DTI: Dynamic Trajectory Initialization for Generative Face Video Super-Resolution

Yingwei Tang[1], Chen Yan[1], Wendi Liu[1], Qiang Hu[1]*, and Xiaoyun Zhang[1]

Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
{xcn_tyw, yc.super, lwd624, qiang.hu, xiaoyun.zhang}@sjtu.edu.cn

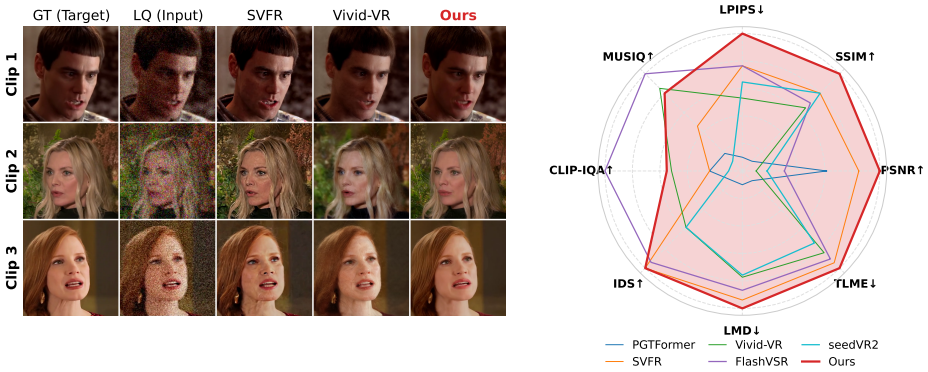


Fig. 1: Left: Qualitative comparison of video restoration results. **Right:** comparison using normalized metrics (max-participant-based normalization). Our method achieves the best balance between fidelity and perceptual quality.

Abstract. As the most perceptually powerful Face Video Super-Resolution (FVSR) method, existing works in Generative FVSR (GFVSR) mainly exploit the generative prior of pretrained diffusion models. However, viewed as full generation, they suffer from fixed sampling and expensive inference costs if without large-scale auxiliary training. Furthermore, an excessive pursuit of generic perceptual metrics often results in low fidelity. To address these issues, we present Dynamic Trajectory Initialization (DTI) paradigm for GFVSR, which reformulates GFVSR as an input-driven directional restoration. With a novel enhancement-and-injection conditioning mechanism for pretrained DiT backbone, fidelity of our model has been significantly improved without compromising perceptual quality. To dynamically set the starting sampling point, we propose a Discriminative Guide (DG) trained via objective Signal-to-Noise

* Corresponding author.

Ratio (SNR) alignment. With only minor model adaptation and fine-tuning, our method achieves a SOTA overall performance across diverse metrics and benchmarks. An analysis of relationship between actual comprehensive quality and common metrics is also conducted, which demonstrates the perception-distortion trade-off and that the LPIPS is the most convincing metric in our case. Code: `https://github.com/MediaX-SJTU/DTI`

Keywords: Generative face video super-resolution · Discriminative guidance · Dynamic Trajectory Initialization · Perception-distortion trade-off

1 Introduction

With its wide-ranging applications such as restoring speeches in news media, enhancing historical footage, and reconstructing forensic evidence, **Face Video Super-Resolution** (FVSR) presents a highly challenging ill-posed problem, for it must recover high-quality spatial-temporal information close to the ground truth from unknown real-world degradations. Formally, FVSR focus on reconstructing high-quality face videos (HQ) from degraded low-quality inputs (LQ). In these years, **Generative Face Video Super-Resolution** (GFVSR) present the best restoration results, but numerous issues remain.

Face Super-Resolution has long been an essentially important field in computer vision. Based on data types, we can classify it into: face image super-resolution (FISR) processing static images and FVSR treating videos with temporal dimension. Early FISR studies attempted to leverage various types of facial prior information, such as geometric priors [7, 49], reference priors [19, 20] and generative priors [3,31]. Despite the reconstruction of majority low-frequency information in one-step mapping, they struggle to restore high-quality details due to the irreversibility of information loss, the focus on certain priors also reduces their robustness when applied to more general real-world scenarios. This direction can be referred to as **discriminative**. With the continuous advancement of large generative models, especially diffusion models, numerous high-performance generative FISR have been developed [11,33,40]. The powerful generative prior of diffusion models has significantly improved the visual quality of restored details. In contrast to extensively studied FISR, there are less research attention towards FVSR, yet it is an increasingly important field today as videos become a central component of streaming media and social networks. For FVSR, main works are still conventional discriminative [5, 12, 46], several works explored the generative approaches [31, 41], but the majority among them conduct multi-frame restoration using existing FISR models or directly adapting general Generative Video Super-Resolution (GVSR) models [6, 44].

The most powerful GFVSR methods undoubtedly surpass other discriminative approaches, yet it also inherits the shortcomings of diffusion-based Video Super-Resolution (VSR) techniques. Similar to advanced GVSR models [1, 37, 42,42,54], GFVSR models also necessitate a huge inference computation burden, heavy auxiliary training or distillation owing to characteristics of diffusion-based

multi-step sampling. Furthermore, the face-unique characteristics of structural clarity and well-defined subject areas have not been sufficiently exploited for deep mining of LQ information. Despite enhanced detail generation capabilities from pretrained models, fidelity has not been assured in the reckless pursuit of perceptual quality. This phenomenon can be formulated as distribution generation trap: emergence of unnatural distortions or repetitive visual artifacts, resulting in a lower consistency with ground truth (GT).

