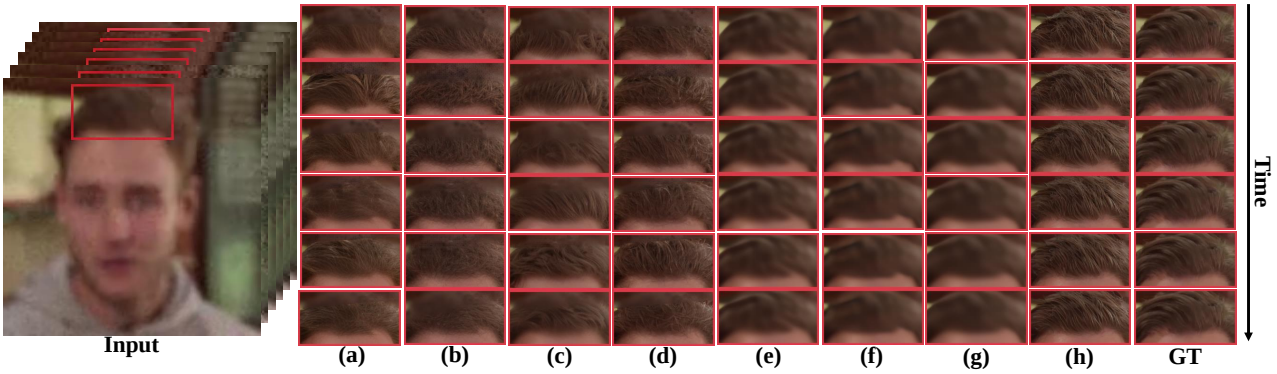

Figure 6: Visual comparisons of the temporal consistency for restored videos. (a) GFP-GAN, (b) CodeFormer, (c) DiffBIR, (d) RestoreFormer, (e) BaiscVSR++, (f) RVRT, (g) DSTNet, (h) Ours. Zoom in for best view.

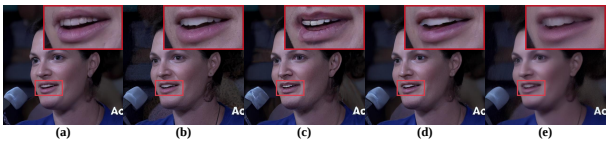

Figure 7: Visual comparison results of ablation study. (a) Full Method, (b) w/o Shift-Resblock, (c) replace NFA with SA, (d) replace DAPM with CLIP, (e) w/o DAPM. Zoom in for best view.

Table 2: Ablation studies of our StableBFVR on VFHQ-Test.

Configuration	LPIPS↓	MUSIQ↑	WE↓
replace NFA with SA	0.3312	73.18	14.89
w/o Shift-Resblock	0.3541	72.33	15.44
w/o DAPM	0.3263	74.69	13.48
replace DAPM with CLIP	0.3289	74.86	13.52
inference frames 8	0.3162	74.82	14.91
inference frames 16	0.3144	75.19	13.96
inference frames 24	0.3120	75.23	13.65
Full method	0.3119	75.33	13.45

tend to be excessively smooth, lacking textures. StableBFVR strikes a favorable balance, reconstructing more textures, while simultaneously preserving temporal consistency.

4.3 Ablation Studies

Effectiveness of Temporal Layers. As depicted in Tab. 2, we explore the significance of the temporal layers. Specifically, we first remove Shift-ResBlock, resulting in a noticeable decline in both temporal consistency and perceptual metrics. The impact of this configuration change is shown in Fig. 7, where not only are the details of the teeth inadequately restored, but artifacts also emerge in the background. It implies that Shift-Resblock can improve perceptual quality and consistency through the aggregation of long-term information. Subsequently, we replace NFA with the original Self-Attention from Stable Diffusion. This replacement leads to a performance decrease in terms of both temporal consistency and perceptual metrics. Fig. 7 illustrates that, while the texture of the teeth can be generated, the quality is notably diminished. This observation emphasizes that the rich contextual information from neighboring frames extracted by NFA plays a crucial role in refining details.

Effectiveness of DAPM. We then investigate the significance of the DAPM. We first directly remove DAPM. This alteration leads to a drop in perceptual performance. Subsequently, we replace DAPM with CLIP. Following other Stable Diffusion-based restoration methods [35, 51], we set the input of CLIP as an empty string. This substitution similarly leads to a reduction in perceptual performance. Fig. 7 visually demonstrates that both configurations fail to restore the texture of teeth, implying the instrumental role of DAPM in enhancing restoration.

The Number of Inference Frames. As detailed in Tab. 2, we observe that the input number of frames during inference can affect the performance. The more the input frames are, the better the performance is. Especially for temporal consistency, there is a significant enhancement when the number of input frames increases from 8 to 16. These results also illustrate that propagating information about distant frames helps improve restoration performance and temporal consistency.

5 Conclusion

In this work, we tackle the BFVR problem for the first time. We propose StableBFVR leveraging the strong generative prior from the pre-trained generative model Stable Diffusion to restore face videos with realistic details. To ensure content consistency among frames and use multi-frame information for improved restoration, we develop Shift-Resblock and Nearby-Frame Attention to aggregate both long-term and short-term information. Additionally, we propose a Degradation-Aware Prompt Module to dynamically guide the restoration process and further enhance performance. Extensive experiments show that our StableBFVR achieves superior performance than video restoration methods and blind face image restoration methods.

