Improving Temporal Consistency and Fidelity at Inference-time in Perceptual Video Restoration by Zero-shot Image-based Diffusion Models

Nasrin Rahimi, A. Murat Tekalp, *Life Fellow, IEEE*

***Abstract*—Diffusion models have emerged as powerful priors for single-image restoration, but their application to zero-shot video restoration suffers from temporal inconsistencies due to the stochastic nature of sampling and complexity of incorporating explicit temporal modeling. In this work, we address the challenge of improving temporal coherence in video restoration using zero-shot image-based diffusion models without retraining or modifying their architecture. We propose two complementary inference-time strategies: (1) Perceptual Straightening Guidance (PSG) based on the neuroscience-inspired perceptual straightening hypothesis, which steers the diffusion denoising process towards smoother temporal evolution by incorporating a curvature penalty in a perceptual space to improve temporal perceptual scores, such as Fréchet Video Distance (FVD) and perceptual straightness and (2) Multi-Path Ensemble Sampling (MPES), which aims reducing stochastic variation by ensembling multiple diffusion trajectories to improve fidelity (distortion) scores, such as PSNR and SSIM, without sacrificing sharpness. Together, these training-free techniques provide a practical path toward temporally stable high-fidelity perceptual video restoration using large pretrained diffusion models. We performed extensive experiments over multiple datasets and degradation types, systematically evaluating each strategy to understand their strengths and limitations. Our results show while PSG enhances temporal naturalness particularly in case of temporal blur, MPES consistently improves fidelity and spatio-temporal perception–distortion trade-off across all tasks.**

***Index Terms*—Temporal consistency, Perceptual straightening hypothesis, Zero-shot video restoration, Diffusion models, Multi-path ensembling**

I. Introduction

THE advent of deep learning has revolutionized the field of image and video restoration. Generative models have demonstrated impressive ability to produce perceptually pleasing single-image restoration results. In particular, diffusion models have emerged as state-of-the-art generative priors for image generation and restoration. Their strong generative capacity and stable training makes them attractive for supervised or unsupervised solutions of inverse problems such as image and video super-resolution, deblurring, and inpainting.

Adapting large pretrained image diffusion models to video restoration is particularly appealing, as paired training data for video degradations is often scarce or unavailable. Even when video data is available, training large high-resolution video generative models is prohibitively expensive. Therefore, a line of zero-shot methods, which apply text-to-image diffusion prior to multiple video restoration tasks without task-specific retraining, have been developed. These task-agnostic, zero-shot large generative methods are valuable for real-world video restoration, where task-specific models may generalize poorly. However, applying image-based diffusion models to zero-shot video restoration pose several challenges:

First, zero-shot image diffusion models cannot capture the spatio-temporal dynamics of degraded videos. Therefore, the stochastic nature of sampling leads to independent hallucinations in consecutive frames, which results in texture flicker or jitter, and inconsistent motion patterns. A central challenge in video restoration using image-based restoration models is ensuring temporal consistency across frames while maintaining sharpness of each frame. The apparent trade-off between sharp spatial detail and coherent temporal dynamics has motivated a growing line of research focused on addressing temporal consistency as a core component of video restoration.

Second, most methods that enforce temporal coherence in zero-shot methods, require architectural modifications or training for added recurrent modules, temporal attention blocks, or motion guidance which are costly. Inference-time strategies, on the other hand, are lightweight and task-agnostic and there is still significant room for improving their ability to directly enforce temporal consistency and naturalness. Overall, the key challenge is how to improve temporal coherence while preserving the high spatial quality of pretrained diffusion models, without retraining or sacrificing generality across different degradation types.

To address these challenges, we propose two complementary perspectives: First, the neuroscience-inspired perceptual straightening hypothesis (PSH) [1] suggests that natural video sequences which follow curved trajectories in pixel space become straighter when processed by the hierarchical stages of the human visual system. This hypothesis provides a potential proxy for temporal naturalness. Second, the stochastic nature of diffusion sampling causes variations in the outputs across different runs, even with the same conditioning. This motivates variance-reduction strategies such as ensembling, which improves fidelity without retraining. In this work, we leverage VISION-XL [2] as the baseline, which is a zero-shot video restoration framework based on text-to-image latent diffusion. VISION-XL applies measurement consistency during sampling; however, the temporal consistency mechanisms used are heuristic leaving room for more principled and broadly

Authors are with the Department of Electrical and Electronics Engineering and KUIS AI Center, Koç University, 34450 Istanbul, Turkey. N.R. was partly supported by a KUIS AI Center Fellowship. A.M.T. acknowledges support from the Turkish Academy of Sciences (TUBA). Manuscript submitted October 20, 2025.

applicable inference-time strategies. Our methods can be integrated into VISION-XL or other similar zero-shot pipelines in a coherent manner to enhance temporal stability during inference, without modifying the model architecture.