To address the aforementioned issues and to enhance GFVSR, in this work we propose a **Dynamic Trajectory Initialization** (DTI) paradigm for GFVSR. We argue that the GFVSR process should not be regarded as a purely generative task as in previous studies, but rather as a dynamic restoration process driven by the extent of information loss in the input: preserving existing GT information and repairing damaged information. In other words, this paradigm transforms GFVSR from a simple conditional generation into a directional restoration, which places greater emphasis on fidelity and consistency with facial information, as well as authentic perceptual quality rather than pursuing improvements on generic metrics.

Given that LQ is lossy information, we innovatively introduce visual feature extractors into the framework, for face video features are highly suitable for their application. By redesigning the conditioning methods of generative approaches along with adaptation of model architecture, we obtain a GFVSR model with largely enhanced fidelity and reduced training overhead (accelerated model convergence) while maintaining the theoretical consistency with flow-matching diffusion. By theoretically deriving new training paradigm of discriminative components through objective signal-to-noise ratio alignment, we realize an **interpretable and controllable dynamic-starting restoration framework**. An analysis of relationship between metrics and true comprehensive quality is also carried out after the evaluation.

We conclude our main contributions below:

- We present Dynamic Trajectory Initialization (DTI) paradigm for GFVSR, reformulating GFVSR from "full generation" to "input-driven restoration", successfully decoupling it from pure generation tasks.
- We introduce novel enhancement-and-injection method for model conditioning with extracted visual features and conditions parallelism. We propose a Discriminative Guide (DG) and design its interpretable training by objective supervision, which estimates the reasonable intermediate distribution of diffusion trajectory which LQ should belong to.
- Through only minor model adaptation and fine-tuning, our approach achieves state-of-the-art performance with significantly improved efficiency and fidelity. By analyzing phenomena observed during evaluation, we also discuss the relationship between metrics and actual quality, demonstrating the optimality of LPIPS in perception-distortion trade-off for this task.

2 Related Work

2.1 Face Super-Resolution

Face Super-Resolution aims to recover ground-truth HQ from their LQ counterparts, where LQ stems from the intractable degradations in the real-world. Formerly, the highly structural nature of human faces allows researchers to incorporate different types of prior information into restoration models or frameworks: geometry priors such as parsing maps or facial landmarks [5, 7], reference-based priors such as high-quality images [20, 21], and generative priors including codebook priors [12–14, 25, 46, 51, 52] such as pretrained StyleGAN model [32]. Most early methods tend to produce low visual quality and lack of universality. Recently, restoration methods based on advanced deep neural networks or large models have been developed [41], some adapt directly general VSR networks to face video datasets [44], some extend the image-based models or approaches to video tasks [12, 17].

2.2 Diffusion-based Video Generation Models

Diffusion models tends to generate clean data from Gaussian noise by gradually denoising noisy data. Based on score matching or flow matching, probabilistic or deterministic diffusion models represent the most powerful visual generation models available today. From the original UNet-based backbones (DDPM/LDM) [15, 28] to the current transformer-based backbones (DiT) [26], architectural advancements have enabled scalability, paving the way for large-scale video diffusion generative models. These models can learn high-quality generative patterns for video data, which possesses both spatial and temporal dimensions. In recent years, continuous efforts from the open-source community have steadily narrowed the gap between open-source and closed-source video generation models [18, 34, 47].

2.3 Real-World Video Super-Resolution

Early researches in general VSR primarily focused on predefined idealized degradations, their performance deteriorate significantly when handling unknown degradations in reality [4, 45, 48]. In recent years, the development of diffusion-based large generative models has ultimately enabled a leap in visual quality for high-resolution VSR. Their powerful generative pattern learning abilities allow fine-tuned models to produce acceptable restoration results for complex real-world degradations, but also inherit their shortcomings. Extensive inference cost, insufficient fidelity and unnatural artifacts remain critical challenges after simply standard fine-tuning [1, 37, 42]. Heavy auxiliary components training and model post-training using large-scale data [36, 54] deviate from the inherent difficulty of restoration problem, offering poor cost-effectiveness.

