This WACV paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

# FLAIR: A Conditional Diffusion Framework with Applications to Face Video Restoration

Zihao Zou<sup>1,\*</sup>, Jiaming Liu<sup>2,\*</sup>, Shirin Shoushtari<sup>2</sup>, Yubo Wang<sup>2</sup>, and Ulugbek S. Kamilov<sup>2</sup> <sup>1</sup>University of North Carolina, Chapel Hill, NC, USA <sup>2</sup>Washington University in St. Louis, MO, USA

\*These authors contributed equally.

### Abstract

Face video restoration (FVR) is a challenging but important problem where one seeks to recover a perceptually realistic face videos from a low-quality input. While diffusion probabilistic models (DPMs) have been shown to achieve remarkable performance for face image restoration, they often fail to preserve temporally coherent, high-quality videos, compromising the fidelity of reconstructed faces. We present a new conditional diffusion framework called FLAIR for FVR. FLAIR ensures improved temporal alignments across frames in a computationally efficient fashion by converting a traditional image DPM into a video DPM. The proposed conversion uses a recurrent video refinement layer and a temporal self-attention at different scales. FLAIR also uses a conditional iterative refinement process to balance the perceptual and distortion quality during inference. This process consists of two key components: a data-consistency module that analytically ensures that the generated video precisely matches its degraded observation and a coarse-to-fine image enhancement module specifically for facial regions. Our extensive experiments show superiority of FLAIR over the current state-of-the-art (SOTA) for video super-resolution, deblurring, JPEG restoration, and space-time frame interpolation on two high-quality face video datasets.

# 1. Introduction

As a subcategory of the general image and video restoration [47,48,69,87], face restoration is an active research area in computer vision [34,38,40,44,58,71]. Image and video restoration is usually *ill-posed* due to the information loss induced by degradation (*e.g.*, resolution loss, blur, encoding artifacts, and noise), with multiple plausible high-quality (HQ) objects leading to the same low-quality (LQ) observation. Face restoration has recently been greatly improved by using generative priors [28,74,84] and pre-trained face dictionary priors [25,42,77,93]. While SOTA methods—such as Codeformer [93], VQFR [25], and RestoreFormer [77] can restore high-quality results with fine details, they usually hallucinate HQ faces that diverge from the original subjects in the presence of severe degradation [91], leading to large distortion, as can be seen in Fig 1 (*Top*). On the other hand, our method generates more realistic facial appearances and better preserves identity, which may benefit downstream tasks such as face recognition [50] and surveillance and security [3, 26].

Diffusion probabilistic models (DPMs) [31,67] have attracted significant attention as an alternative to traditional generative models due to their excellent performance in image and video generation [5, 21, 27, 57, 60, 89]. DPMs have been applied to a range of imaging problems, showing impressive results for face restoration. These methods generally fall into two categories: model-based unsupervised methods [16, 36, 39, 66, 75, 79] and conditional training methods [56, 59, 61, 80]. Despite recent activity in the area, there are very few DPM-based frameworks for video restoration, especially in the context of face video restoration (FVR). The key challenges are the significant computational cost of training on video data and the lack of largescale, publicly available HQ face video datasets. Given the stochasticity of the generative process in DPMs, another challenge is the effective use of nearby, similar but misaligned frames for reconstructing temporally aligned HQ reference frames [51,73]. Unlike DDNM [75], one of the latest conditional image DPMs, our method consistently restores facial features across frames, enhancing temporal consistency, as illustrated in the zoomed-in region in Fig.1 (Middle).

**Proposed Work:** We present *Diffusion Probabilistic Face Video Restoration (FLAIR)*, a conditional generative model for FVR that generates multiple distinct, high-quality enhanced face videos from degraded sequential data. FLAIR is designed as a "repeated-refinement" conditional DPM. Instead of directly training on high-resolution videos, we first pre-train our conditional DPMs on images, leveraging large-scale HQ image datasets efficiently. The image DPMs



Figure 1. Qualitative evaluation of the proposed FLAIR method. *Top:* FLAIR can restore high-quality facial details and preserve the data fidelity across frames, while both CodeFormer [93] and RestoreFormer++ [78] hallucinate faces that diverge from the original subject. *Middle:* FLAIR produces better temporal consistency than existing conditional diffusion method DDNM [75]. *Bottom:* FLAIR delivers superior perceptual quality results than video restoration method VRT [46] on *real-world* web video.

take degraded estimations as auxiliary inputs for conditional restoration similar to [52, 61, 80]. Given a pre-trained image DPM backbone based on UNet [21], we modify it into a video restoration model by introducing a temporal dimension into the feature space and training these temporal layers on video sequences. Specifically, we propose a flow-guided video enhancement layer with a multi-scale recurrent module at high-resolution scales and several temporal self-attention blocks for low-resolution features in a sliding-window fashion. FLAIR captures long-range temporal dependencies, using information from multiple neighboring frames for the restoration of each frame during inference.

To better balance perceptual quality and data fidelity [6], we propose a two-stage refinement process at each reverse diffusion step. The first stage uses an interpretable dataconsistency (DC) module to ensure that the generated coarse, clean intermediate results precisely match their LQ counterparts, despite various mixed real-world degradations (e.g., resolution loss, blur, JPEG). In the second stage, the DC outputs are processed by an enhancement module for highquality facial details (see Fig. 2). This design ensures compatibility with various restoration methods, enabling FLAIR to produce both perceptually realistic and data-consistent results. Our method introduces novel technical ideas to the field of conditional DPMs by integrating the DC module and facial priors.

Our main contributions can be summarized as follows: (1) We propose FLAIR as the first conditional diffusion

framework for the recovery of long-term consistent, highquality face videos from their LQ observations. Our key insight is to convert conditionally pre-trained image DPMs into video restoration models by inserting temporal layers that learn to align images from severely degraded video *clips* in a temporally consistent manner (Fig. 3). (2) Together with a data-consistency module and an enhancement module, we employ FLAIR in a two-stage conditional refinement process at each iteration of the reverse diffusion to further improve the perception and fidelity simultaneously. (3) We show through extensive experiments that FLAIR outperforms SOTA methods for composite noisy degradation on two high-quality face video datasets both quantitatively and qualitatively, showing great potential for practical applications. (4) Our code is available at https://github.com/wustl-cig/FLAIR.