Contributions. We propose two complementary inference-time optimization strategies that explicitly address temporal inconsistency in zero-shot diffusion video restoration:

- **Perceptual Straightening Guidance (PSG):** Inspired by the neuroscience-based perceptual straightening hypothesis, we introduce a diffusion denoising guidance mechanism that penalizes sequences with large curvature in a perceptual space and encourages frames to evolve more linearly in this embedding space. Evaluation results show that PSG improves perceptual straightness and other temporal metrics, particularly under degradations involving temporal blur and reduces frame-to-frame jitter.
- **Multi-Path Ensemble Sampling (MPES):** We propose ensembling multiple diffusion trajectories to reduce the variance of posterior sampling. The proposed MPES offers a tradeoff between sampling from the posterior and MMSE estimation in a controlled manner. We evaluate pixel-space vs. latent fusion strategies. Our results show pixel-space fusion consistently improves both perceptual and distortion metrics across spatial and temporal dimensions -albeit with higher runtime- and significantly improves spatio-temporal perception-distortion trade-off.

Together, these contributions demonstrate that lightweight, training-free techniques can substantially improve temporal consistency in zero-shot diffusion video restoration—without retraining or architectural changes.

II. Related Work

A. Video Restoration and Temporal Consistency

Video restoration approaches exploit temporal correlations between frames to improve quality beyond single-image restoration. Early classical methods employed explicit motion estimation and multi-frame modeling [3, 4]. Temporal regularization was also exploited to penalize deviations between consecutive reconstructed frames [5] to reduce flicker. Alternative strategies modeled video as a 3D volume using spatio-temporal priors [6]. Another approach was using recursive propagation across time [7], though it was prone to error accumulation.

The advent of deep learning has enabled learning rich spatio-temporal representations from training data that lead to significantly enhanced quality. A diverse set of strategies have been employed to achieve temporal coherence. Alignment-based methods such as EDVR [8], BasicVSR [9], and BasicVSR++ [10] explicitly align features through deformable convolutions and optical flow, while recurrent methods like FRVSR [11] encourage temporally consistent results using the previously inferred estimate to restore the subsequent frame. Attention mechanisms including VRT [12], RVRT [13] model long-range temporal dependencies through self-attention and space-time attention.

Generative models, such as GANs, allow optimization of perceptual realism rather than focusing only on pixel-wise fidelity. A GAN-based video super-resolution framework incorporating edge enhancement and motion-compensated multi-frame fusion was proposed in [14]. TecoGAN [15] pioneered temporally consistent video super-resolution using a spatio-temporal discriminator and ping-pong consistency loss. [16] proposes a perceptual VSR model with two discriminators, where a flow discriminator encourages naturalness of motion.

Denoising diffusion probabilistic models (DDPMs) [17] and variants have recently become state-of-the-art generative approaches for image and video generation. These models have been adapted to specific supervised image and video restoration tasks by conditioning the denoising process on paired (original, degraded) observations. Compared to the GANs, diffusion-based frameworks offer greater stability and typically generate higher-quality results, especially under heavy degradations. Upscale-A-Video [18], a text-guided VSR diffusion model preserves temporal consistency through incorporating temporal layers inside the UNet and VAE-Decoder and global latent propagation mechanism to ensure temporal coherence. DiffVSR [19] exhibits stable performance against complex degradations through multi-scale temporal alignment. [20] generate fine-grained details while preserving temporal stability via a trained temporal module. FLAIR [21] provides a conditional framework for face video restoration with facial-specific temporal dynamics. More recent video diffusion methods introduce temporal priors through architectural modifications—such as 3D convolutions or temporal attention [22], or motion-aware modules guided by optical flow [23].

B. Diffusion-Based General Inverse-Problem Solvers (DIS)

Diffusion-based inverse problem solvers (DIS) exploit pre-trained diffusion models to capture the the prior image distribution $p_\theta(x)$. Given a degradation model, DIS methods approximate the posterior $p_\theta(x|y)$, where y is a degraded observation. This is achieved by combining the data prior $p_\theta(x)$ with the likelihood $p(y|x)$, enabling iterative sampling guided by the measurement y. In the video domain, diffusion models face additional challenges due to the temporal dimension.