3 Reasoning and Design

3.1 Preliminaries: Diffusion Transformer as Flow-Matching Diffusion Model

Unlike original probabilistic diffusion models based on stochastic Langevin dynamics [30], flow-matching (FM) diffusion models simulate the generative process of video latents by deterministically transforms a simple prior noise distribution p_1 (e.g., $\mathcal{N}(0, I)$) into the complex data distribution p_0 via an Ordinary Differential Equation (ODE) [22].The trajectory of data samples is governed by a time-dependent vector field v_t, defined by the ODE $dx_t/dt = v_t(x_t)$, where $t \in [0, 1]$. Usually, Conditional Flow Matching objective with an Optimal Transport path [23] are used to enable efficient training, which corresponds to a linear interpolation between the high-quality data x_0 and the noise $x_1 \sim \mathcal{N}(0, I)$. Specifically, the intermediate state x_t at time t is constructed as:

$$x_t = (1 - t)x_0 + tx_1. \tag{1}$$

This linear path implies a constant target velocity field for a given data-noise pair. Differentiating the path with respect to t yields the ground-truth velocity $u_t(x|x_0, x_1) = x_1 - x_0$. Consequently, a neural velocity estimator v_θ conditioned on time t and other conditions c is trained to regress this target direction. The training objective is to minimize the expected mean squared error between the predicted velocity and the ground-truth vector pointing from data to noise:

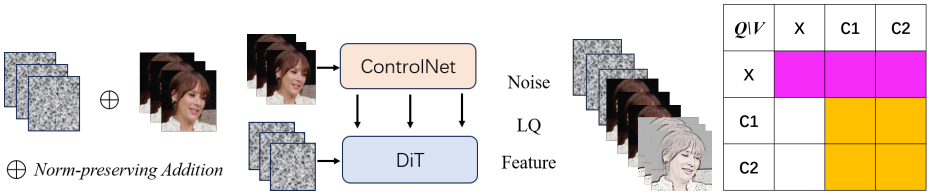
$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,x_0,x_1}\left[\omega(t)\|v_\theta(x_t, t, c) - (x_1 - x_0)\|^2\right], \tag{2}$$

where v_θ is predicted velocity, $\omega(t)$ represents a time-dependent weight. During inference, output is sampled by solving the ODE from noise x_1 to $t = 0$, guided by velocity field v_θ and conditions c using a numerical solver (e.g., Euler method).

Diffusion Transformer (DiT) [26] is a diffusion model design with transformer backbone. Its conditional controllability during generation primarily stems from the attention blocks within the transformer structure. This mechanism enables the sampling entity to compute weighted attention over conditions, the condition information is then injected as residual increments.

3.2 Conditional Generation as Restoration

To transform a DiT-based diffusion generation model into a VSR model, past studies have already some practices: by modifying the starting point of denoising process, it becomes a norm-preserving linear combination (Fig. 2a) of pure Gaussian noise and LQ [54]; by adopting ControlNet architecture (Fig. 2b), learnable control networks embed LQ and inject information into diffusion backbone as an additional factor to timestep condition [42]; adopting both noisy LQ as input and ControlNet-style injection mechanism [1]; concatenate LQ and noise then passing it into MLP layers after flattening [36, 37]. They each have their own shortcomings: under standard sampling, linear combination fails to align with diffusion



(a) Addition **(b)** ControlNet **(c)** Our Conditioning **(d)** Attention Matrix

Fig. 2: (a)(b) some conventional conditioning methods; (c) our novel condition injection method; (d) represents the real calculated attention parts in adapted DiT, *orange* ones are only calculated once in a forward.

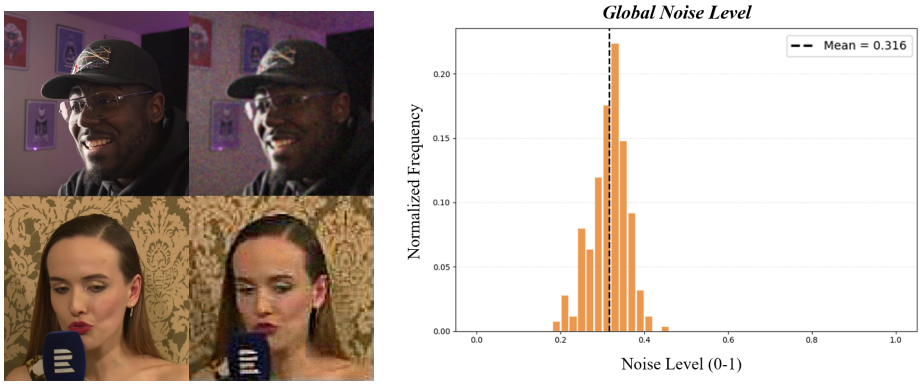
model due to input shifting; ControlNet requires heavy auxiliary components and training; concatenation-flattening-MLP only performs channel-level information exchange, resulting in relatively low quality. Besides, all previous works directly employ LQ without considering the possibility of extracting information from it to enhance the condition.

The primary consideration is that LQ, as our sole information source, is inherently lossy data. Much useful information lies buried beneath degradation such as blurring or noise, making it insufficient as a direct condition. Naturally, to enhance this condition, **visual feature extraction models** occur to our mind: human faces are highly suitable for application of such networks, for they possess strong visual structural characteristics such as distinct main areas and clear subjects. Theoretically, as extractors embed patches, extracted features not only are enhanced visual information, but also can be seen partially purified.

Secondly, to efficiently and succinctly inject conditions while maintaining alignment with Diffusion theory, we propose connecting pure noise (the sampling subject) end-to-head with condition tokens, treating them as **one input sequence** to DiT (Fig. 2c), and achieving conditional generation through attention calculations. Moreover, based on the fact that conditions don't need to be influenced by noise, bidirectional attention inside conditions and unidirectional attention between noise and conditions can realize sufficient conditioning and reduce the computational burden (Fig. 2d), for the complexity of bidirectional attention increases quadratically with the length of sequence.

