TIGER: Taming Identity, Geometry, and Generative Priors for High-Quality Face Video Restoration

Yang Zhou[1]†, Wenxue Li[1,2†], Peng Zhang[1⋆], Yifei Chen[1], Fei Wang[1], and Daiguo Zhou[1]

[1] MiLM Plus, Xiaomi Inc., Wuhan, China
[2] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
zhouyang46@xiaomi.com

Abstract. Face Video Restoration (FVR) aims to recover high-fidelity facial videos from degraded input while preserving identity and semantic consistency across frames. Existing methods often struggle to simultaneously address three key challenges: identity shift, viewpoint-entangled guidance, and perceptual realism. To tackle these issues, we propose TIGER, a structured tri-prior fusion framework that **T**ames **I**dentity, **G**eometry, and g**E**nerative p**R**iors for high-quality FVR. Specifically, an *Identity Prior* is first established by injecting subject-discriminative embeddings into the latent space, effectively anchoring the subject's identity against severe degradations. Then, to provide temporally consistent structural guidance for dynamic videos, TIGER constructs a *Geometry Prior* by lifting 2D reference cues into a disentangled 3D parameter space, creating a geometric anchor through cross-source parameter fusion. Moreover, to achieve better efficiency without compromising realism, we harness the video generation model's *Generative Prior* through a one-step rectified flow. We further design a progressive three-stage training optimazition strategy that refines structural fidelity, textural reconstruction, and distribution-level realism to ensure robust optimization. We also construct a large-scale FVR dataset to facilitate robust training and standardized evaluation. Extensive experiments demonstrate that TIGER achieves state-of-the-art performance in both identity fidelity and temporal stability, delivering a high-quality, efficient and identity-consistent FVR. Project page: https://yzhoulv.github.io/Tiger/.

Keywords: Face Video Restoration · Rectified Flow · Diffusion Model · Low-level Vision · Geometry Prior

[1] †Equal contribution.
⋆ Corresponding author.

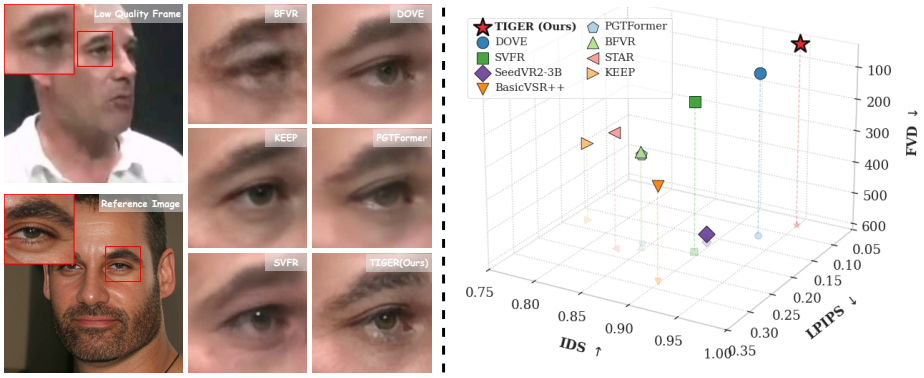

Fig. 1: Qualitative and quantitative comparisons with state-of-the-art video restoration methods. **Left**: Visual comparison of restored facial details, particularly eye regions, demonstrating that TIGER generates more realistic and perceptually pleasing results. **Right**: 3D scatter plot showing the trade-offs among IDS (identity similarity), LPIPS (perceptual quality), and FVD (temporal consistency). TIGER achieves superior performance in both visual quality and identity preservation.

1 Introduction

Recent advances in video restoration have demonstrated remarkable progress in reconstructing high-fidelity content from severely degraded inputs [3, 4, 48]. Among these tasks, Face Video Restoration (FVR) is particularly challenging due to its structural and semantic constraints. Unlike generic video restoration, FVR must simultaneously recover fine-grained facial details, preserve the subject's identity-defining characteristics, and maintain temporal coherence across dynamic expressions and pose variations [13, 56]. The difficulty is further amplified when degradations are severe, as the available observations become insufficient to reliably preserve identity.

The advent of powerful generative foundation models [32, 38] has shifted the paradigm from mere degradation removal to photorealistic synthesis, significantly boosting perceptual quality [14, 51]. However, naively adapting these generative paradigms to FVR reveals three critical failures that constitute a formidable trilemma: ***(1) Identity Drift:*** Under severe degradations, identity-discriminative cues are heavily corrupted. Consequently, strong generative priors tend to dominate the reconstruction process, producing visually plausible yet subject-inconsistent facial traits. ***(2) Viewpoint-Entangled Guidance:*** While reference images can anchor identity, a static 2D reference inherently entangles identity with a single pose and expression. When large viewpoint or motion variations occur, the model must extrapolate unseen structures from learned appearance statistics, reintroducing identity ambiguity and temporal inconsistency. ***(3) Perceptual Quality and Realism:*** To achieve efficient inference, many one-step acceleration techniques rely on distillation or regression-based objectives, which suffer from a regression-to-the-mean effect, resulting in softened textures

and diminished photorealistic realism. Balancing generative realism with computational efficiency therefore remains an open challenge.

To resolve this trilemma, we propose **TIGER**, a structured framework that **T**ames **I**dentity, **G**eometry, and g**E**nerative p**R**iors for high-quality FVR. We reformulate FVR as prior-conditioned latent transport, where identity cues, geometric structure, and generative knowledge are utilized to stabilize identity, promote geometry-consistent temporal dynamics, and preserve perceptual richness within a unified framework.

Specifically, we first establish a ***Identity Prior*** by leveraging ArcFace-based [9] identity embeddings extracted from the reference image. By injecting the subject-discriminative anchors into the diffusion backbone, TIGER ensures that the synthesized features remain unequivocally bound to the target subject's unique traits, effectively eliminating identity drift. To bridge the viewpoint-entangled guidance gap, we lift 2D reference cues into a 3D parameter space to construct a ***Geometry Prior***. By performing cross-source parameter fusion—recombining the static identity of the reference with the frame-specific motion of the degraded sequence—we generate a sequence of viewpoint-agnostic normal maps. This geometry prior serves as a dense, temporally consistent structural scaffold, ensuring that the latent transport process is strictly bounded by 3D-aware facial dynamics. Finally, we harness the foundation model's ***Generative Prior*** by formulating the restoration as a one-step rectified flow. Guided by the aforementioned geometry and identity priors, TIGER learns a direct velocity field that transports the degraded latent to the high-fidelity manifold in a single inference step. To ensure the robust optimization of this synergistic prior integration, we design a progressive three-stage training strategy that sequentially refines structural fidelity, textural reconstruction, and holistic realism.