# 2. Related Work

**Face Restoration.** Traditional approaches for face restoration are based on the incorporation of prior knowledge and degradation models [10, 26, 68]. The quality of restored faces has been progressively improving after adoption of convolutional neural networks (CNNs) [32, 70, 86, 88]. Recent work has investigated various deep priors for face image restoration, including geometric and reference priors [8, 13, 14, 22, 43]. The restoration quality has been further improved by adapting pre-trained GANs, such as StyleGAN [35], as generative priors [1, 28, 74, 82, 84]. This



Figure 2. Overview of the proposed FLAIR framework. At the *t*-th sampling step, FLAIR uses the degraded video frame y as the guidance for the video DPM to denoise the latent video sequence  $x_t$ . The estimated  $x_{0t}$  is passed through the data-consistency module to ensure that its low-frequencies are consistent with y. The enhancement module then improves faces from  $\tilde{x}_{0t}$  for next sampling step.

line of works treats face restoration as a conditional image generation problem by projecting the LQ faces into a compact, low-dimension space of the pre-trained generator. Another line of works, e.g., VQFR [25], CodeFormer [93], ResotreFormer [77] and its variant [78], leverages pre-trained Vector-Quantization (VQ) codebooks [23] as dictionaries learned on facial regions, achieving SOTA results in blind face restoration. Diffusion Models. Apart from unconditional image generation, diffusion models have been extensively investigated in various imaging restoration tasks. One line of works has focused on designing conditional training methods in a supervised fashion [20, 56, 59, 61, 80]. Another line of work has focused on keeping the training of an unconditional image DPM intact, and only modify the inference procedure to enable sampling from a conditional distribution [15, 17, 18, 36, 53, 75, 79]. However, few DPM methods [12, 19, 92] have been explored for image video enhancement and restoration. Notably, none have directly addressed video restoration tasks focused on FVR.

Recent works convert pre-trained text-to-image DPMs into video DPMs for video editing and generation. Instead of training video DPMs from scratch [30], recent methods [5, 24, 33] add temporal attention layers to pre-trained image DPMs and fine-tune on video datasets. Some zeroshot methods [54, 81, 83] reshape self-attention to spatial-temporal self-attention without changing pre-trained weights during inference. However, we observe that using temporal attention alone results in sub-optimal outcomes for severely degraded FVR (see Table 3). Therefore, we focus on a novel conditional sampling algorithm for FVR and study video diffusion priors to learn temporal coherence from degraded face videos.

## 3. Preliminaries

**Diffusion Probabilistic Models.** The forward process of DPMs [31, 64] is a Markov Chain that gradually adds noise  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  to data  $\mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_0 \in \mathbb{R}^d$  according to the variance schedule  $\beta_t \in (0, 1)$  for all  $t = 1, \dots, T$ . Using the notation  $\alpha_t := 1 - \beta_t$  and  $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$ , sampling of

 $\boldsymbol{x}_t$  given  $\boldsymbol{x}_0$  can be expressed in a closed form

$$q(\boldsymbol{x}_t | \boldsymbol{x}_0) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t} \boldsymbol{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}).$$
(1)

The unconditional generative reverse process is a Gaussian transition that samples from  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  to  $x_0$  as

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) := \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \sigma_t^2 \mathbf{I}), \quad (2)$$

where  $\mu_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$  and  $\sigma_t$  depend on  $\boldsymbol{x}_t, \boldsymbol{x}_0$ , and  $\beta_t$ . DPMs train  $\epsilon_{\boldsymbol{\theta}}$  to learn the Gaussian transition  $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$  as an approximation of reverse diffusion  $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ . By training the residual denoiser network  $\epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$  to predict the total noise  $\epsilon_t$ , one can estimate  $\boldsymbol{x}_{0t}$  through

$$\boldsymbol{x}_{0t} = (\boldsymbol{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)) / \sqrt{\bar{\alpha}_t}, \qquad (3)$$

where  $x_{0t}$  denotes the first prediction of  $x_0$  given the noisy observation  $x_t$ . One can use the DDIM [65] strategy to sample from the generative process more efficiently

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \boldsymbol{x}_{0t} + \sqrt{1 - \bar{\alpha}_{t-1}} (\sqrt{1 - \eta_t} \boldsymbol{\epsilon}_t + \sqrt{\eta_t} \boldsymbol{\epsilon}), \quad (4)$$

where the magnitude of  $\eta_t = \eta \sigma_t^2 / (1 - \bar{\alpha}_{t-1})$  controlled by  $\eta \in \mathbb{R}_0^+$  determines how stochastic the forward process is (e.g., when  $\eta = 0$ , (4) becomes deterministic).

**Inverse Problems.** The FVR can be formulated as an inverse problem involving the recovery of a sequence  $\{X^n\}_{n=1}^N \in \mathbb{R}^{H \times W \times C}$  of video frames from a series of LQ measurements, where N, H, W, and C are the video length, height, width, and channel, respectively. For  $\boldsymbol{x} = [\boldsymbol{x}^1, \dots, \boldsymbol{x}^N] \in \mathbb{R}^{Nd}$  defined in a vector form, we have  $\boldsymbol{x}^n = \text{vec}(\boldsymbol{X}^n)^{\mathsf{T}} \in \mathbb{R}^d$ . The measurements can be represented as  $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{e}$ , where  $\mathcal{A} = [\mathcal{A}_1, \dots, \mathcal{A}_N] : \mathbb{R}^{Nm} \to \mathbb{R}^{Nd}$  ( $m \ll d$ ) is the measurement operator modeling the degradation process, and  $\boldsymbol{e} = [\boldsymbol{e}^1, \dots, \boldsymbol{e}^N] \in \mathbb{R}^{Nm}$  denotes the measurement noise. In this paper, we consider the scenario in which video quality suffers from spatial and temporal degradation of images due to factors such as out-of-focus, motion, limited sensor array intensity, and JPEG encoding [9,49,73].



Figure 3. Overview of our video DPM. (*a*) Layer-wise information of UNet model. The image DPM backbone  $\theta$  is fixed and only temporal layers  $\phi$  are fine-tuned. The recurrent feature enhancement (RFE) and temporal attention are selected for different resolutions. (*c*) Illustration of the RFE module, where each recurrent block takes the flow estimation from y for feature alignment.

#### 4. Proposed Approach: FLAIR

In this section, we describe the training and testing details of FLAIR tailored for FVR. Fig. 2 illustrates the overview of the proposed method. FLAIR is defined as a generative process over T steps conditioned on degraded video sequence  $\boldsymbol{y}, p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}|\boldsymbol{y})$ . The conditional generative process  $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{y})$  is learned to approximate the intractable conditional reverse process  $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y})$  for the inference, similar to unconditional DPMs.

#### 4.1. Diffusion Video Restoration Network

We leverage pre-trained DPMs for images to efficiently train the video diffusion model [5,63]. Our proposed method extends a DPM designed for image restoration, denoted as  $\epsilon_{\theta}$ , into a video diffusion restoration network represented as  $\epsilon_{\theta,\phi}$ . We introduce additional *temporal* neural network layers parameterized by  $\phi$  to  $\epsilon_{\theta}$  and fine-tune them to align individual frames for temporal consistency. We adopt UNet architecture in [21] for network  $\epsilon_{\theta}$ . The training of the conditional model requires concatenation of the input image  $x_t \in \mathbb{R}^d$  and condition  $c \in \mathbb{R}^d$  along the channel dimension. The condition c represents the up-scaled LQ measurements  $y \in \mathbb{R}^m$  to the same dimension as  $x_{0:T}$  (see supplements for more details).