3.3 Objectively Dynamic Initialization by SNR Alignment

As the examples show (Fig. 3a), previous related work [40] has observed that LQ retains much low-frequency information (overall visual structure, regional colors) from GT, with the primary information loss occurring in high-frequency domain (texture features, edge sharpness). This observation supports that mapping neural networks (NN) in discriminative methods suffice to an efficient coarse-grained refinement in low-frequency domain with acceptable quality, while struggling to restore high-frequency details. This insight inspires us to employ a discriminative approach to guide generative restoration, ensuring proximity to low-frequency ground-truth information and avoiding unnecessary generative sampling.



(a) Low-frequency Info Consistency **(b)** SNR-aligned Noise Level

Fig. 3: (a) represents that LQ preserves perceptually more low-frequency information than high-frequency; (b) visualizes the SNR-aligned timesteps corresponding to real-world degradations on a test subset.

Inspired by ideas of related studies [43] in vision field, the **signal-to-noise ratio** (SNR) is used in this work to objectively estimate the relationship between a diffusion timestep which represents the injection proportion of Gaussian noise and the global degradation level of real-world randomly degraded samples. Here, we extend theoretically the proposal:

As mentioned in Sec. 3.1, noisy latent z_t at timestep t in diffusion process is a linear combination of Gaussian noise ϵ and the ground-truth HQ latent z_H. Similarly, LQ are obtained by applying real-world degradation to HQ, where z_L can be regarded as combination of residual difference $z_L - z_H$ and z_H. Their SNR can be expressed as:

$$SNR(z_t) = \frac{(1-t)^2 \mathbb{E}\left[z_H^2\right]}{t^2 \mathbb{E}\left[\epsilon^2\right]}, \tag{3}$$

$$SNR(z_L) = \frac{\mathbb{E}\left[z_H^2\right]}{\mathbb{E}\left[(z_L - z_H)^2\right]}, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\mathbb{E}\left[\epsilon^2\right] = 1$. Defining $RMSE(x, y) = \sqrt{\mathbb{E}\left[(x-y)^2\right]}$ as square root of mean square error, by linking Eq. (3) and Eq. (4), we can derive the following equation:

$$t = \frac{RMSE(z_L, z_H)}{1 + RMSE(z_L, z_H)}. \tag{5}$$

In Fig. 3b, the noise level distribution (equivalent to timestep in diffusion process) calculated using Eq. (5) effectively demonstrates the preservation of low-frequency information in LQ, further validating the significance of discriminative component design. The global noise level of real-world degradations are SNR-equivalently to the 20%-45% of Gaussian diffusion process.

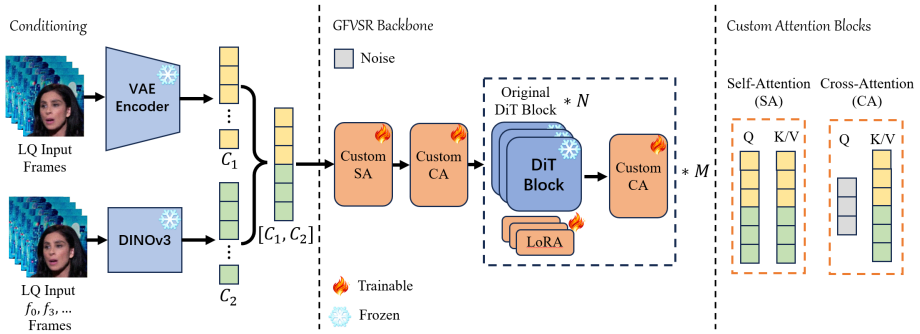


Fig. 4: Model Structure & Pipeline

4 Methodology

4.1 Conditioning and Architecture of GFVSR Model

Human faces exhibit a clearly defined subject, an overall structured composition and pronounced edges. Given the need to extract as much useful visual information as possible from LQ, **fine-grained visual feature extractors** are our preferred choice. Compared to semantic-aligned contrastive learning models like CLIP [27], the DINO family of self-supervised pure vision models clearly better fits our requirements [24]. Since the VAE encoder also downsamples the input along the temporal dimension, to naturally align with the sequence shape, we apply the feature extractor to the intermediate frames at the downsampling intervals, serving as a pillar regularization for the overall trend (Fig. 4).

In addition to Sec. 3.2, to further minimize modifications to the backbone, we let subsequence (LQ,Extracted-Feature) do only *one* self-attention computation at the beginning of a model forward. As shown in Fig. 4, formally, input sequence is (X, C1, C2), where X represents pure noise tokens, C1 denotes LQ tokens, and C2 signifies extracted feature tokens. One block is added at top of DiT to realize self-attention of (C1, C2) and first cross-attention of X to (C1, C2). A cross-attention block of X to (C1, C2) is also added after every several original DiT blocks, which is equivalent to a unidirectional attention block. It should be noted here that testing confirms our method doesn't make observable change in model's wall-clock time per forward.