Our main contributions are:

- We propose TIGER, a FVR framework that effectively resolves the long-standing trilemma between identity fidelity, viewpoint-entangled guidance, and perceptual realism through the integration of identity, geometry, and generative priors.
- We introduce a prior-conditioned one-step rectified-flow formulation. Equipped with cross-source 3D parameter fusion and hybrid prior injection, it achieves high-fidelity, identity-preserving restoration in a single forward pass.
- We design a progressive three-stage optimization strategy that sequentially enhances structural fidelity, fine-grained detail, and distribution-level realism.
- To facilitate robust training and standardized evaluation, we construct a new large-scale FVR dataset comprising 28,491 training samples and establish a comprehensive benchmark. Extensive experiments demonstrate that TIGER achieves state-of-the-art performance.

2 Related Work

2.1 General Video Enhancement

Early video super-resolution (VSR) and video restoration methods primarily relied on convolutional networks [2, 42], optical flow–based alignment [12], and later, Transformer-based designs [26]. To handle real-world degradations, subsequent approaches incorporated pre-cleaning modules [4], degradation kernel estimation [16], synthetic degradation pipelines [5, 35], and GAN-based objectives [6]. Temporal consistency was further enforced through alignment error minimization or motion compensation [19].

More recently, diffusion models—leveraging strong generative priors—have been adopted for image and video restoration [20,28,34,45]. In VSR, this paradigm has led to methods employing diffusion backbones [22, 33], high-capacity GANs [50], autoregressive generation [36, 54], and explicit temporal modeling [39]. To address the computational burden of iterative sampling, acceleration techniques such as distillation [29, 57] and efficient stochastic differential equation (SDE) solvers [27, 55] have been proposed. Despite these advances, general VSR frameworks often lack identity-aware modeling, resulting in identity drift and suboptimal performance in portrait-centric scenarios.

2.2 Face Video Enhancement

Blind FVR requires simultaneously preserving identity, recovering fine textures, and maintaining temporal stability under unknown and severe degradations. However, general VSR models lack explicit facial priors and often exhibit identity inconsistency, texture artifacts, and temporal flickering. This limitation has motivated a shift toward face-specific designs. Early efforts enhance inter-frame correlation through temporal modeling [11], while more recent approaches integrate generative priors with facial semantics. For example, some methods learn identity-aware representations end-to-end without explicit alignment [49]; others incorporate motion or diffusion priors to stabilize dynamic details [43], or leverage structured codebooks to improve the robustness of feature representations [47]. Moreover, model performance remains bottlenecked by training data quality. Widely used datasets like VoxCeleb [30] and CelebV-HQ [58], though large, contain in-the-wild degradations (e.g., motion blur, occlusion, and low illumination) that hinder the learning of clean facial distributions. This degraded-to-clean learning gap introduces a distribution mismatch that limits the recovery of realistic high-frequency details, highlighting the need for higher-fidelity training data.

2.3 Reference-Guided Face Restoration

Incorporating identity information from a reference image has emerged as a promising strategy to enhance identity fidelity in blind restoration. Early methods transferred identity cues via pose alignment or keypoint matching [23, 24],

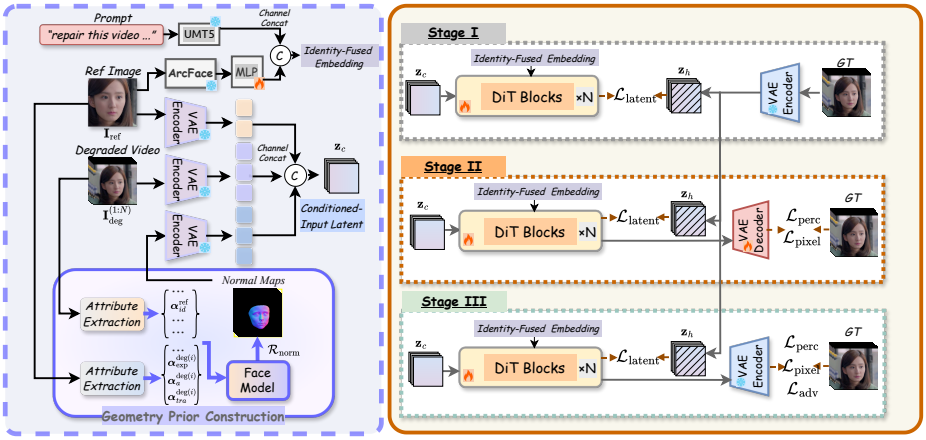

Fig. 2: Overview of the proposed **TIGER** framework, which **T**ames **I**dentity, **G**eometry, and g**E**nerative p**R**iors for high-quality FVR. Given a degraded video and a high-quality reference image, we estimate 3D facial parameters to construct geometry-aware constraints that fuse the reference identity with the degraded video's pose and camera. The resulting geometry prior, together with an identity embedding, conditions a one-step rectified-flow mapper in latent space. We adopt a progressive three-stage optimization strategy: (I) geometry-identity conditioned latent mapping, (II) decoder-coupled detail refinement, and (III) adversarial distribution alignment.

but were sensitive to input quality. Extensions to multiple references improved robustness but remained dependent on reliable landmark detection, which often fails under severe degradation and introduces temporal jitter in videos [25].

Alternative approaches construct personalized identity spaces in GAN latent domains using numerous subject images [31], achieving high fidelity at the cost of per user optimization and poor scalability. In the video domain, reference guided solutions are still limited. Existing attempts built on diffusion models [13] often neglect explicit temporal modeling and struggle with data scarcity during fine-tuning. In contrast, our method aligns the degraded video with the reference via geometric cues, enabling robust identity transfer without requiring explicit correspondence. Combined with stochastic conditioning dropout, the framework naturally supports both reference-guided and reference-free inference, achieving a flexible balance between identity fidelity, temporal consistency, and deployment efficiency.