Naïvely applying image-trained diffusion models frame-by-frame often leads to flicker and motion inconsistency. To address this, DIS-based video restoration approaches [24, 25] employ explicit strategies to control temporal consistencies. DIS-based approaches have been successfully applied to a wide range of tasks, including deblurring, super-resolution, inpainting, colorization, and compressed sensing. ZVRD [26] uses a pre-trained single image diffusion model for video in a zero-shot setting by mechanisms like short/long-range temporal attention, consistency guidance, and spatio-temporal noise coordination to achieve temporally coherent restoration. DiffIR2VR-Zero [27] performs zero-shot temporally consistent video restoration by using any pre-trained image diffusion model. It combines hierarchical latent warping with hybrid token merging and maintains temporal consistency while preserving restoration quality for different restoration tasks. VISION-XL [2] a Zero-shot video restoration model utilizes latent diffusion model with pseudo-batch consistent

sampling and data consistency refinement to improve temporal coherence and fidelity.

Despite recent progress, preserving temporal coherence in zero-shot video restoration remains challenging. Many existing methods rely on retraining or architectural changes, while general inference-time strategies for temporal consistency are still limited. We address this gap by proposing inference-time (training-free) methods to improve temporal consistency and fidelity without modifying the underlying diffusion model.

C. Perceptual Straightening Hypothesis (PSH)

The perceptual straightening hypothesis (PSH), introduced by neuroscientist Hénaff, et al. [1], states that the human visual system transforms natural video sequences to a perceptual feature space, where motion follows straighter trajectories compared to the pixel-intensity domain. In contrast, unnatural or inconsistent visual sequences tend to exhibit greater curvature in this perceptual feature space.

The PSH has inspired several works in video processing. Kancharla and Channappayya [28] used PSH in a video frame prediction model and showed improvement in visual quality by minimizing curvature of the trajectory in the perceptual space. They also proposed a no-reference quality measure based on PSH to evaluate naturalness of motion in user-generated videos [29]. In our earlier work [16], we employed perceptual straightness as an evaluation metric for temporal naturalness in video super-resolution (VSR). More recently, Internò et al. [30] applied the PSH to detect AI-generated videos, showing that unnatural temporal dynamics introduced by generative models result in higher perceptual curvature, which can serve as a discriminative signal.

In this work, we employ perceptual straightness as an optimization criterion at inference-time to enforce perceptual straightness of videos generated by zero-shot VISION-XL pipeline. We show the proposed method can avoid unnatural or inconsistent restored sequences, straightening their perceptual trajectory without explicit motion supervision.

D. Ensembling in Generative Models

Due to the stochastic nature of diffusion models, each time the diffusion model processes the same input, the noisy latent follows a slightly different trajectory because of the randomness in the denoising process. This trajectory variability can cause small differences in fine-grained image details, as well as inconsistencies in temporal dynamics of the video. The iterative nature of diffusion sampling can make the situation worse, where small errors accumulate at each denoising step.

Ensembling methods combine multiple models or predictions to reduce variance, and improve robustness for a variety of machine learning and generative tasks. In Ensemble-GAN [31] and Dropout-GAN [32], ensembling techniques, such as combining multiple discriminators or generators, have been explored to stabilize training and improve visual fidelity. In diffusion models, prior works have explored multiple samplings for uncertainty estimation [33], classifier-free guidance tuning [34], and best-of-N selection [35]. eDiff-I [36] constructs ensembles specialized for different synthesis stages to enhance text alignment in text-to-image generation. Adaptive Feature Aggregation (AFA) [37] proposes dynamically merging multiple diffusion models' feature outputs using a spatial-aware attention mechanism, improving generation robustness. Recently, Korkmaz et al. [38] applied ensembling to diffusion-based image super-resolution by generating multiple outputs, ranking them using a vision-language model (VLM), and fusing the top-ranked ones via pixel-wise averaging.

These studies motivate us to investigate multi-path ensemble sampling in zero-shot image-based video restoration, where improving fidelity and temporal consistency are critical.

III. Baseline: VISION-XL Video Inverse Problem Solver

VISION-XL [2] is a state-of-the-art video inverse problem solver that adapts pretrained SDXL [39] for zero-shot video restoration. It introduces algorithmic strategies to extend single-frame latent diffusion models to multi-frame video tasks.