**Temporal Layers Implementation.** Input feature maps in the pixel space are processed using the layers of image DPM denoted as *spatial* layers  $\{\mathcal{H}_{\theta}^i\}_{i=1}^L$ , while each interleaved *temporal* layer is denoted as  $\mathcal{H}_{\phi}^i$ . We use two distinct types of temporal layers depicted in Fig. 3: recurrent feature enhancement (RFE) and temporal attention. Each temporal module contains several 3D convolutional residual blocks. The spatial layers  $\mathcal{H}_{\theta}^i$  process the video as a collection of individual images within a batch by rearranging the temporal dimension into the batch axis, i.e.,  $\mathbb{R}^{B \times C \times N \times H \times W} \rightarrow \mathbb{R}^{(BN) \times C \times H \times W}$ , where *B* is the batch size. Subsequently, we reshape it back to the original video dimensions for each temporal layer  $\mathcal{H}^i_{\phi}$ .

At *i*-th resolution (*i.e.*, [512, 256]) of the RFE, a 3D convolutional residual block first extracts temporal features  $\tilde{f}_i$  from the spatial output  $f_i$  of  $\mathcal{H}_{\theta}^i$ . The extracted features are then propagated and aligned by several recurrent blocks. Each block consists of a flow-guided deformable feature alignment (DFA) inspired by [11] and a propagation annealing output layer. The DFA is designed for bidirectional propagation, aiming to enhance the robustness of the recurrent network against error accumulation and alteration in appearance. Given the fused feature outputs  $\{g_{i,j}^n\}_{n=1}^N$  in the *j*-th propagation branch, the annealing balances the smoothness of the background scenes in the feature space as

$$\boldsymbol{g}_{i,j}^{n} = \tilde{w} * (\boldsymbol{1} - \tilde{\boldsymbol{m}}_{t,i}^{n}) \odot \tilde{\boldsymbol{g}}_{i,j}^{n} + (\tilde{\boldsymbol{m}}_{t,i}^{n}) \odot \tilde{\boldsymbol{g}}_{i,j}^{n}, \quad (5)$$

where  $\tilde{w} \in [0,1]$  is a learnable parameter. The masks  $\{\tilde{m}_{t,i}^n\}_{n=1}^N$  are the downscale version (i-th scale) of facial region masks  $\{m_t^n\}_{n=1}^N$  estimated from  $x_{0t}$  at the *t*-th reverse diffusion step. The main motivation behind this design is to enhance robustness against appearance changes within the recurrent network. We have observed that this annealing can notably improve the temporal consistency of background scenes across frames while preserving the sharpness of facial region. We detail the modified DFA and facial mask  $m_t$  warping process in Sec. A.2 and A.6 in the supplemental material, respectively.

In addition, we integrate temporal attention following each  $\mathcal{H}^i_{\theta}$  to concurrently process N frames locally in parallel within low-resolution blocks (e.g., [32, 16, 8]). This can notably reduce memory complexity compared to directly integrating temporal attention into high-resolution scales. To enhance the expressiveness of modeling sequential representation, we include sinusoidal positional embeddings [31] into the attention blocks. Our video temporal backbone is then trained with the same noise schedule as in (1). We optimize the temporal layers' weights with the objective function

$$\mathcal{L}_{\boldsymbol{\phi}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{c}, \boldsymbol{\epsilon}, t \sim [1, T]} \left[ \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}, \boldsymbol{\phi}}(\boldsymbol{x}_t, \boldsymbol{c}, t)\|^2 \right], \quad (6)$$

while the spatial layers are frozen.

#### 4.2. Analytical Data Consistency Module

We initially consider a linear forward-model  $y^n = (h_n * x^n) \downarrow_s$  without noise added to individual frames  $\{y^n\}_{n=1}^N \in \mathbb{R}^m$ . In this expression,  $h_n * x^n$  denotes twodimensional convolution of clean image  $x^n$  and the blur kernel associated with the point-spread function (PSF) of the camera at frame n and  $\downarrow_s$  represents a s-fold down-sampler. For convenience, we denote the forward-model as y = Ax.

Mathad	Test			CelebV-	Text [85]			CelebV-HQ [94]						
weulou	Task	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓	
$A^+y$		20.81	0.721	0.542	1278.46	182.22	150.72	21.32	0.704	0.567	2145.50	260.01	183.21	
VQFR [25]		23.49	0.746	0.362	500.76	97.27	32.12	22.78	0.716	0.407	1103.10	180.93	59.50	
RestoreFormer++ [78]	bic	23.29	0.732	0.368	518.95	92.86	26.20	22.75	0.706	0.414	1154.06	175.27	50.04	
CodeFormer [93]	icu	23.58	0.738	0.374	507.03	101.20	32.90	22.89	0.711	0.419	1155.91	178.08	55.84	
DR2E [79]	В	24.38	0.755	0.314	456.12	81.42	21.81	23.73	0.726	0.349	984.00	148.80	37.55	
DDNM [75]	- <u>6</u>	25.78	0.789	0.337	617.05	82.07	41.80	24.85	0.753	0.368	1264.72	148.13	67.27	
ILVR [15]	-	25.56	0.777	0.285	635.46	68.90	23.83	24.74	0.743	0.312	1306.38	128.67	47.93	
FLAIR (Ours)		26.70	0.800	0.216	158.05	58.09	8.94	25.49	0.758	0.222	442.55	99.15	17.56	
$\mathcal{A}^+ y$		17.21	0.287	0.832	1905.12	143.77	85.97	17.77	0.299	0.827	3022.43	204.81	108.90	
VQFR [25]	E I	27.54	0.810	0.195	385.35	50.22	9.62	27.87	0.816	0.200	628.88	84.72	16.94	
RestoreFormer++ [78]	a b	28.13	0.818	0.193	322.94	47.15	7.99	27.90	0.813	0.191	527.08	78.29	13.85	
CodeFormer [93]	sia 0.0	28.64	0.825	0.193	294.92	50.09	9.09	28.04	0.816	0.192	494.30	81.99	15.65	
DR2E [79]	= ans	27.43	0.802	0.220	564.43	56.15	12.45	27.01	0.788	0.218	909.62	100.89	20.86	
DDNM [75]	Öь	30.24	0.863	0.250	320.77	74.11	34.40	29.20	0.846	0.265	629.74	112.86	50.14	
DiffPIR [95]	×	28.93	0.838	0.210	672.55	43.80	6.29	28.04	0.815	0.223	1051.06	83.07	16.86	
FLAIR (Ours)		29.87	0.856	0.149	82.82	39.54	8.25	28.15	0.818	0.179	255.44	74.47	14.40	
$\mathcal{A}^+y$	H Q	19.53	0.481	0.710	1856.50	141.39	90.56	20.15	0.472	0.696	2990.26	205.48	113.83	
VQFR [25]	i di B	27.15	0.807	0.214	483.55	54.09	10.59	26.68	0.798	0.215	807.43	94.40	19.59	
RestoreFormer++ [78]	IPE	27.12	0.806	0.214	427.63	52.58	9.42	26.83	0.797	0.214	739.01	89.94	17.20	
CodeFormer [93]	5, J	27.71	0.814	0.211	385.63	55.24	10.74	27.05	0.802	0.215	720.54	94.25	19.15	
DR2E [79]	0.0 0.0	26.58	0.789	0.242	695.99	60.39	12.89	26.01	0.773	0.243	1091.43	116.38	21.90	
DDNM [75]	× II	29.02	0.851	0.271	509.15	74.48	35.89	27.63	0.818	0.317	1067.57	126.23	57.91	
FLAIR (Ours)	.4	29.39	0.857	0.178	126.36	45.90	9.11	28.40	0.841	0.185	316.89	74.12	14.04	