3 Method

3.1 Overview

As illustrated in Fig. 2, we present TIGER, a framework designed to resolve the FVR trilemma by synergizing identity, geometry, and generative priors. Given a degraded video $\mathbf{I}_{\text{deg}}^{(1:N)}$ and a reference image $\mathbf{I}_{\text{ref}}$, we first extract their facial

attributes into a disentangled 3D parameter space. By performing cross-source fusion, we combine the static identity from $\mathbf{I}_{\text{ref}}$ with the dynamic motion from $\mathbf{I}_{\text{deg}}^{(1:N)}$ to render a sequence of geometry-consistent normal maps $\mathbf{I}_{\text{norm}}^{(1:N)}$. These geometric cues, along with an identity embedding generated by ArcFace [9], form the structured priors for our restoration network. We adopt a DiT-based one-step rectified-flow mapper to perform latent transport. By leveraging the generative priors of the pretrained foundation model, TIGER directly maps the degraded latent $\mathbf{z}_l$ to the clean manifold in a single forward pass, ensuring high-fidelity results with maximum efficiency.

3.2 Geometry Prior Construction

A major hurdle in FVR is that a static reference image provides identity cues that are fundamentally entangled with its specific viewpoint and expression. Relying on 2D-based reference guidance often fails when the video subject undergoes significant pose variations, as the fixed-viewpoint prior cannot provide consistent structural constraints across dynamic frames. To address this, we construct a geometry prior in the disentangled 3D parameter space, where identity- and motion-related factors can be explicitly factorized, enabling consistent identity preservation under dynamic pose variations.

Following the 3D Morphable Model (3DMM) [1], the 3D face vertices are parameterized as:

$$\mathbf{V}_{3d}(\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{\text{exp}}, \boldsymbol{\alpha}_a, \boldsymbol{\alpha}_{tra}) = \mathbf{R}(\boldsymbol{\alpha}_a)(\overline{\mathbf{V}} + \boldsymbol{\alpha}_{id}\mathbf{A}_{id} + \boldsymbol{\alpha}_{\text{exp}}\mathbf{A}_{\text{exp}}) + \boldsymbol{\alpha}_{tra}, \tag{1}$$

where $\overline{\mathbf{V}}$ denotes the mean face shape, $\mathbf{A}_{id}$ and $\mathbf{A}_{exp}$ are the identity and expression bases learned from statistical face models. The parameters $\boldsymbol{\alpha}_{id}$, $\boldsymbol{\alpha}_{exp}$, $\boldsymbol{\alpha}_a$, and $\boldsymbol{\alpha}_{tra}$ control identity, expression, rotation, and translation, extracted from the pretrained 3DDFA-V3 [44] respectively.

Decoupled Attribute Extraction. To obtain the coefficients for identity and motion, we employ a pretrained 3D face reconstructor (3DDFA-V3 [44]) as a spatial parameter encoder $\mathcal{E}_{3d}$. For the high-quality reference image $\mathbf{I}_{\text{ref}}$, we extract its static identity representation: $\boldsymbol{\alpha}_{id}^{\text{ref}} = \mathcal{E}_{3d}(\mathbf{I}_{\text{ref}})$. Simultaneously, for each degraded frame $\mathbf{I}_{\text{deg}}^{(i)}$, we extract the time-varying motion parameters: $\{\boldsymbol{\alpha}_{\text{exp}}^{\text{deg}(i)}, \boldsymbol{\alpha}_a^{\text{deg}(i)}, \boldsymbol{\alpha}_{tra}^{\text{deg}(i)}\} = \mathcal{E}_{3d}(\mathbf{I}_{\text{deg}}^{(i)})$. This step maps raw pixels into a common, disentangled parameter space, enabling the subsequent cross-source fusion.

Cross-source Parameter Fusion. To construct geometry-aware guidance that preserves identity and tracks motion dynamics, we explicitly decompose facial geometry estimation into complementary identity and motion streams. This design addresses the limited identity guidance of 2D references: while $\mathbf{I}_{\text{ref}}$ provides clean identity traits, it lacks motion; conversely, $\mathbf{I}_{\text{deg}}^{(1:N)}$ contains the target motion but suffers from identity ambiguity.

Specifically, we decompose the 3DMM parameter estimation into two complementary streams:

- ***Identity from Reference:*** We extract identity coefficients $\boldsymbol{\alpha}_{id}^{\text{ref}}$ from the high-quality $\mathbf{I}_{\text{ref}}$. These coefficients represent view-invariant facial traits and are kept fixed across all frames to ensure rigorous identity preservation.
- ***Motion from Degraded Video:*** For each frame $i \in \{1, \ldots, N\}$ in $\mathbf{I}_{\text{deg}}^{(1:N)}$, we estimate the dynamic parameters: pose $\boldsymbol{\alpha}_{a}^{\text{deg}(i)}$, expression $\boldsymbol{\alpha}_{\text{exp}}^{\text{deg}(i)}$, and translation $\boldsymbol{\alpha}_{tra}^{\text{deg}(i)}$. These capture the precise motion dynamics of the input video despite the presence of degradations.

By fusing the static identity from the reference with the dynamic motion from the degraded video, we construct a hybrid 3D geometry for each frame:

$$\mathbf{V}_{3d}^{(i)} = \mathbf{V}_{3d}\big(\boldsymbol{\alpha}_{id}^{\text{ref}}, \boldsymbol{\alpha}_{\text{exp}}^{\text{deg}(i)}, \boldsymbol{\alpha}_{a}^{\text{deg}(i)}, \boldsymbol{\alpha}_{tra}^{\text{deg}(i)}\big). \tag{2}$$

Finally, we employ a fixed perspective camera following [44] to project the 3D vertices onto the 2D image plane. We render surface normal maps using the differentiable renderer $\mathcal{R}_{\text{norm}}$, as

$$\mathbf{I}_{\text{norm}}^{(i)} = \mathcal{R}_{\text{norm}}\big(\mathbf{V}_{3d}^{(i)}\big), \tag{3}$$

These normal maps serve as a view-agnostic geometric prior, providing the restoration network with a structurally consistent scaffold that maintains the subject's identity across diverse poses and expressions.