Given a degraded video $\{\mathbf{y}^{[n]}\}_{n=1}^{N}$ and degradation operator $A : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H'\times W'\times 3}$, the goal is to restore clean frames $\{\mathbf{x}_0^{[n]}\}_{n=1}^{N}$.

One conditioning strategy used in VISION-XL is initializing the diffusion sampling process using DDIM inversion [40] of the measurements. Instead of starting from pure noise, VISION-XL applies DDIM inversion to the latent representation of the first measurement frame, and uses the resulting noisy latent $\mathbf{z}_\tau$ to initialize all frame latents by repetition $\mathbf{Z}_\tau := [\mathbf{z}_\tau, \ldots, \mathbf{z}_\tau]$.

Starting the denoising process from timestep τ, at each timestep $t \in [\tau, \ldots, 0]$, the noisy latent $\mathbf{z}_t$ is passed to SDXL UNet to predict noise $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$, then the clean latent $\hat{\mathbf{z}}_0^{(t)}$ is predicted using Tweedie's formula in 21.

In Data Consistency Optimization step, the latent $\hat{\mathbf{z}}_0^{(t)}$ is decoded using VAE decoder and are refined by minimizing a data-consistency loss using N_{GC}-step conjugate gradient (CG) optimization:

$$\bar{\mathbf{x}}_0 := \arg \min_{X \in \hat{X}_0 + \mathcal{K}_l} \|Y - A(X)\|_2^2 \tag{1}$$

To mitigate VAE errors and suppress high-frequency artifacts, a Gaussian low pass filter is applied on refined frames and then they are encoded to latent space and are re-noised into $\mathbf{z}_{t-1}$ to continue the sampling.

$$\begin{aligned} \bar{\mathbf{z}}_0 &= E_\phi\left(\bar{\mathbf{x}}_0\right) \\ \mathbf{z}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\bar{\mathbf{z}}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \boldsymbol{\epsilon}_t, \end{aligned} \tag{2}$$

where $\boldsymbol{\epsilon}_t$ is composed of batch-consistent noise and deterministic noise:

$$\begin{aligned} c_1 &= \sqrt{1-\bar{\alpha}_{t-1}} \cdot \eta \\ c_2 &= \sqrt{1-\bar{\alpha}_{t-1}} \cdot \sqrt{1-\eta^2} \\ \mathbf{z}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\bar{\mathbf{z}}_0 + c_1 \cdot \boldsymbol{\epsilon}_{\text{fixed}} + c_2 \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) \end{aligned} \tag{3}$$

Temporal Consistency Mechanisms. VISION-XL uses several inference-time strategies to enforce temporal consistency:

- Consistent DDIM Initialization: First-frame DDIM inversion is reused across frames.

- Synchronized Sampling: Shared noise schedule and conditioning across frames.
- Low-Pass Filtering: Reduces flickering by suppressing high-frequency noise.
- CG Optimization: Joint data consistency improves coherence under temporal degradation.

Limitations. Although VISION-XL's inference-time strategies improve temporal coherence and reduce flickering, there exists some limitations: shared latent initialization can over-smooth motion, there is no inter-frame communication during sampling, and CG updates lack explicit temporal modeling. While stronger temporal priors have been proposed in other diffusion frameworks, they often require architectural modifications or retraining on large video datasets—making them incompatible with zero-shot settings.

Temporal inconsistencies in zero-shot diffusion models are also caused by perceptual instability and stochastic variation during sampling. This motivates our work towards stronger temporal coherence: In the following, we propose a set of lightweight, inference-time strategies designed to improve the temporal coherence of zero-shot diffusion-based video restoration models, such as VISION-XL, without sacrificing sample quality or generality.

IV. Perceptual Straightening Guidance (PSG)

A. Motivation and Objective

A major limitation of applying image-based diffusion models to video restoration is the potential of temporal incoherence. While the VISION-XL framework achieves state-of-the-art performance in zero-shot video inverse problems, it primarily focuses on data consistency and lacks an explicit mechanism to enforce temporal consistency. This is important, as the human visual system exhibits strong sensitivity to temporal artifacts. We propose using the PSH as a guiding principle to reduce unnatural fluctuations in the video while preserving sharpness. To this effect, we introduce Perceptual Straightening Guidance (PSG), an inference-time optimization strategy, where video samples obtained by the reverse diffusion process are further optimized using perceptual straightness as a criterion. Unlike changes to model architecture or retraining, PSG is lightweight and task-agnostic, making it suitable across different video degradations in a zero-shot setting.