4.2 Dynamic Initialization

Our conditioning method ensures the consistency with Diffusion theory, yet the sampling process is expensive. Previous studies on diffusion models indicate that during denoising, different timesteps primarily target distinct domains in information: high-noise steps construct low-frequency information first, while low-noise steps focus on generating high-frequency details [10,15,28]. As LQ already closely matches GT in low-frequency domain, it should not require a complete

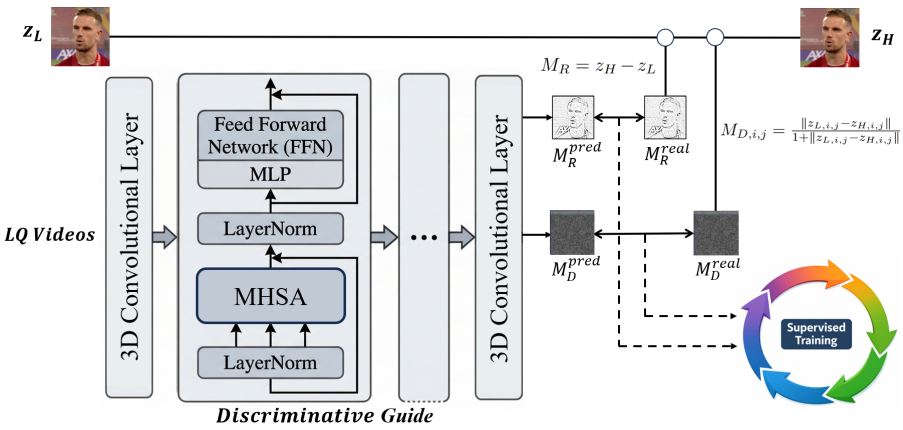


Fig. 5: Supervised Training for Discriminative Guide

generation process starting from pure noise. But it can't be treated directly as an intermediate state of noisy latent neither, as LQ is in untractable real-world degradation distribution, not the specific parameterized distribution (Gaussian distribution) learned by diffusion model. So, if we can transfer LQ to a reasonable nearest point in the probability flow field, the generation process can be reformulated to a restoration process with dynamic starting point.

Reasoning in Sec. 3.2 provides us with the validation and strategy for a supervised training of **discriminative guide** (DG). (Fig. 5) DG is a lightweight mapping NN to estimate two latents with both the same shape as target latent: one for global noise-level, another for efficient low-frequency refinement. The reason of full-dimension design for first latent is to alleviate the error of global prediction. The first latent is trained to approximate the matrix M_D where $M_{D,i,j} = \frac{\|z_{L,i,j} - z_{H,i,j}\|}{1+\|z_{L,i,j} - z_{H,i,j}\|}$, which represents the local information-loss level; the second to approximate the element-wise residual between HQ and LQ $M_R = z_H - z_L$. As the lightweight DG has limited capability, the predicted M_R is with high error which can only serves as a low-frequency coarse-grained refinement. This objective supervision makes our DG training more interpretable, which is never concerned in any previous work for restoration guidance.

The dynamic initialization is realized in the following framework:

The predicted starting timestep is calculated by M_D:

$$t_{pred} = \frac{RMSE(z_L, z_H)}{1 + RMSE(z_L, z_H)} = \frac{\sqrt{\mathbb{E}\left[(\frac{M_D}{1-M_D})^2\right]}}{1 + \sqrt{\mathbb{E}\left[(\frac{M_D}{1-M_D})^2\right]}}. \tag{6}$$

The noise proportion in the $z_{start}(t_{start})$ determines a trade-off: relatively, more sampling steps with noisier starting point yield higher perceptual quality but

lower fidelity, and vice versa. Therefore, by setting

$$t_{start} = (1 - \lambda)t_{pred} + \lambda t_{max}, \tag{7}$$

where $t_{max} = 1, \lambda \in [0, 1]$, we obtain a hyperparameter that can flexibly adjust the trade-off: 1 to perceptual quality, 0 to fidelity. We calculate the starting latent before DiT sampling as:

$$z_{start} = (1 - t_{start})z_{anchor} + t_{start}\epsilon, \tag{8}$$

where $z_{anchor} = z_L + M_R$ and $\epsilon \sim \mathcal{N}(0, I)$.

In this manner, on one hand we employ discriminative prediction to efficiently influence the repair starting point of generative sampling, bringing it closer to HQ while preserving more low-frequency information that does not require much refinement; on the other hand, by initiating from a reasonable timestep (SNR aligned), we leverage DiT's robust generative prior and multi-step denoising characteristics to mitigate the adverse effects of discriminative prediction errors on the final outcome.

5 Experiments

5.1 Datasets and Metrics

Training Data All our training are implemented on the train split of **VFHQ** dataset [44], which includes 16k high quality face clips as GT. During training, the LQ video data are synthesized by the standard Real-ESRGAN degradation pipeline [4,38] with lightweight random temporal corrupt: stochastic multi-frame pooling and frame replication to simulate frame drops and temporal stuttering, enabling a more realistic approximation of real-world video degradation.