3.3 Prior-Conditioned One-Step Rectified Flow

To bridge the gap between degraded inputs and photorealistic reconstructions, we formulate FVR as a prior-conditioned latent transport problem. Unlike standard diffusion models that rely on stochastic multi-step sampling, we leverage a one-step rectified flow to directly map the degraded distribution to the clean data manifold.

Specifically, given the ground-truth video $\mathbf{I}^{(1:N)}$ and its degraded version $\mathbf{I}_{\text{deg}}^{(1:N)}$, we operate in the latent space of a pretrained video VAE with encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$. The clean and degraded sequences are encoded as:

$$\mathbf{z}_h = \mathcal{E}(\mathbf{I}^{(1:N)}), \quad \mathbf{z}_l = \mathcal{E}(\mathbf{I}_{\text{deg}}^{(1:N)}), \tag{4}$$

where $\mathbf{z}_h \sim p_{\text{data}}$ represents the latent distribution of natural videos. We define the latent residual $\mathbf{y} = \mathbf{z}_h - \mathbf{z}_l$ as the restoration target. Following the rectified flow [10], we construct a linear probability path between the degraded and clean latents:

$$\mathbf{z}_t = (1-t)\,\mathbf{z}_l + t\,\mathbf{z}_h, \quad t \in [0,1]. \tag{5}$$

The theoretical velocity of this path is constant: $\mathbf{v}_t = \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_h - \mathbf{z}_l$. The velocity predictor $\mathbf{v}_\theta$ is implemented using a Diffusion Transformer (DiT).

To provide dense geometric and visual constraints frame-by-frame, we encode geometry-aware normal maps $\mathbf{I}_{\text{norm}}^{(1:N)}$ and the reference image $\mathbf{I}_{\text{ref}}$ into the latent

space. These are concatenated with the degraded latent $\mathbf{z}_l$ along the channel dimension, forming a spatially-aligned input

$$\mathbf{z}_c = [\mathbf{z}_l, \mathcal{E}(\mathbf{I}_{\text{norm}}^{(1:N)}), \mathcal{E}(\mathbf{I}_{\text{ref}})]. \tag{6}$$

This ensures that the velocity $\mathbf{v}_\theta$ is strictly bounded by the 3D geometry and reference appearance. To further anchor the subject's identity, we leverage the projected ArcFace [9] embedding $\mathbf{f}_{\text{arc}}$ extracted from $\mathbf{I}_{\text{ref}}$. We fuse the identity embedding with text embeddings to construct the identity-fused embedding $\mathbf{f}_{\text{id}}$, which is injected into the DiT blocks via the cross-attention mechanism. This design forces the generative process to attend to identity features for eliminating identity drift.

Instead of relying on unguided velocity estimation, we incorporate the proposed priors to rectify the transport direction as:

$$\hat{\mathbf{z}}_h = \mathbf{z}_l + \mathbf{v}_\theta\big(\mathbf{z}_c, \mathbf{f}_{\text{id}}, t^*\big), \tag{7}$$

where t^* is the fixed rectification time (set to 0.4 for both training and inference).

3.4 Progressive Three-Stage Optimization

To synergize the structured priors and the generative capabilities of the foundation model, we adopt a coarse-to-fine optimization strategy that progressively refines identity fidelity, textural detail, and holistic realism.

Stage I: Geometry-Identity Conditioned Latent Mapping. The initial stage focuses on establishing a robust one-step mapping from the degraded latent $\mathbf{z}_l$ to the clean manifold $\mathbf{z}_h$. We feed the spatially-aligned input $\mathbf{z}_c$ into the DiT backbone, while anchoring the identity via cross-attention with $\mathbf{f}_{\text{arc}}$. The model is optimized by minimizing the latent-space MSE:

$$\mathcal{L}_{\text{latent}} = \mathbb{E}\left[\|\hat{\mathbf{z}}_h - \mathbf{z}_h\|_2^2\right]. \tag{8}$$

The training objective of Stage I is defined as $\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{latent}}$. This stage tames the model to reconstruct basic facial structures and identity consistent with the guidance.

Stage II: Decoder-Coupled Detail Refinement. While Stage I achieves structural restoration, the pretrained VAE decoder $\mathcal{D}$ may introduce texture softening when processing restored latents. To alleviate this, we decode $\hat{\mathbf{z}}_h$ into frames $\hat{\mathbf{I}}_{\text{rec}}^{(1:N)} = \mathcal{D}(\hat{\mathbf{z}}_h)$ and jointly optimize DiT and $\mathcal{D}$ with a perceptual loss (LPIPS [52]) in addition to latent and pixel losses:

$$\mathcal{L}_{\text{Stage2}} = \lambda_1 \mathcal{L}_{\text{latent}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{pixel}}. \tag{9}$$

$\mathcal{L}_{pixel}$ is the MSE loss between the output and the ground truth. This strengthens texture and edge fidelity while preserving identity and pose accuracy.

Stage III: Adversarial Distribution Alignment. To bridge the perceptual gap, we employ adversarial training with a curated face-video dataset to better

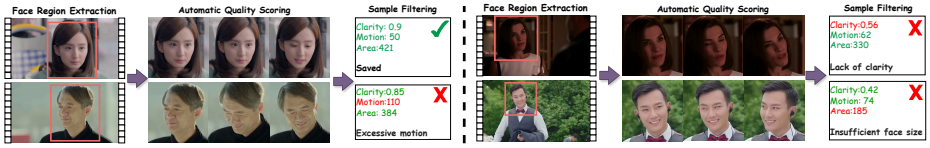

Fig. 3: Pipeline of the proposed high-quality face video filtering with three core dimensions: (1) motion amplitude, (2) face clarity, and (3) face proportion.

fit the distribution of real portrait videos. A spatiotemporal discriminator $\mathcal{D}_\phi$ is employed to supervise the realism and temporal coherence. The overall objective is:

$$\mathcal{L}_{\text{Stage3}} = \lambda_1 \mathcal{L}_{\text{latent}} + \lambda_2 \mathcal{L}_{\text{perc}} + \lambda_3 \mathcal{L}_{\text{pixel}} + \lambda_4 \mathcal{L}_{\text{adv}}. \tag{10}$$

This improves global coherence and perceptual quality while maintaining identity faithfulness and temporal stability.