Table 1. Quantitative results on two face video datasets (short clips). Our method generates better perceptual quality and data-fidelity results than SOTA face restoration baselines. **Best** and second-best values for each metric are color-coded.



Figure 4. Visual comparisons on FVR  $4 \times$  SR, motion deblurring. Our FLAIR produces higher restoration quality and better data-consistent than existing DPM-based restoration methods. See supplements for additional numerical results.

We enforce consistency of reconstructed  $\tilde{x}$  (e.g.,  $x_{0t}$  in (3)) by using a projection onto the subspace spanned by Ax, where rank $(A) = Nm \le Nd$ . We recover the consistent reconstruction by solving the following minimization problem

$$\tilde{\boldsymbol{x}}_{0t} = \operatorname*{arg\,min}_{\tilde{\boldsymbol{x}}} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}_{0t}\|_2^2 \quad \text{s.t.} \quad \mathcal{A}\tilde{\boldsymbol{x}} = \boldsymbol{y}, \tag{7}$$

corresponding to least-norm problem with equality constraints. This problem can be solved analytically [7] as

$$\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{x}_{0t} - \mathcal{A}^+ (\mathcal{A} \boldsymbol{x}_{0t} - \boldsymbol{y}), \qquad (8)$$

where  $\mathcal{A}^+ = \mathcal{A}^T (\mathcal{A}\mathcal{A}^T)^{-1} \in \mathbb{R}^{Nd \times Nm}$  is the Moore-Penrose pseudo-inverse of  $\mathcal{A}$  and satisfies  $\mathcal{A}\mathcal{A}^+ = \mathbf{I}_{Nm}$ . By substituting the estimated  $\mathbf{x}_{0t}$  with  $\tilde{\mathbf{x}}_{0t}$  in (4), we enforce the low-frequency content of  $\tilde{\mathbf{x}}_{0t}$  to align with that of the ground-truth video sequence  $\mathbf{x}$  (*i.e.*,  $\mathcal{A}\tilde{\mathbf{x}}_{0t} = \mathcal{A}\mathbf{x} = \mathbf{y}$ ), while allowing the reverse diffusion process to recover the high-frequency components. We reformulate (8) by calculating  $\mathcal{A}^+$  according to [2] for each individual frame

$$ilde{oldsymbol{x}}_{0t}^n = oldsymbol{x}_{0t}^n - ilde{oldsymbol{h}}_n st (oldsymbol{k}_n st ((oldsymbol{h}_n st oldsymbol{x}_{0t}^n) \downarrow_s - oldsymbol{y}^n)) \uparrow^s,$$

where  $\hat{h}_n$  is the mirrored version of the blur kernel  $h_n$ , and  $\uparrow^s$  denotes spatial upsampling by zero-filling of new entries.  $k_n$  is used to replace the multiplication by  $(\mathcal{A}\mathcal{A}^T)^{-1}$  and corresponds to the inverse of filter  $(h_n * \tilde{h}_n) \downarrow_s$  in Fourier domain.

Noisy FVR Degradation. In the presence of noise, the forward model is  $\boldsymbol{y} = \mathcal{A}\boldsymbol{x} + \boldsymbol{e}$ , where  $\boldsymbol{e} = \{\boldsymbol{e}^n\}_{n=1}^N$  represents additive white Gaussian noise (AWGN) with  $\boldsymbol{e}^n \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{e}}^2 \mathbf{I}_m)$ . Directly applying (8) to noisy measurements  $\boldsymbol{y}$  will result in an additional noise term  $\mathcal{A}^+\boldsymbol{e}$  in  $\tilde{\boldsymbol{x}}_{0t}$ , consequently affecting the reverse diffusion  $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0, \boldsymbol{y})$ . We can approximate  $\mathcal{A}_n^+\boldsymbol{e}^n$  as a AWGN  $\mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{e}}^2 \mathbf{I}_m)$ , given  $\mathcal{A}^+$  in FVR closely resembles a copy operation [75]. Thus (8) and (4) can be modified into

$$\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{x}_{0t} - \gamma_t \mathcal{A}^+ (\mathcal{A} \boldsymbol{x}_{0t} - \mathcal{A} \boldsymbol{x}) + \gamma_t \mathcal{A}^+ \boldsymbol{e}, \qquad (9)$$
$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \tilde{\boldsymbol{x}}_{0t} + \sqrt{1 - \bar{\alpha}_t} (\sqrt{1 - \rho_t} \tilde{\boldsymbol{\epsilon}}_t + \sqrt{\rho_t} \boldsymbol{\epsilon}),$$

where  $\gamma_t \geq 0$  and  $\rho_t > 0$  are user-defined hyperparameters such that  $\sigma_t = \sqrt{\bar{\alpha}_{t-1}\gamma_t^2\sigma_e^2 + \rho_t}$ , and  $\tilde{\epsilon}_t = \frac{1}{\sqrt{1-\bar{\alpha}_t}}(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\tilde{\boldsymbol{x}}_{0t}) \in \mathbb{R}^{Nd}$  is the recalculated noise estimate. By



Figure 5. Qualitative comparisons. Our method with different enhancement module backbones achieve higher restoration quality while maintaining data fidelity well. *Top*: FLAIR + RestoreFormer++. *Bottom*: FLAIR + CodeFormer.