B. Formulation of PSG

In order to guide the diffusion process toward restoring frames that follow straighter trajectories in the perceptual representation space, we incorporate perceptual straightening as a refinement step during VISION-XL sampling. This involves defining a perceptual straightening loss that penalizes sequences with large curvature in perceptual space, encouraging the generation of temporally natural frames aligned with motion patterns as perceived by the human visual system.

Perceptual Representation Computation Given a sequence of N refined frames $\{\bar{x}_0^{[n]}\}_{n=1}^N$ obtained from the VISION-XL sampling process at timestep t, we compute hierarchical perceptual representations at different stages of the visual processing pipeline. To extract the perceptual representation for a video sequence, we use the two-stage nonlinear model (RetinalDN and V1) suggested by [1] that simulates the early processing of human visual system.

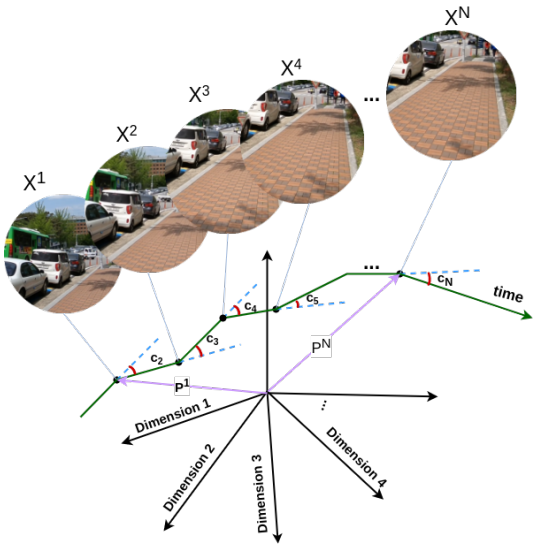

Fig. 1. Visualization of a high-dimensional perceptual representation of a video sequence and the concept of curvature

$$R_{\text{pixel}} = \{\bar{x}_0^{[n]}\}_{n=1}^N \tag{4}$$

$$R_{\text{retina}} = \{f_{\text{retina}}(\bar{x}_0^{[n]}\}_{n=1}^N \tag{5}$$

$$R_{\text{V1}} = \{f_{\text{V1}}(f_{\text{retina}}(\bar{x}_0^{[n]}\}_{n=1}^N \tag{6}$$

where $f_{\text{retina}}(\cdot)$ represents the RetinalDN transformation and $f_{\text{V1}}(\cdot)$ denotes the steerable pyramid processing that simulates V1 complex cell responses. The details can be found in Appendix B. As illustrated in Figure 1, each representation forms a trajectory in high-dimensional space, where temporal consistency is measured through trajectory curvature.

Curvature Computation

Given perceptual embeddings $P^n \in R_{\text{V1}}$ for frame n, we define displacement vectors as $V^n = P^n - P^{n-1}$. The curvature at frame n is computed as the angle between successive displacement vectors between frame representations in V1 space:

$$\text{Curv}(n) = \arccos\left(\frac{V^n \cdot V^{n+1}}{\|V^n\|\|V^{n+1}\|}\right) \tag{7}$$

where the dot product measures the similarity between consecutive displacement directions. The overall trajectory curvature is obtained by averaging across the sequence:

$$\text{Curv}(R_{V1}) = \frac{1}{N-2}\sum_{n=2}^{N-1}\text{Curv}(n) \tag{8}$$

This high-dimensional curvature computation detect subtle temporal inconsistencies that may not be visible in pixel space and provides a more perceptually meaningful measure of temporal naturalness [16].