To improve robustness, we apply conditional dropout to $\mathbf{z}_c$ during training: with a certain probability, we drop the identity embedding or the geometry control stream, making the model resilient to missing or unreliable priors at inference. At the inference time, the model can operate with reference guidance when available, or fall back to geometry-only or unguided restoration under challenging conditions.

3.5 Quality-Aware Face-Centric Dataset Construction

Although existing mainstream face video datasets such as VFHQ [46] and CelebVHQ [58] contain high-resolution portrait videos, they still exhibit notable limitations for FVR. First, the facial regions often occupy a limited proportion of each frame, causing models to implicitly emphasize background reconstruction at the expense of fine-grained facial textures. Second, many samples are affected by motion blur, low illumination, or compression artifacts. Without reliable high-frequency facial details as supervision, models are prone to generating spurious textures during inference, thereby degrading restoration fidelity.

To address these challenges, we construct a quality-aware, face-centric training dataset via a dedicated video selection pipeline (Fig. 3). We collect raw videos from three public datasets: OpenHumanVid [21], CelebV-HQ [58], and VFHQ [46]. In the quality filtering stage, we establish quantitative criteria across three core dimensions: motion amplitude, face clarity, and face proportion. Motion amplitude is quantified by facial centroid displacement, with a threshold of less than 100 pixels to ensure temporal stability. Face clarity is evaluated via normalized Laplacian variance, scaled to $[0, 1]$ using a logarithmic function with a variance upper bound of 1000. Samples are filtered to retain only those with a clarity score above 0.7, thereby excluding blurry frames. Face proportion requires the minimum side length of the facial region to exceed 320 pixels, ensuring that faces occupy a significant proportion of frames and provide sufficient texture

information. Through this process, we provide high-quality training data with geometric consistency, rich details, and controllable degradation levels for FVR.

After filtering, we retain 28,491 high-quality face video samples for training. Reference images are obtained by random sampling from video frames followed by background removal, ensuring that the model focuses on learning geometric priors and reconstructing facial details. The resulting dataset provides high-resolution, geometrically consistent, and detail-rich supervision tailored for FVR.

4 Experiments

4.1 Implementation details

Training Details. We utilize Wan2.1-Fun-1.3B [38] as the base model. The DiT backbone is fine-tuned via LoRA [15] with a rank of 384, whereas the VAE decoder is optimized with full-parameter fine-tuning. Training is conducted for 10 epochs on 8 GPUs with a batch size of 8. For all training phases, we set the learning rate to 10^{-4} and adopt the Adam [18] optimizer. In Stage I, each video clip consists of 81 frames. In Stages II and III, the clip length is set to 21 frames to reduce computational overhead. All videos are resized to a spatial resolution of 512×512 for training.

Datasets. For Stage I, we use 28,491 filtered facial video samples from three datasets (OpenHumanVid [21], CelebV-HQ [58], and VFHQ [46]) to train the model's ability to adhere to geometry priors. The reference images are sampled randomly from video frames with background removed. Following common practices [47], Stages II and III are trained on filtered video samples from VFHQ. In addition to the standard 50 test samples from VFHQ-Test, we construct a new test benchmark for reference-guided FVR without ground truth (GT) based on the VoxCeleb2 [8] dataset. This benchmark consists of 100 low-quality real-world video samples and corresponding extra-scene photos of the same subjects as reference images.

Evaluation Metrics. For synthetic VFHQ-Test dataset with GT, we evaluate performance using PSNR, SSIM, LPIPS [52], and FVD [37]. For real-world datasets, we adopt CLIP-IQA [40], MUSIQ [17], and LIQE [53]. Furthermore, we use IDS (identity similarity distance) to assess identity preservation, which is defined as the cosine similarity with ArcFace [9]. To measure inter-frame identity consistency, we propose VIRD (video identity and reference distance), representing the identity similarity between each frame in the facial video and the reference image. The metric is defined as:

$$\text{VIRD} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{F}_{Arc}(I_i), \mathcal{F}_{Arc}(I_{ref})\|_2, \tag{11}$$

where we use ArcFace [9] as the feature extractor $\mathcal{F}_{Arc}(\cdot)$ to obtain discriminative identity embeddings.

Table 1: Quantitative comparison results on the VFHQ-Test dataset. ↑ / ↓ denote higher / lower is better, respectively. We highlight the best and second metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IDS ↑	FVD ↓	LIQE ↑	MUSIQ ↑	CLIP-IQA ↑	VIRD ↓
DOVE	30.4082	0.8751	0.1189	0.9283	103.892	3.7983	69.9454	0.4654	0.4891
BasicVSR++	27.9281	0.8216	0.2822	0.8970	325.657	1.6914	43.9436	0.3498	0.5355
STAR	24.6980	0.7821	0.2307	0.8259	251.939	3.8819	68.4678	0.5889	0.6481
SeedVR2-3B	23.1543	0.7848	0.1638	0.8931	593.618	3.7019	67.8074	0.5076	0.5438
KEEP	28.0798	0.8488	0.1776	0.7682	370.081	2.9652	61.3693	0.4097	0.7430
BFVR	26.8917	0.8377	0.2128	0.8454	316.215	1.8835	53.3688	0.3725	0.6265
PGTFormer	28.7534	0.8466	0.2052	0.8421	334.736	2.9345	64.5613	0.5059	0.6310
SVFR	25.6862	0.8019	0.1936	0.8943	147.999	3.2908	65.8764	0.4588	0.5496
TIGER(Ours)	30.7289	0.8773	0.0697	0.9481	40.4222	4.0919	70.1086	0.4756	0.4536

4.2 Comparison With State-of-the-Art

Compared Methods. We conducted a comparison with current state-of-the-art video enhancement methods, including BasicVSR++ [3], DOVE [7], STAR [48], and SeedVR2-3B [41], as well as blind FVR algorithms KEEP [11], BFVR [47], PGTFormer [49], and SVFR [43]. Among these, DOVE, STAR, SeedVR2-3B, and SVFR are all built upon generative priors of video generation models, leveraging their capacity to model complex video distributions and recover high-quality facial details from degraded inputs.