Test Data Evaluation of our GFVSR model is conducted on three benchmarks: the official **VFHQ** test dataset [44], **CelebV-HQ** [53] for fine texture recovery and **VoxCeleb2** [8] for severe in-the-wild degradations. As our DG is only trained from scratch on VFHQ dataset with a small amount of iterations (15k), we conducted its evaluation (ours w/ DG) on corresponding test dataset.

Metrics For datasets with GT, we adopt widely used **PSNR**, **SSIM**, and **LPIPS** [50] as general full-reference metrics to evaluate the fidelity; **IDS** [9] quantifying the biometric consistency of the reconstructed faces,**LMD** [39] evaluating the alignment accuracy of facial landmarks and **TLME** [46] measuring the consistency of facial dynamics are evaluated as facial full-reference metrics. We adopt also generic no-reference metrics on all test datasets: **MUSIQ** [16] for multi-scale aesthetic perception, and **CLIP-IQA** [35] for high-level semantic quality.

Table 1: Quantitative comparison with state-of-the-art methods on three face video benchmarks. The best and second-best performances are marked in **red** and **blue** respectively. Here ours is without DG.

Datasets	Metrics	PGTFormer	SVFR	Vivid-VR	FlashVSR	SeedVR2	Ours
VFHQ	PSNR ↑	**25.63**	25.54	19.82	19.57	19.80	**26.35**
	SSIM ↑	0.66	0.72	0.72	0.71	**0.73**	**0.74**
	LPIPS ↓	0.30	0.22	0.21	**0.18**	0.21	**0.17**
	MUSIQ ↑	64.22	65.60	71.21	**75.33**	60.01	**72.54**
	CLIP-IQA ↑	**0.56**	0.50	0.55	**0.71**	0.48	**0.56**
	IDS ↑	0.79	**0.87**	0.83	0.86	0.85	**0.87**
	LMD ↓	6.055	4.71	4.64	6.10	**4.59**	**4.20**
	TLME ↓	5.72	4.11	**3.98**	4.29	4.19	**3.72**
CelebV-HQ	PSNR ↑	22.75	**24.95**	23.91	24.59	24.66	**25.52**
	SSIM ↑	0.47	**0.71**	0.64	0.68	0.70	**0.75**
	LPIPS ↓	0.55	**0.28**	0.40	0.32	0.34	**0.22**
	MUSIQ ↑	47.1	57.73	63.81	**69.69**	48.72	**66.26**
	CLIP-IQA ↑	0.44	0.47	**0.54**	**0.70**	0.44	**0.54**
	IDS ↑	0.55	**0.77**	0.66	**0.77**	0.64	**0.77**
	LMD ↓	45.05	**7.54**	15.30	9.44	16.04	**5.26**
	TLME ↓	29.16	**5.34**	8.59	6.34	11.39	**4.01**
VoxCeleb2	MUSIQ ↑	58.71	63.02	**73.87**	**72.57**	56.62	68.41
	CLIP-IQA ↑	0.49	0.52	**0.67**	**0.75**	0.45	**0.67**

5.2 Implementation

The backbone of our GFVSR model is the official Wan2.1-1.3B T2V DiT model [34]. The selected visual feature extractor is DINOv3 [29], embedding the first and every third among four frames in the rest, aligned with temporal downsampling of Wan VAE encoder. All experiments are implemented in 512 * 512 spatial resolution, which is near the officially recommended resolution for backbone and sufficient for demonstration of any method's practicality. Existing DiT blocks are fine-tuned using LoRA. The DG is implemented as a lightweight Vision Transformer (ViT) after comparing several architectures' performance.

5.3 Comparison with State-of-the-Art Methods

We compare first our method against state-of-the-art (SOTA) FVSR methods including **PGTFormer** [46] and **SVFR** [41], then against SOTA general VSR methods to more universally prove the performance: **Vivid-VR** [1], **FlashVSR** [54] and **SeedVR2** [36].

Performance As shown in Table 1, our model shows an overall SOTA performance. It achieves the lowest LPIPS on all datasets with GT, with an excellent average lower than 0.2. It has the best performance on all general and facial

Table 2: Quantitative result on VFHQ benchmark and the NFE counting. The best and second-best performances are marked in **red** and **blue**, respectively. λ is set to 0.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑	CLIP-IQA ↑	IDS ↑	LMD ↓	TLME ↓	NFE
SVFR	25.54	0.72	0.22	65.60	0.50	**0.87**	4.71	4.11	80
Vivid-VR	19.82	0.72	0.21	**71.21**	**0.55**	0.83	4.64	3.98	50
Ours w/o DG	**26.35**	**0.74**	**0.17**	**71.54**	**0.56**	**0.87**	**4.20**	**3.72**	50
Ours w/ DG	**26.85**	**0.76**	**0.17**	65.17	0.48	0.86	**4.03**	**3.69**	12

fidelity metrics, which demonstrates its capabilities in structural consistency, spatio-temporal consistency, and facial feature reconstruction. It remains competitive on no-reference metrics. Given that our approach is based on a small amount of fine-tuning (only 20k iterations), the results undoubtedly validate our conditioning paradigm. Note that FlashVSR attains the best performance on generic perceptual metrics (no-reference), which is reasonable since it has conducted the largest-scale model post-training and auxiliary components' training on a closed-source 160k image-video dataset, generating distribution that most closely resembles real-world data distribution.