Acknowledgments

This work is funded in part by the National Natural Science Foundation of China (Grant No. 62372480, 62302532), in part by Theme-based Research Scheme (T45-205/21-N) from Hong Kong RGC, in part by CCF-Tencent Rhino-Bird Open Research Fund (No. CCF-Tencent RAGR20230118), in part by Generative AI Research and Development Centre from InnoHK.

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018. The 2018 PIRM challenge on perceptual image super-resolution. In *European Conference on Computer Vision Workshop*.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[4] Adrian Bulat and Georgios Tzimiropoulos. 2018. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[5] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[8] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. 2021. Progressive semantic-aware style transformation for blind face restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[9] Xiaoxu Chen, Jingfan Tan, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaochun Cao. 2024. Towards Real-World Blind Face Restoration with Generative Diffusion Prior. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[10] Zijun Deng, Xiangteng He, Yuxin Peng, Xiongwei Zhu, and Lele Cheng. 2023. MV-Diffusion: Motion-aware Video Diffusion Model. In *Proceedings of the ACM International Conference on Multimedia*.

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023).

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

[14] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. 2023. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2024. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*.

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. In *Advances in Neural Information Processing Systems*.

[18] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. 2023. Diversity-Aware Meta Visual Prompting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[19] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. 2020. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*.

[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*.

[21] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*.

[24] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning blind video temporal consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[28] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. 2023. A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[29] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. 2020. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*.

[30] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. [n. d.]. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. 2022. Learning Dual Memory Dictionaries for Blind Face Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[32] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. 2022. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288* (2022).

[33] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. 2022. Recurrent video restoration transformer with guided deformable attention. In *Advances in Neural Information Processing Systems*.

[34] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

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

[36] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761* (2023).

[37] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602* (2021).

[38] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. 2023. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018* (2023).

[39] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* (2012).

[40] Andres Munoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. 2021. Temporal shift GAN for large scale video generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[41] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. 2023. Conditional Image-to-Video Generation with Latent Flow Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. 2023. Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[43] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. 2023. PromptIR: Prompting for All-in-One Image Restoration. In *Advances in Neural Information Processing Systems*.

[44] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing Supplemental Material. In *IEEE International Conference on Computer Vision*.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[46] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[47] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

[48] Jingfan Tan, Xiaoxu Chen, Tao Wang, Kaihao Zhang, Wenhan Luo, and Xiaocun Cao. 2023. Blind Face Restoration for Under-Display Camera via Dictionary Guided Transformer. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[49] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[50] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images.

[51] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015* (2023).

[52] Tao Wang, Kaihao Zhang, Ziqian Shao, Wenhan Luo, Bjorn Stenger, Tong Lu, Tae-Kyun Kim, Wei Liu, and Hongdong Li. 2024. Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions. *International Journal of Computer Vision* (2024).

[53] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599* (2023).

[54] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

[55] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[56] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *IEEE International Conference on Computer Vision*.

[57] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. 2022. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[58] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. 2022. Deblurring via stochastic refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[59] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[60] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

[61] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. 2023. SimDA: Simple Diffusion Adapter for Efficient Video Generation. *arXiv preprint arXiv:2308.09710* (2023).

[62] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision* (2019).

[63] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. 2021. Gan prior embedded network for blind face restoration in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[64] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. 2023. Motion-Guided Latent Diffusion for Temporally Consistent Real-world Video Super-resolution. *arXiv preprint arXiv:2312.00853* (2023).

[65] Yanjiang Yu, Puyang Zhang, Kaihao Zhang, Wenhan Luo, and Changsheng Li. 2023. Multi-Prior Learning via Neural Architecture Search for Blind Face Restoration. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[66] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. 2020. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[67] Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, and Wei Liu. 2023. Enhanced Spatio-Temporal Interaction Learning for Video Deraining: A Faster and Better Framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[68] Kaihao Zhang, Wenhan Luo, Yanjiang Yu, Wenqi Ren, Fang Zhao, Changsheng Li, Lin Ma, Wei Liu, and Hongdong Li. 2022. Beyond monocular deraining: Parallel stereo deraining network via semantic prior. *International Journal of Computer Vision* (2022).

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[70] Ruofan Zhang, Jinjin Gu, Haoyu Chen, Chao Dong, Yulun Zhang, and Wenming Yang. 2023. Crafting training degradation distribution for the accuracy-generalization trade-off in real-world super-resolution. In *International Conference on Machine Learning*.

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[72] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. In *Advances in Neural Information Processing Systems*.

[73] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. 2019. Spatio-temporal filter adaptive network for video deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[74] Chao Zhu, Hang Dong, Jinshan Pan, Boyang Liang, Yuhao Huang, Lean Fu, and Fei Wang. 2022. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring.

[75] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. 2022. Blind face restoration via integrating face shape and generative priors. In *IEEE Conference on Computer Vision and Pattern Recognition*.