Method	$16 \times$ Bicubic							$4\times$ , Gaussian blur, $\sigma = 0.05$ , JPEG 60						
Wediou	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓		
$\mathcal{A}^+ y$	16.89	0.657	0.621	4456.46	216.34	145.39	16.11	0.426	0.728	3574.66	189.92	126.43		
VRT [46]	29.38	0.866	0.287	580.65	114.91	54.26	31.18	0.889	0.238	357.04	97.58	43.50		
BasicVSRPP [11]	26.91	0.836	0.308	634.21	143.10	63.24	31.02	0.880	0.243	337.28	102.65	46.87		
RestoreFormer++ [78]	23.80	0.742	0.353	518.95	92.86	26.20	28.15	0.818	0.213	761.40	78.68	22.57		
CodeFormer [93]	23.80	0.738	0.366	1059.39	141.79	51.20	28.58	0.824	0.203	698.86	73.96	22.17		
DR2E [79]	24.83	0.764	0.312	903.16	113.63	35.93	24.83	0.764	0.312	903.16	113.63	35.93		
DDNM [75]	26.28	0.809	0.343	846.88	104.91	48.88	29.72	0.849	0.275	954.21	99.86	50.78		
FLAIR (Ours)	28.23	0.842	0.240	358.72	84.40	27.26	29.99	$-\overline{0.860}$	0.175	235.86	62.10	17.73		
FLAIR-SA (Ours)	28.96	0.855	0.268	571.36	95.90	35.97	30.57	0.873	0.199	295.60	77.17	30.51		
FLAIR+CodeFormer (Ours)	27.57	0.830	0.212	344.99	80.47	24.71	29.96	0.858	0.174	246.77	59.48	16.49		
FLAIR+RestoreFormer++ (Ours)	27.31	0.819	0.233	352.00	78.68	23.78	29.95	0.857	0.172	229.69	62.27	17.48		

Table 2. Quantitative results on CelebV-Text [85] (long clips). Our method generates better temporal coherence in the video contents compared to SOTA video restoration baselines [11,46], and it yields improved perceptual results than existing face restoration methods.

appropriately setting  $\gamma_t$  and  $\rho_t$ , we make the total noise variance in  $\boldsymbol{x}_{t-1}$  conform to the forward diffusion  $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)$  in (1). This allows for an effective estimation of noise by  $\epsilon_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}_{t-1},\boldsymbol{c},t)$  at next step.

**Composite FVR Degradation.** FLAIR is also applicable to more complicated FVR degradation

$$\boldsymbol{y}^{n} = \mathcal{E}_{n}\left(\left(\boldsymbol{h}_{n} \ast \boldsymbol{x}^{n}\right) \downarrow_{s} + \boldsymbol{e}^{n}\right), \qquad (10)$$

where  $\mathcal{E} = \{\mathcal{E}_n\}_{n=1}^N$  denotes the JPEG encoding with quality factors  $Q \ge 0$ . It is worth to note that the degradation setting in (10) aligns with current face restoration literature [25, 41, 78, 84, 93]. While JPEG is non-linear, we can construct JPEG decoding operator  $\mathcal{D}$ , such that  $\mathcal{E}(\mathcal{D}(\mathcal{E}(\boldsymbol{x}))) = \mathcal{E}(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^{Nn}$ , similar to [37], which is analogue to the matrix pseudo-inverse  $\mathcal{AA}^+\mathcal{Ax} = \mathcal{Ax}$ . For composite forward operator  $\mathcal{A} = \mathcal{A}_1 \circ ... \circ \mathcal{A}_k$ , we may approximate  $\mathcal{A}^+$  with  $\mathcal{A}^+ = \mathcal{A}_k^+ \circ ... \circ \mathcal{A}_1^+$ . Hence,  $\tilde{\boldsymbol{x}}_{0t}$ in (9) under composite degradation in (10) can be efficiently solved using

$$\tilde{\boldsymbol{x}}_{0t} = \boldsymbol{x}_{0t} - \gamma_t \mathcal{A}^+ \mathcal{D}(\mathcal{E}(\mathcal{A}\boldsymbol{x}_{0t}) - \boldsymbol{y}).$$
(11)

The full algorithm of FLAIR is detailed in the supplements.

#### 4.3. Efficient Spatial Enhancement Module

Finally, we introduce a coarse-to-fine image enhancement module designed for refinement of estimated  $\tilde{x}_{0t}$ , as

$$\tilde{\boldsymbol{x}}_{0t} = (\boldsymbol{1} - w_t \boldsymbol{m}_t) \odot \tilde{\boldsymbol{x}}_{0t} + w_t \boldsymbol{m}_t \odot \mathcal{G}(\tilde{\boldsymbol{x}}_{0t}), \qquad (12)$$

where  $w_t$  balances the importance of the facial enhancement region  $\boldsymbol{m} \odot \mathcal{G}(\tilde{\boldsymbol{x}}_{0t}) \in \mathbb{R}^{Nd}$  and originally estimated  $\boldsymbol{m}_t \odot \tilde{\boldsymbol{x}}_{0t} \in \mathbb{R}^{Nd}$  at each step, and  $(1 - \boldsymbol{m}_t) \odot \tilde{\boldsymbol{x}}_{0t}$  denotes the background scenes. Note that we *do not* impose any specific constraints on the method or architecture of  $\mathcal{G}$ , allowing the enhancement module to be trained independently. For our enhancement module, we consider two well-established backbones: Restorformer++ [78] and CodeFormer [93]. This shows the compatibility of FLAIR with a diverse range of existing methods. Both backbones make use of pre-trained high-quality VQ codebooks [23] specifically designed for face images. We refer to these methods as *FLAIR* + *Restore-Former*++ and *FLAIR* + *CodeFormer*, respectively.

#### **5.** Experiments

#### 5.1. Experimental Setup

**Datasets.** We use FFHQ [35] and 7200 clips from CelebV-Text [85] for training image DPMs. These CelebV-Text clips are then used to fine-tune the video DPMs. We choose 125 short clips and 6 long clips from the unused identities of the CelebV-Text for testing. We also consider 20 clips from CelebV-HQ [94] for testing. We additionally crawled four real-life video clips from YouTube videos, each around 150 frames for testing. See supplements for more details.

**Evaluation Metrics.** Our evaluation is based on both perception and distortion metrics. For perception, we choose three frame-wise perceptual metrics: FID [29], LPIPS [90],



Figure 6. Visual results on real-world FVR. Note that FLAIR outperforms the baselines even when the degradation is not known exactly.



Figure 7. (a): Comparison of PSNR  $\uparrow$  (left) and LPIPS  $\downarrow$  (right) of FLAIR with various controlling schedules  $\{w_t\}_{t=\tau}^{K-1}$  in (12), where K = 25 and  $\tau = 5$  for all experiments. Note the improved perception (LPIPS) quality by increasing  $w_{\tau}$ . (b): Comparison of PSNR (left) and LPIPS (right) of FLAIR w/ and w/o data-consistency module for FVR with uniform re-scheduling strategy starting from K = 5 to K = 100. Note the improved data-fidelity (PSNR) by imposing data-consistency and the trade-off between perception and distortion [6].

and KID [4]. Following the video quality assessment literature [5,62,94], we report FVD [72]. For distortion, we adopt two pixel-wise metrics: PSNR and SSIM [76].