Efficiency Among these methods, PGTFormer is originally discriminative, SeedVR2 and Flash VSR is one-step generative after their large-scale post-training, while SVFR, Vivid-VR and our method are multi-step generative. As shown in Table 2, discriminative guidance mechanism largely reduces the NFE (Number of Function Evaluations) of our method by dynamically set the starting point, with an average reduction of 76%. Meanwhile, this framework significantly improved full-reference metrics while slightly declining no-reference metrics, thereby validating our reasoning about that discriminative guidance shifts the perception-distortion trade-off towards fidelity.

6 Discussions

6.1 Effect of Extracted Face Visual Information

We conduct an ablation experiment under identical conditioning paradigm for adapted DiT model. One is trained with the LQ latent as condition (single-condition), the other with also face visual information encoded by DINOv3 (double-condition). Loss evolution (Fig. 6) and metrics evaluation (Tab. 3) reveal that injection of extracted information condition not only elevates the ultimate performance ceiling but also accelerates model convergence.

6.2 Discussion on Metrics

Past works in computer vision have already proved that there is a trade-off between fidelity and perceptual quality: the famous **perception-distortion** trade-off [2]. Based on this theory, performance evaluation should use two types of metrics: full-reference metrics for the proximity to GT; no-reference metrics for the

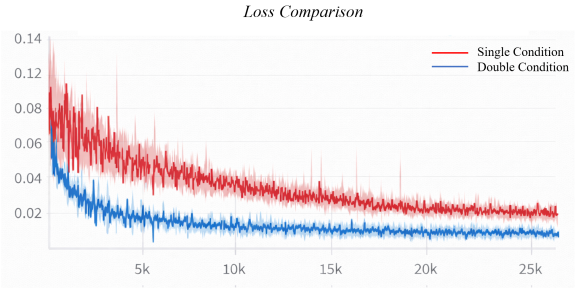


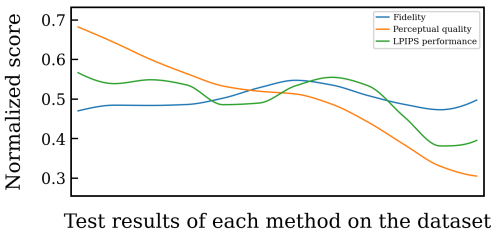
Fig. 6: Loss comparison

	Single	Double
PSNR ↑	22.15	26.35
SSIM ↑	0.61	0.74
LPIPS ↓	0.28	0.17
MUSIQ ↑	69.20	71.54
CLIP-IQA ↑	0.56	0.56

Table 3: Performance

alignment with the real-world data distribution and the visual quality. During evaluation, we identify certain metrics exhibiting unreasonably inflated values in both types, consistent with the trade-off theory: restorations with high no-reference metrics are riddled with repetitive artifacts which is a consequence of overemphasizing realistic details, while restorations with high full-reference metrics appear blurry which reflects mediocrity due to excessive pursuit of global fidelity. Some qualitative examples are shown in Fig. 7a, we confirm that this phenomenon is fairly common across all test data and effectively avoided by our method, which we attribute to the regularization imposed by visual feature condition and discriminative guidance. Among all the metrics, LPIPS demonstrates the most balanced performance in this trade-off and best aligned with human visual perception, because it is only better when both aspects are relatively good (Fig. 7b).

(a) Qualitative Exhibition



(b) Metrics Relationship

Fig. 7: Metrics Discussion. (a): At the *top*, higher MUSIQ sample has evident visual artifacts; at the *bottom*, higher PSNR sample is blurry. (b): The relative normalized score trend among LPIPS, fidelity metrics and perceptual metrics.

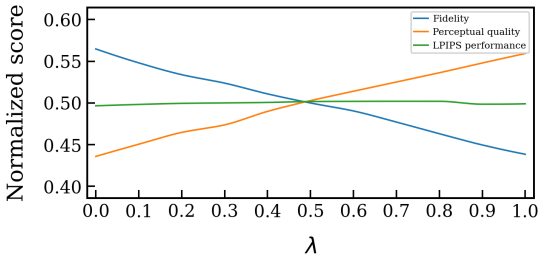


Fig. 8: Hyperparameter λ for Perception-Distortion Trade-off

6.3 Controllable Perception-Distortion Trade-off

In Fig. 8, normalized scores of metric groups show that the hyperparameter λ successfully regulates the perception-distortion trade-off, providing a controllable external interface that can be manipulated as needed, thereby validating our prior deduction about dynamic initialization.

7 Conclusion

In this work, we propose DTI, a paradigm allowing dynamic starting for GFVSR, which makes it conform to inherent logic of the restoration task. Our method is built upon a pretrained DiT model with only minor model adaptation and fine-tuning. Dynamic starting is realized by discriminative guidance, with model performance improvement attributed to novel method for condition enhancing and injecting. The proposed DG simultaneously performs coarse-grained low-frequency information anchoring and global noise level prediction, enabling DTI, dynamic NFE reduction and a little shift towards fidelity in perception-distortion trade-off. Extensive experiments on diverse benchmarks demonstrate SOTA performance of our method. The perception-distortion trade-off among metrics is also observed, which illustrates the unreliability of single type of metrics and the comprehensive credibility of LPIPS.