Quantitative Results. Our method demonstrates superior performance over SoTAs across both VFHQ-Test and VoxCeleb2 datasets, excelling in critical metrics that measure visual quality, perceptual fidelity, identity preservation, and temporal consistency. On VFHQ-Test, it outperforms SoTAs in eight of nine metrics—significantly improving on LPIPS, IDS, and temporal stability (FVD, VIRD)—while establishing new benchmarks in visual quality (PSNR, SSIM) and LIQE. Similarly, on VoxCeleb2, it dominates three metrics, showcasing exceptional perceptual quality (LIQE, MUSIQ) and consistency (VIRD). Collectively, these results validate its efficacy in balancing identity retention, temporal coherence, and visual realism, establishing its superiority across FVR task.

Qualitative Results. We provide visual comparisons in Fig. 4. Our method produces more realistic and coherent results compared to other approaches. For instance, our model successfully reconstructs fine details such as skin texture and facial features, while other methods often produce overly smooth or distorted results. Similarly, in complex motion scenarios, our approach maintains sharper restoration with better preservation of dynamic elements across frames.

We also visualize the temporal profile in Fig. 5, demonstrating that our method achieves excellent temporal consistency even under challenging conditions. Existing approaches often struggle with complex degradations, exhibiting visible artifacts such as misalignment or blurring between consecutive frames. This results in a more natural and visually pleasing restoration that better aligns with human perception.

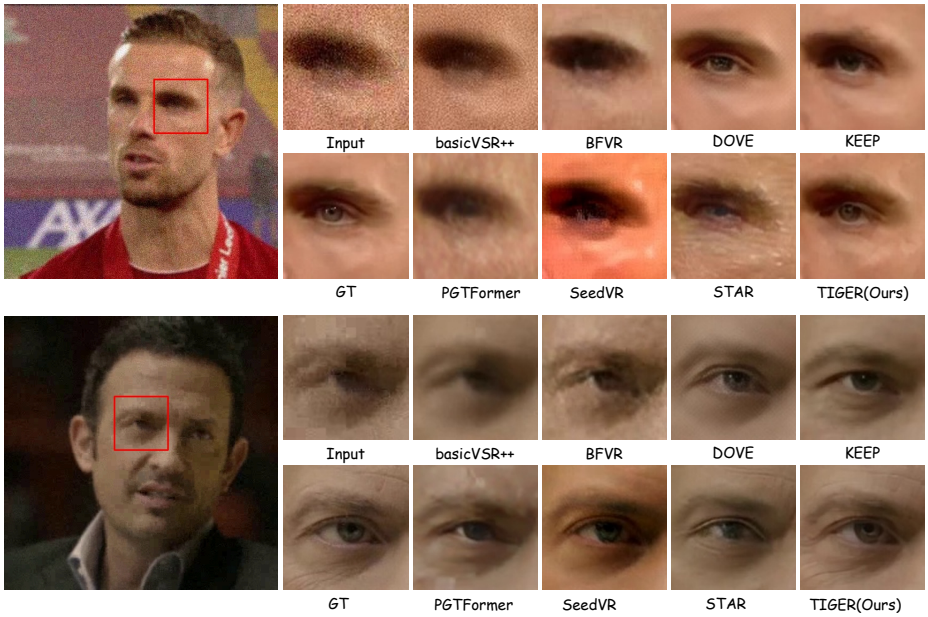

Fig. 4: Qualitative comparison with other state-of-the-art models. Our approach better preserves identity-discriminative features and restores more fine-grained facial details.

4.3 Ablation Studies

We conduct ablation studies to validate the effectiveness of our approach. Specifically, we evaluate the impact of reference priors, the contribution of each training stage, cross-source parameter fusion, and identity conditioning on VFHQ-Test. Additionally, the loss weight balancing strategy is assessed on VoxCeleb2.

Effectiveness of Different Priors. We analyze the contributions of the identity and geometry priors. "w/o Ide." removes the identity input, while "w/o Geo." disables the 3D normal-based geometry prior. VIRD evaluates identity fidelity, LPIPS measures perceptual similarity, and FVD reflects overall video consistency. As shown in Tab. 3, removing both priors leads to the worst performance, with degraded identity preservation and visual quality. Using only a reference image improves results significantly, as it provides strong identity cues through identity embeddings. However, without explicit geometric guidance, the model lacks structural consistency across poses and expressions. The best performance is achieved when both reference and geometry priors are employed, demonstrating their complementary roles: the reference anchors identity, while geometry enforces cross-frame structural coherence.

Effectiveness of Three Stage Pipeline. We conduct a ablation study by incrementally introducing each stage of the proposed three-stage training pipeline. Specifically, we evaluate: (i) training with only Stage I, (ii) Stage I + Stage II, and (iii) the full three-stage pipeline. As shown in Tab. 4, Stage I training only

Table 2: Quantitative comparison of our method with SoTA approaches on the VoxCeleb2 dataset.

Method	LIQE ↑	MUSIQ ↑	CLIP-IQA ↑	VIRD ↓
DOVE	2.0396	58.8717	0.3731	0.9360
BasicVSR++	1.0208	37.3428	0.2847	0.9470
STAR	2.8615	59.9601	0.5706	0.9625
SeedVR	1.3459	48.9973	0.3143	0.9408
KEEP	2.1701	55.3556	0.3630	0.9767
BFVR	1.0706	43.0584	0.3191	0.9819
PGTFormer	1.8864	56.8901	0.4370	0.9641
SVFR	2.1780	57.7379	0.4031	0.9536
TIGER(Ours)	3.7412	66.2382	0.4933	0.9322

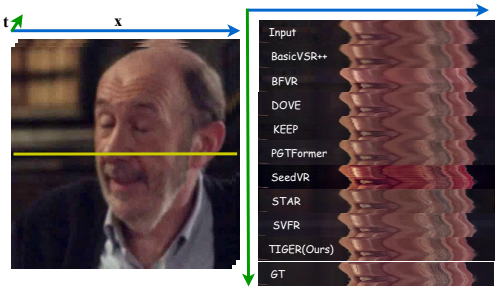

Fig. 5: Qualitative comparison of temporal consistency, stacking the yellow line across frames.