**Training and Inference Details.** We consider three types of degradation models: video super-resolution (SR), deblurring, and JPEG restoration. For video SR, we pre-train a conditional image DPM backbone (spatial layers) using downsampling factors s = 8 with bicubic degradation and then fine-tune the video DPMs using the loss function in (6)with  $s \in 4, 8, 16$ . For video deblurring, we pre-train a conditional image DPM with scale factors s = 4 and AWGN  $\sigma_{e} \in [0, 25]$  using anisotropic Gaussian and motion kernels. For video JPEG restoration, we use the same settings as for deblurring with additional JPEG quality factor  $Q \in [60, 100]$ . For motion deblurring, we generate 100 distinct motion blur kernels and apply them to frames using (10) without JPEG encoding. We use pre-trained SPyNet [55] as our flow estimation network. For the enhancement module, we employ the original pre-trained models of RestoreFormer++ [78] and CodeFormer [93].

# 5.2. Comparisons with SOTA Methods

We present quantitative comparisons <sup>1</sup> between our FLAIR and several methods across various degradation settings in Table 1 and Table 2. VQFR [25], CodeFormer [93], and RestoreFormer++ [78] are SOTA face restoration meth-

ods using pre-trained high-quality facial dictionary priors. As there is no existing work using video diffusion models for FVR, we compare FLAIR with the latest conditional image DPMs using unconditionally trained diffusion models for inverse problems, including ILVR [15], DR2E [79], DDNM [75], and DiffPIR [95]. DR2E consists of a degradation removal module built on image DPM and an enhancement module similar to FLAIR; we use VQFR as the enhancement module for DR2E. For the DPM baselines, we pre-train an unconditional image DPM on the same training images as FLAIR. For each task, we omit any method not implemented in the original work for fair comparison.

The quantitative results on short video clips from CelebV-Text and CelebV-HQ are listed in Table 1. Overall our FLAIR achieves similar or much superior results to SOTA baselines on all evaluation metrics for different restoration tasks. The quantitative results on long video clips from CelebV-Text are listed in Table 2. The visual comparisons with DPM baselines are in Fig. 4. We additionally compare two video restoration methods: VRT [46] and BasicVS-RPP [11]. VRT is a supervised deep learning approach to video SR and deblurring, while BasicVSRPP is a deep recurrent network method specifically designed for video SR. We include FLAIR-SA (sampling average) to illustrate that different samples generated by our method achieve pixel-wise consistency in performance. By carefully constructing the forward model in (10), one may directly applying FLAIR for real-world FVR (see supplementary material for more

<sup>&</sup>lt;sup>1</sup>The comparisons only within face regions are presented in Table 1 in the supplements.

Method	$E_{warp} \downarrow (\times 10^{-3})$	Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID.
CodeFormer [93] VRT [46]	3.928 2.639	FLAIR (Image DPM) FLAIR (3D Res + RFF)	27.04	0.808	0.247	680.17 387.40	87.00 87.44	28.14
FLAIR (Image DPM) FLAIR (Video DPM)	5.625 2.546	FLAIR (3D Res. + Temp. Attr	a.) 27.64	0.830	0.251	408.85	87.46	28.72
FLAIR (Video DPM) + CodeFormer	2.531	FLAIR (All)	28.23	0.842	0.240	358.72	84.40	27.20

Table 3. Additional numerical studies. *Left*: Temporal inconsistency measured by warping error  $E_{warp}$ , lower value corresponding to smoother temporal results. *Right*: Ablation study on different FLAIR temporal layers.

Method		AMT [45] + Restoration							Restoration + AMT [45]					
	<b>PSNR</b> ↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	KID↓		
VRT [46]	30.03	0.884	0.259	509.92	90.52	42.84	30.87	0.904	0.190	275.78	62.75	27.92		
BasicVSRPP [11]	29.75	0.871	0.251	495.87	88.12	40.04	30.15	0.899	0.187	272.96	59.25	28.12		
VQFR [25]	26.19	0.799	0.248	487.78	111.00	35.71	26.35	0.816	0.251	473.47	111.51	35.93		
RestoreFormer++ [78]	26.68	0.805	0.236	415.03	100.91	32.49	26.85	0.822	0.237	477.78	99.05	31.54		
CodeFormer [93]	26.52	0.801	0.246	535.59	105.60	35.68	26.67	0.819	0.244	596.61	104.53	35.09		
DR2E + VQFR [79]	27.08	0.823	0.218	533.37	81.38	25.36	27.34	0.841	0.220	470.57	83.47	26.78		
DDNM [75]	28.93	0.863	0.266	435.32	94.87	52.76	29.19	0.872	0.247	329.15	91.06	52.84		
FLAIR (Ours)	29.04	0.866	$\bar{0}.\bar{1}60$	242.60	51.42	12.24	29.74	0.885	$\bar{0}.\bar{1}8\bar{2}$	271.80	57.70	18.43		

Table 4. Quantitative results for space-time video super-resolution (time:  $4\times$ , space:  $8\times$ ) on CelebV-Text [85] (long clips). AMT [45] is a SOTA frame interpolation method. Note that our FLAIR is only trained on spatial  $8\times$  SR task.

details), as shown in Fig 6. The running time comparisons are shown in *Sec. C* in the supplements. Note our method outperforms SOTA face restoration methods, CodeFormer and RestoreFormer++, in out-of-distribution datasets. Additional quantitative and visual comparisons are shown in the supplements.

## 5.3. Ablation Studies

Effect of Enhancement Module. We report PSNR and LPIPS results for our method in Fig. 7 (a) by adjusting the weighted schedule  $\{w_t\}_{t=\tau}^{K-1}, \tau \in [0, K-1]$  in (12). For simplicity, we have selected RestoreFormer++ for 4× SR and CodeFormer [93] for  $16 \times$  SR and JPEG Q = 60. We consider growth sequences  $1 \ge w_{\tau} > \cdots > w_{K-1} = 0$ from K - 1 to  $\tau$  and  $w_t = 0$  for  $t < \tau$ . Note that w adjusts the relative weights of the enhancement module at each intermediate step. By setting an appropriate  $w_{\tau}$ , one can achieve perception (LPIPS) improvement for all three video tasks, with a slight compromise on PSNR performance. Qualitative results in Fig. 5 show that FLAIR with the enhancement module yields superior visual outcomes.

Effect of Data-Consistency. We report PSNR and LPIPS results of our method in Fig. 7 (*b*) left and right for a mix of degradation consisting of  $4 \times$  SR, Gaussian blur and JPEG Q = 60. We see that, while both FLAIR w/ and w/o data-consistency module achieve similar LPIPS scores, FLAIR w/ data-consistency module better preserves the PSNR results. Effect of Temporal Layers. For completeness, Table 3 (*left*) shows that FLAIR with video DPMs outperforms its image DPM counterpart in temporal consistency (e.g., head movement) for video restoration. Temporal consistency is measured by the averaged flow warping error  $Ewarp(x) = \frac{1}{N-1} \sum n = 1^{N-1}E_{warp}(x^n, x^{n+1})$ , where a lower value indicates smoother temporal results. Our temporal layers enhance sequential consistency, outperforming the SOTA video restoration method, VRT. Table 3 (*right*) reports the effects of different temporal layers on performance (see

supplements for visuals). Clearly, FLAIR with both temporal layers significantly reduces the learning burden, leading to more accurate, efficient estimations, and superior results across all metrics.