For DG, one limitation lies in the fact that from-scratch base restricts the generalizability of lightweight training on one dataset. Although we selected the current design after ablation study among LoRA fine-tuning and full-parameter fine-tuning of VAE Encoder, from-scratch training of UNet-like (Encoder-like) NNs, we lack investigation and fine-tuning experiments on pretrained ViT-like models. Leveraging pretrained priors may further improve general performance and training efficiency of DG, which we leave for future work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62271308, 62571322), STCSM (24ZR1432000, 22DZ2229005), 111 plan (BP0719010), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

1. Bai, H., Chen, X., Yang, C., He, Z., Deng, S., Chen, Y.: Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. In: International Conference on Learning Representations (ICLR) (2026)
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6228-6237, 2018 (2018)
3. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2021)
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
5. Chen, C., Li, X., Lingbo, Y., Lin, X., Zhang, L., Wong, K.Y.K.: Progressive semantic-aware style transformation for blind face restoration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
6. Chen, X., Tan, J., Wang, T., Zhang, K., Luo, W., Cao, X.: Towards real-world blind face restoration with generative diffusion prior (2023)
7. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
8. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. arXiv preprint arXiv:2403.03206 (2024)
11. Fang, Y., Chen, Y., Yin, S., Hu, Q., Yao, J., Zhang, Y., Zhang, X., Wang, Y.: One-step diffusion transformer for controllable real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23440–23450 (2026)
12. Feng, R., Li, C., Loy, C.C.: Kalman-inspired feature propagation for video face super-resolution. In: European Conference on Computer Vision (ECCV) (2024)
13. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: CVPR (2020)
14. Gu, Y., Wang, X., Xie, L., Dong, C., Li, G., Shan, Y., Cheng, M.M.: Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In: ECCV (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
16. 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)
17. Kelvin C.K. Chan, Shangchen Zhou, X.X., Loy, C.C.: Basicvsr++: Improving video super resolution with enhanced propagation and alignment". arXiv preprint arXiv:2104.13371 (2021)

18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
19. Li, X., Chen, C., Zhou, S., Lin, X., Zuo, W., Zhang, L.: Blind face restoration via deep multi-scale component dictionaries. In: ECCV (2020)
20. 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: CVPR (2020)
21. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: The European Conference on Computer Vision (ECCV) (September 2018)
22. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
23. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
25. Pan, X., Zhan, X., Dai, B., Lin, D., Loy, C.C., Luo, P.: Exploiting deep generative prior for versatile image restoration and manipulation. In: European Conference on Computer Vision (ECCV) (2020)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
29. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025)
30. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
31. Tao Yang, Peiran Ren, X.X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
32. Tero Karras, Samuli Laine, T.A.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
33. Varma, K., Reddy, G.S., Subramanyam, N.: Face image super resolution using a generative adversarial network. In: 2021 Smart Technologies, Communication and Robotics (STCR). pp. 1–8 (2021)
34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W.,

Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
36. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., Xiao, X., Loy, C.C., Jiang, L.: Seedvr2: One-step video restoration via diffusion adversarial post-training. In: ICLR (2026)
37. Wang, J., Lin, Z., Wei, M., Zhao, Y., Yang, C., Loy, C.C., Jiang, L.: Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In: CVPR (2025)
38. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW)
39. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
40. Wang, Z., Zhang, X., Zhang, Z., Zheng, H., Zhou, M., Zhang, Y., Wang, Y.: Dr2: Diffusion-based robust degradation remover for blind face restoration. arXiv preprint arXiv:2303.06885 (2023)
41. 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 (2025)
42. Weisong Zhao1, Jingkai Zhou, X.Z.W.C.X.Y.Z.Z.L.F.W.: Realisvsr: Detail-enhanced diffusion for real-world 4k video super-resolution. arXiv:2507.19138 (2025)
43. Wu, Z., Sun, Z., Zhou, T., Fu, B., Cong, J., Dong, Y., Zhang, H., Tang, X., Chen, M., Wei, X.: Omgsr: You only need one mid-timestep guidance for real-world image super-resolution (2025)
44. Xie, L., Wang, X., Zhang, H., Dong, C., Shan, Y.: Vfhq: A high-quality dataset and benchmark for video face super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022)
45. Xintao Wang, Kelvin CK Chan, K.Y.C.D., Loy, C.C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (2019)
46. 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)
47. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
48. Younghyun Jo, Seoung Wug Oh, J.K., Kim, S.J.: Deep video super-resolution network using dy namic upsampling filters without explicit motion compensa tion. CVPR (2018)
49. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
51. Zhang, Z., Gao, X., Wang, Z., Hu, Q., Zhang, X.: Td-bfr: Truncated diffusion model for efficient blind face restoration. 2025 IEEE International Conference on

Multimedia and Expo (ICME) pp. 1–6 (2025), https://api.semanticscholar.org/CorpusID:277322774
52. Zhou, S., Chan, K.C., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS (2022)
53. 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)
54. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)