Table 3: Ablation study on the effectiveness of identity (Ide.) and geometry (Geo.) priors.

Ide.	Geo.	VIRD ↓	FVD ↓	LPIPS ↓
×	×	0.4940	101.5778	0.1942
×	✓	0.4791	70.4368	0.1275
✓	×	0.4783	71.1703	0.1272
✓	✓	**0.4492**	**40.4222**	**0.0697**

Table 4: Ablation study on the effectiveness of the progressive three-stage optimization.

I	II	III	VIRD ↓	FVD ↓	LPIPS ↓
✓			0.4774	70.0517	0.1261
✓	✓		0.4628	60.5291	0.0911
✓	✓	✓	**0.4492**	**40.4222**	**0.0697**

the latent mapper yields limited perceptual quality despite reasonable identity preservation. In Stage II, joint fine-tuning of the mapper and the VAE decoder significantly improves visual fidelity by refining textures and edges. To stabilize adversarial training in Stage III, we freeze the decoder parameters to prevent mode collapse and performance degradation, while optimizing the mapper with GAN loss. This final stage further boosts both identity consistency and perceptual quality, demonstrating the necessity of staged optimization.

Impact of Cross-source Parameter Fusion. We ablate the parameter sources for geometry construction in Tab. 5. Using parameters solely from the degraded video leads to unreliable identity cues due to the severe degradation. Conversely, extracting both from the reference images provides stable identity but fails to capture the true motion dynamics, resulting in degraded performance. Disentangling identity and motion in the 3D parameter space and recombining them from complementary sources provides more reliable geometric guidance for FVR.

Impact of Identity Conditioning. We evaluate the impact of introducing identity embeddings into the cross-attention layers. As shown in Tab. 6, compared to the text-only baseline, incorporating identity embeddings significantly improves overall performance.

Impact of Loss Weights. Tab. 7 presents the ablation results under different weight configurations. Stronger latent supervision or weaker pixel loss degrades identity consistency, while increasing the adversarial weight improves perceptual quality at the cost of slight fidelity drop. The configuration $(1.0, 2.0, 10.0, 0.2)$

Fig. 6: Qualitative of ablation study on the three-stage training pipeline.

Table 5: Effectiveness of cross-source parameter fusion. *ref* and *deg* indicate parameters extracted from reference image and degraded video, respectively.

Identity Source	Motion Source	VIRD ↓	FVD ↓	LPIPS ↓
deg	*deg*	0.4790	71.0134	0.1275
ref	*ref*	0.4781	71.3495	0.1274
ref	*deg*	**0.4492**	**40.4222**	**0.0697**

Table 6: Ablation on the identity-fused embedding.

Text Embedding	Identity Embedding	VIRD ↓	FVD ↓	LPIPS ↓
×	✓	0.6122	618.2963	0.3666
✓	×	0.4501	41.0479	0.0698
✓	✓	**0.4492**	**40.4222**	**0.0697**

achieves the best trade-off between identity preservation and visual realism and is thus adopted.

Table 7: Impact of loss weight configurations (λ_1, λ_2, λ_3, λ_4) on latent supervision, perceptual similarity, pixel reconstruction, and adversarial learning.

λ_1	λ_2	λ_3	λ_4	VIRD ↓	LIQE ↑
1.0	2.0	10.0	0.1	**0.9280**	3.0988
2.0	2.0	10.0	0.1	0.9315	2.8255
1.0	2.0	5.0	0.1	0.9356	3.1156
1.0	3.0	10.0	0.1	0.9452	3.1068
1.0	2.0	10.0	0.2	0.9322	**3.7412**

Table 8: Comparison of inference step (Step), running time (Time), and FVD (Performance) of different diffusion-based methods.

Method	Step	Time (s)	FVD ↓
STAR	15	396.52	251.939
SVFR	16	130.12	147.999
SeedVR2-3B	**1**	18.46	593.618
DOVE	**1**	**11.83**	103.892
TIGER(Ours)	**1**	17.53	**40.4222**

4.4 Running Time Comparisons

We compare the inference step, running time, and performance of different diffusion-based video enhancing methods in Tab. 8. For fairness, all methods are measured running time on the same GPU, generating a 81-frame 512×512 video. Our method achieves the best performence, attributed to our control prior.

Although calculating the control prior incurs an additional overhead of nearly 7s, the overall performance gain is significant, making the trade-off worthwhile.

5 Conclusion

In this work, we introduced TIGER, a diffusion-based, one-step inference framework for FVR that excels in identity preservation, temporal consistency, and visual quality. Our approach synergistically **T**aming **I**dentity, **G**eometry, and g**E**nerative p**R**iors to produce high-fidelity, coherent results. To mitigate identity drift, we anchor the restoration process using robust identity embeddings from a reference image. Coherence across frames is enforced by incorporating 3D geometric priors as structural constraints. The one-step inference design, augmented with adversarial training for texture refinement, ensures both speed and photorealism. Furthermore, we propose a three-stage training strategy and a new large-scale benchmark dataset to facilitate stable optimization and rigorous evaluation. Extensive experiments confirm that TIGER sets a new state-of-the-art in reference-guided FVR. Future work will focus on expanding our training data to encompass a wider range of real-world variations, pushing the boundaries of photorealistic FVR.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 187–194. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co. (1999) 6
2. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4947–4956 (2021) 4
3. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5972–5981 (2022) 2, 11
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5962–5971 (2022) 2, 4
5. Chang, K.C., Wang, R., Lin, H.J., Liu, Y.L., Chen, C.P., Chang, Y.L., Chen, H.T.: Learning camera-aware noise models. In: European Conference on Computer Vision. pp. 343–358. Springer (2020) 4
6. Chen, R., Mu, Y., Zhang, Y.: High-order relational generative adversarial network for video super-resolution. Pattern Recognition **146**, 110059 (2024) 4
7. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: DOVE: Efficient one-step diffusion model for real-world video super-resolution. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 11
8. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018) 10