## 5.4. Space-Time Video Super-Resolution

We demonstrate that pre-trained FLAIR on video SR can be combined with any video frame interpolation method for space-time video SR. Here, we use pre-trained AMT [45] for frame interpolation. In practice, FLAIR can be cascaded in two ways: AMT followed by FLAIR or FLAIR followed by AMT. As shown in Table 4, FLAIR achieves the best LPIPS, FVD, FID, and KID scores compared to existing methods, despite being a two-stage model and not specifically trained for this task. Additional details and video visual comparisons are available in the supplements.

# 6. Conclusion

In this paper, we propose the FLAIR, a novel framework based on diffusion probabilistic models for *face video restoration*. The key idea of FLAIR is to build upon pretrained image diffusion models specialized in face image restoration and to transform them into video diffusion restoration models by incorporating and fine-tuning temporal alignment layers. We further propose a two-stage refinement process at every reverse sampling step. In the first stage, FLAIR analytically imposes reconstruction fidelity by using a data-consistency module that can handle composed degradation in practice. The subsequent stage involves an enhancement module dedicated to regional improvement. Extensive comparisons show that our FLAIR framework provides temporally aligned, high-quality results in face video restoration.

# References

- M. Asim, F. Shamshad, and A. Ahmed. Blind image deconvolution using deep generative priors. *IEEE Trans. on Comput. Imag.*, 6:1493–1506, 2020. 2
- [2] Y. Bahat and T. Michaeli. Explorable super resolution. In Proc. CVPR, pages 2716–2725, 2020. 5
- [3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1167–1183, 2002. 1
- [4] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd GANs. arXiv:1801.01401, 2018. 7
- [5] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. CVPR*, pages 22563–22575, 2023. 1, 3, 4, 7
- [6] Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *Proc. CVPR*, pages 6228–6237, 2018. 2, 7
- [7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004. 5
- [8] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs. In *Proc. CVPR*, pages 109–117, 2018. 2
- [9] J. Cao, J. Liang, K. Zhang, W. Wang, Q. Wang, Y. Zhang, H. Tang, and L. Van Gool. Towards interpretable video superresolution via alternating optimization. In *Proc. ECCV*, pages 393–411. Springer, 2022. 3
- [10] A. Chakrabarti, A. Rajagopalan, and R. Chellappa. Superresolution of face images using kernel PCA-based prior. *IEEE Trans. Multimedia*, 9(4):888–892, 2007. 2
- [11] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proc. CVPR*, pages 5972–5981, 2022. 4, 6, 7, 8
- [12] M. Chang, A. Prakash, and S. Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. arXiv:2305.16301, 2023. 3
- [13] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K. K. Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proc. CVPR*, pages 11896–11905, 2021.
   2
- [14] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-toend learning face super-resolution with facial priors. In *Proc. CVPR*, pages 2492–2501, 2018. 2
- [15] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *Proc. ICCV*, pages 14347–14356, 2021. 3, 5, 7
- [16] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proc. ICLR*, 2023. 1
- [17] H. Chung, B. Sim, D. Ryu, and J. C. Ye. Improving diffusion models for inverse problems using manifold constraints. *Proc. NeurIPS*, 35:25683–25696, 2022. 3
- [18] H. Chung and J. C. Ye. Score-based diffusion models for accelerated mri. *Med. Image Anal.*, page 102479, 2022. 3

- [19] D. Danier, F. Zhang, and D. Bull. LDMVFI: Video frame interpolation with latent diffusion models. arXiv:2303.09508, 2023. 3
- [20] M. Delbracio and P. Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research*, 2023. Featured Certification. 3
- [21] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Proc. NeurIPS*, volume 34, pages 8780–8794, 2021. 1, 2, 4
- [22] B. Dogan, S. Gu, and R. Timofte. Exemplar guided face image super-resolution without facial landmarks. In *Proc. CVPR Workshops*, pages 0–0, 2019. 2
- [23] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proc. CVPR*, pages 12873–12883, 2021. 3, 6
- [24] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J. B. Huang, M. Y. Liu, and Y. Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proc. ICCV*, pages 22930–22941, 2023. 3
- [25] Y. Gu, X. Wang, L. Xie, C. Dong, G. Li, Y. Shan, and M. Cheng. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Proc. ECCV*, pages 126– 143. Springer, 2022. 1, 3, 5, 6, 7, 8
- [26] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Trans. Image Process.*, 12(5):597–606, 2003. 1, 2
- [27] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. *Proc. NeurIPS*, 35:27953–27965, 2022. 1
- [28] J. He, W. Shi, K. Chen, L. Fu, and C. Dong. GCFSR: a generative and controllable face super resolution method without facial and gan priors. In *Proc. CVPR*, pages 1889–1898, 2022. 1, 2
- [29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Proc. NeurIPS*, 30, 2017. 6
- [30] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3
- [31] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, volume 33, pages 6840–6851, 2020. 1, 3, 4
- [32] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based CNN for multi-scale face super resolution. In *Proc. ICCV*, pages 1689–1697, 2017. 2
- [33] H. Jeong, G. Y. Park, and J. C. Ye. Vmc: Video motion customization using temporal attention adaption for text-tovideo diffusion models. *arXiv preprint arXiv:2312.00845*, 2023. 3
- [34] A. Jourabloo, M. Ye, X. Liu, and L. Ren. Pose-invariant face alignment with a single CNN. In *Proc. ICCV*, pages 3200–3209, 2017. 1