9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 3, 6, 8, 10
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 7
11. Feng, R., Li, C., Loy, C.C.: Kalman-inspired feature propagation for video face super-resolution. In: European Conference on Computer Vision. pp. 202–218. Springer (2024) 4, 11
12. Guo, Z., Li, W., Loy, C.C.: Generalizable implicit motion modeling for video frame interpolation. In: Advances in Neural Information Processing Systems. vol. 37, pp. 63747–63770 (2024) 4
13. Han, W., Lin, W., Zhou, Y., Liu, Q., Wang, S., Yao, C., Chen, J.: Show and polish: reference-guided identity preservation in face video restoration. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10315–10324 (2025) 2, 5
14. He, J., Xue, T., Liu, D., Lin, X., Gao, P., Lin, D., Qiao, Y., Ouyang, W., Liu, Z.: Venhancer: Generative space-time enhancement for video generation. arXiv preprint arXiv:2407.07667 (2024) 2
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 10
16. Ji, X., Cao, Y., Tai, Y., Wang, C., Li, J., Huang, F.: Real-world super-resolution via kernel estimation and noise injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 466–467 (2020) 4
17. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021) 10
18. Kinga, D., Adam, J.B., et al.: A method for stochastic optimization. In: International conference on learning representations (ICLR). vol. 5. California; (2015) 10
19. Lei, C., Xing, Y., Chen, Q.: Blind video temporal consistency via deep video prior. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1083–1093 (2020) 4
20. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022) 4
21. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7752–7762 (2025) 9, 10
22. Li, X., Liu, Y., Cao, S., Chen, Z., Zhuang, S., Chen, X., He, Y., Wang, Y., Qiao, Y.: Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. arXiv preprint arXiv:2501.10110 (2025) 4
23. Li, X., Li, W., Ren, D., Zhang, H., Wang, M., Zuo, W.: Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2706–2715 (2020) 4

24. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 272–289 (2018) 4
25. Li, X., Zhang, S., Zhou, S., Zhang, L., Zuo, W.: Learning dual memory dictionaries for blind face restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(5), 5904–5917 (2022) 5
26. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022) 4
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research pp. 1–22 (2025) 4
28. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022) 4
29. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14297–14306 (2023) 4
30. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language **60**, 101027 (2020) 4
31. Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., Cohen-Or, D.: MyStyle: A personalized generative prior. ACM Transactions on Graphics (TOG) **41**(6), 1–10 (2022) 5
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2
33. Rota, C., Buzzelli, M., van de Weijer, J.: Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. In: European Conference on Computer Vision. pp. 36–53. Springer (2024) 4
34. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022) 4
35. Song, Y., Wang, M., Yang, Z., Xian, X., Shi, Y.: Negvsr: Augmenting negatives for generalized noise modeling in real-world video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10705–10713 (2024) 4
36. Sun, M., Wang, W., Li, G., Liu, J., Sun, J., Feng, W., Lao, S., Zhou, S., He, Q., Liu, J.: Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7364–7373 (2025) 4
37. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: FVD: A new metric for video generation (2019) 10
38. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 2, 10
39. Wang, H., Liu, Y., Liu, H., Wang, C.C., Guo, Y., Li, H., Wang, B., Sun, J.: Temporal-consistent video restoration with pre-trained diffusion models. arXiv preprint arXiv:2503.14863 (2025) 4

40. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023) 10
41. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: Seedvr2: One-step video restoration via diffusion adversarial post-training. arXiv preprint arXiv:2506.05301 (2025) 11
42. Wang, X., Chan, K.C., Yu, K., Dong, C., Loy, C.C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1954–1963 (2019) 4
43. Wang, Z., Chen, X., Xu, C., Zhu, J., Hu, X., Zhang, J., Wang, C., Liu, Y., Zhou, Y., Ji, R.: Svfr: A unified framework for generalized video face restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7406–7415 (2025) 4, 11
44. Wang, Z., Zhu, X., Zhang, T., Wang, B., Lei, Z.: 3d face reconstruction with the geometric guidance of facial part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1672–1682 (2024) 6, 7
45. Weng, S., Zheng, H., Zhan, P., Hong, Y., Jiang, H., Li, S., Shi, B.: Vires: Video instance repainting with sketch and text guidance. arXiv preprint arXiv:2411.16199 (2024) 4
46. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: VHFQ: A high-quality dataset and benchmark for video face super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 657–666 (2022) 9, 10
47. Xie, L., Zheng, B., Wu, S., Wong, H.S.: Dynamic content prediction with motion-aware priors for blind face video restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17821–17830 (2025) 4, 10, 11
48. Xie, R., Liu, Y., Zhou, P., Zhao, C., Zhou, J., Zhang, K., Zhang, Z., Yang, J., Yang, Z., Tai, Y.: STAR: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17108–17118 (October 2025) 2, 11
49. Xu, K., Xu, L., He, G., Yu, W., Li, Y.: Beyond alignment: Blind video face restoration via parsing-guided temporal-coherent transformer. IJCAI 2024 (2024) 4, 11
50. Xu, Y., Park, T., Zhang, R., Zhou, Y., Shechtman, E., Liu, F., Huang, J.B., Liu, D.: Videogigagan: Towards detail-rich video super-resolution (2024), unpublished manuscript 4
51. Yeh, C.H., Lin, C.Y., Wang, Z., Hsiao, C.W., Chen, T.H., Shiu, H.S., Liu, Y.L.: Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models. arXiv preprint arXiv:2407.01519 (2024) 2
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 8, 10
53. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14071–14081 (2023) 10
54. Zhang, Z., Liu, K., Chen, Z., Li, X., Chen, Y., Duan, B., Kong, L., Zhang, Y.: Infvsr: Breaking length limits of generic video super-resolution. arXiv preprint arXiv:2510.00948 (2025) 4

55. Zheng, K., Lu, C., Chen, J., Zhu, J.: Dpmsolver-v3: Improved diffusion ode solver with empirical model statistics. In: Advances in Neural Information Processing Systems. vol. 36, pp. 55502–55542 (2023) 4
56. Zhou, S., Chan, K., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. Advances in Neural Information Processing Systems **35**, 30599–30611 (2022) 2
57. Zhou, Z., Chen, D., Wang, C., Chen, C., Lyu, S.: Simple and fast distillation of diffusion models. In: Advances in Neural Information Processing Systems. vol. 37, pp. 40831–40860 (2024) 4
58. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: CelebV-HQ: A large-scale video facial attributes dataset. In: ECCV (2022) 4, 9, 10