- [35] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4401–4410, 2019. 2, 6
- [36] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. *Proc. NeurIPS*, 35:23593–23606, 2022. 1, 3
- [37] B. Kawar, J. Song, S. Ermon, and M. Elad. Jpeg artifact correction using denoising diffusion restoration models. *Proc. NeurIPS Workshops*, 2022. 6
- [38] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, and C. Feng. LUVLi face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proc. CVPR*, June 2020. 1
- [39] C. Laroche, A. Almansa, and E. Coupete. Fast diffusion em: a diffusion model for blind inverse problems with application to deconvolution. *arXiv:2309.00287*, 2023. 1
- [40] H. Li, Z. Guo, S. Rhee, S. Han, and J. Han. Towards accurate facial landmark detection via cascaded transformers. In *Proc. CVPR*, pages 4176–4185, June 2022. 1
- [41] W. Li, M. Wang, K. Zhang, J. Li, X. Li, Y. Zhang, G. Gao, W. Deng, and C. W. Lin. Survey on deep face restoration: From non-blind to blind and beyond. *arXiv preprint arXiv:2309.15490*, 2023. 6
- [42] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang. Blind face restoration via deep multi-scale component dictionaries. In *Proc. ECCV*, pages 399–415. Springer, 2020. 1
- [43] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang. Learning warped guidance for blind face restoration. In *Proc. ECCV*, pages 272–289, 2018. 2
- [44] Y. Li, S. Liu, J. Yang, and M. Yang. Generative face completion. In *Proc. CVPR*, pages 3911–3919, 2017. 1
- [45] Z. Li, Z. Zhu, L. Han, Q. Hou, C. Guo, and M. Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *Proc. CVPR*, pages 9801–9810, 2023. 8
- [46] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *arXiv:2201.12288*, 2022. 2, 6, 7, 8
- [47] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van G., and R. Timofte. Swinir: Image restoration using swin transformer. In *Proc. ICCV*, pages 1833–1844, 2021. 1
- [48] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool. Recurrent video restoration transformer with guided deformable attention. *NeurIPS*, 35:378–393, 2022. 1
- [49] J. Liang, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In *Proc. ICCV*, pages 4096– 4105, 2021. 3
- [50] C. Liu, H. Y. Shum, and W. T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 75:115–134, 2007. 1
- [51] H. Liu, Z. Ruan, P. Zhao, C. Dong, F. Shang, Y. Liu, L. Yang, and R. Timofte. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 55(8):5981–6035, 2022. 1

- [52] J. Liu, R. Anirudh, J. J. Thiagarajan, S. He, K. A. Mohan, U. S. Kamilov, and H. Kim. DOLCE: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proc. ICCV*, pages 10498–10508, 2023. 2
- [53] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proc. CVPR*, pages 11461–11471, 2022. 3
- [54] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proc. ICCV*, pages 22930–22941, 2023. 3
- [55] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proc. CVPR*, pages 4161– 4170, 2017. 7
- [56] M. Ren, M. Delbracio, H. Talebi, G. Gerig, and P. Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proc. ICCV*, pages 10721–10733, 2023. 1, 3
- [57] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022. 1
- [58] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. In *Proc. CVPR*, pages 4197–4206, 2016. 1
- [59] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *Proc. ACM SIGGRAPH 2022*, pages 1–10, 2022. 1, 3
- [60] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. NeurIPS*, 2022.
- [61] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 1, 2, 3
- [62] G. Shrivastava, S. N. Lim, and A. Shrivastava. Video dynamics prior: An internal learning approach for robust video enhancements. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Proc. NeurIPS*, volume 36, pages 34228–34246. Curran Associates, Inc., 2023. 7
- [63] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Textto-video generation without text-video data. In *Proc. ICLR*, 2022. 4
- [64] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, pages 2256–2265, 2015. 3
- [65] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proc. ICLR*, 2021. 3
- [66] J. Song, A. Vahdat, M. Mardani, and J. Kautz. Pseudoinverseguided diffusion models for inverse problems. In *Proc. ICLR*, 2022. 1
- [67] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1
- [68] X. Tang and X. Wang. Face sketch synthesis and recognition. In Proc. ICCV, pages 687–694. IEEE, 2003. 2

- [69] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. MAXIM: Multi-axis mlp for image processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5769–5780, 2022. 1
- [70] O. Tuzel, Y. Taguchi, and J. R. Hershey. Global-local face upsampling network. arXiv:1603.07235, 2016. 2
- [71] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proc. ICCV*, pages 3659– 3667, 2015. 1
- [72] T. Unterthiner, S. Van S., K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv:1812.01717*, 2018.
  7
- [73] T. Wang, K. Zhang, X. Chen, W. Luo, J. Deng, T. Lu, X. Cao, W. Liu, H. Li, and S. Zafeiriou. A survey of deep face restoration: Denoise, super-resolution, deblur, artifact removal. *arXiv:2211.02831*, 2022. 1, 3
- [74] X. Wang, Y. Li, H. Zhang, and Y. Shan. Towards real-world blind face restoration with generative facial prior. In *Proc. CVPR*, pages 9168–9178, 2021. 1, 2
- [75] Y. Wang, J. Yu, and J. Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *Proc. ICLR*, 2023. 1, 2, 3, 5, 6, 7, 8
- [76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, Apr 2004. 7
- [77] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proc. CVPR*, pages 17512–17521, 2022. 1, 3
- [78] Z. Wang, J. Zhang, T. Chen, W. Wang, and P. Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE TPAMI*, 2023. 2, 3, 5, 6, 7, 8
- [79] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang. DR2: Diffusion-based robust degradation remover for blind face restoration. In *Proc. CVPR*, pages 1704– 1713, 2023. 1, 3, 5, 6, 7, 8
- [80] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar. Deblurring via stochastic refinement. In *Proc. CVPR*, pages 16293–16303, 2022. 1, 2, 3
- [81] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proc. ICCV*, pages 7623–7633, 2023. 3
- [82] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, and P. Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In ACM Multimedia, pages 1551–1560, 2020.
- [83] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *Proc.* ACM SIGGRAPH Asia 2023, 2023. 3
- [84] T. Yang, P. Ren, X. Xie, and L. Zhang. GAN prior embedded network for blind face restoration in the wild. In *Proc. CVPR*, pages 672–681, 2021. 1, 2, 6

- [85] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebvtext: A large-scale facial text-video dataset. In *Proc. CVPR*, pages 14805–14814, 2023. 5, 6, 8
- [86] X. Yu and F. Porikli. Ultra-resolving face images by discriminative generative networks. In *Proc. ECCV*, pages 318–333. Springer, 2016. 2
- [87] K. Zhang, L. V. Gool, and R. Timofte. Deep unfolding network for image super-resolution. In *Proc. CVPR*, pages 3217– 3226, Jun. 2020. 1
- [88] K. Zhang, Z. Zhang, C. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang. Super-identity convolutional neural network for face hallucination. In *Proc. ECCV*, pages 183–198, 2018.
- [89] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, pages 3836–3847, 2023. 1
- [90] R. Zhang, P. Isola, A. A Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, pages 586–595, 2018. 6
- [91] Y. Zhao, T. Hou, Y. Su, X. Jia, Y. Li, and M. Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *Proc. ICCV*, pages 7312–7322, 2023.
- [92] Q. Zhou, R. Li, S. Guo, Y. Liu, J. Guo, and Z. Xu. CaDM: Codec-aware diffusion modeling for neural-enhanced video streaming. arXiv:2211.08428, 2022. 3
- [93] S. Zhou, K. Chan, C. Li, and C. C. Loy. Towards robust blind face restoration with codebook lookup transformer. *Proc. NeurIPS*, 35:30599–30611, 2022. 1, 2, 3, 5, 6, 7, 8
- [94] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *Proc. ECCV*, pages 650–667. Springer, 2022. 5, 6, 7
- [95] Y. Zhu, K. Zhang, J. Liang, J. Cao, B. Wen, R. Timofte, and L. Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proc. CVPR Workshops*, pages 1219– 1229, 2023. 5, 7