Perceptual Straightening Penalty Formulation The goal of PSG is to penalize video sequences with large V1 curvature,

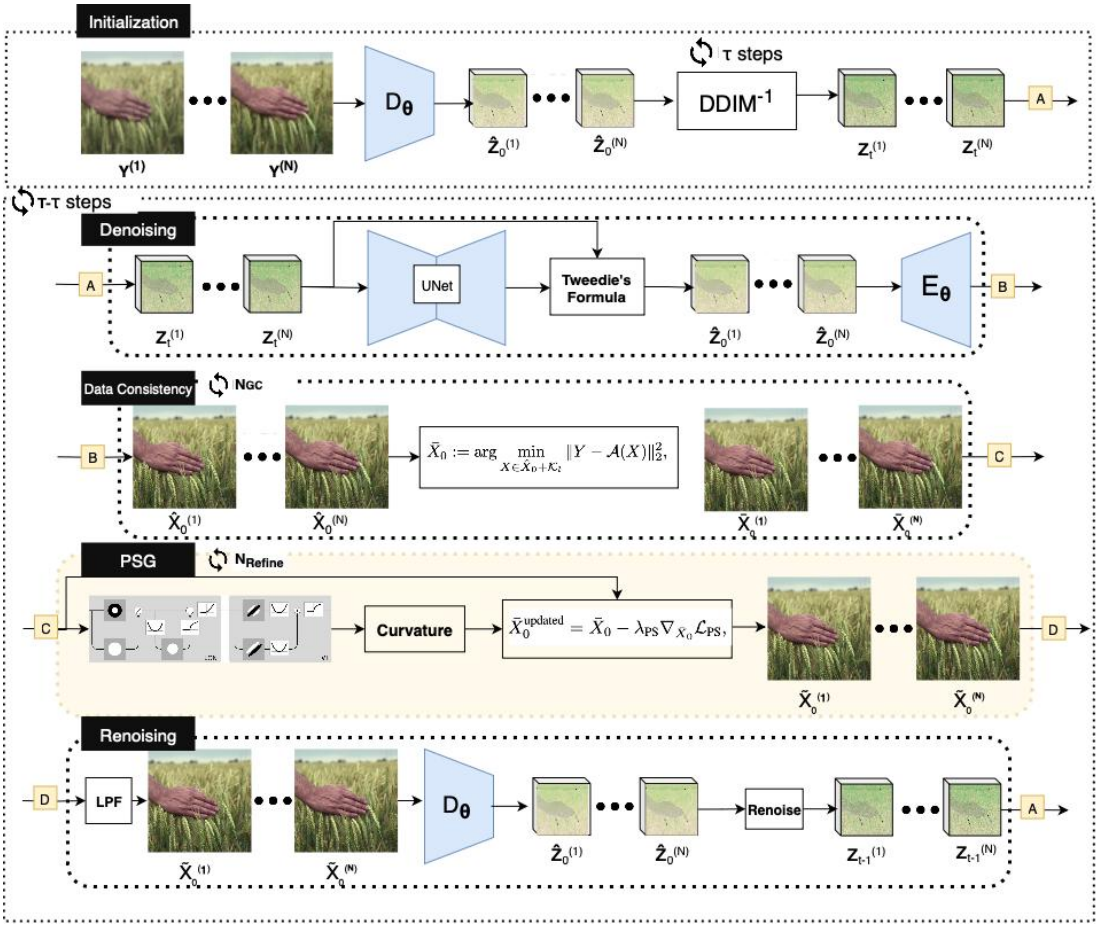

Fig. 2. The proposed PSG block within the inference process of VISION-XL.

thereby encouraging smoother temporal trajectories in the perceptual space. We formulate the perceptual straightening penalty $\mathcal{L}_{PS}$ as a differentiable guidance term that can be seamlessly integrated into the VISION-XL inference pipeline. Given a sequence of predicted clean frames $\bar{X}_0 = \{\bar{x}_0^{[n]}\}_{n=1}^{N}$ at timestep t during the reverse diffusion process, the penalty is defined as:

$$\mathcal{L}_{\text{PS}}(\bar{X}_0) = \text{ReLU}(\text{Curv}(f_{\text{V1}}(f_{\text{retina}}(\bar{X}_0)))) \tag{9}$$

the ReLU activation is used primarily for numerical stability. Since curvature values are inherently non-negative (angles between consecutive displacement vectors range from 0° to 180°), its practical impact is minimal.

We incorporate the perceptual straightening penalty as guidance during diffusion sampling. As illustrated in Figure 2, at each denoising timestep after predicting the clean sequence, we iteratively compute $\mathcal{L}_{\text{PS}}$ and backpropagate the penalty computed in the perceptual space to refine the predicted clean frames using gradient descent:

$$\bar{X}_0^{\text{updated}} = \bar{X}_0 - \lambda_{\text{PS}} \nabla_{\bar{X}_0} \mathcal{L}_{\text{PS}}, \tag{10}$$

where λ_{PS} is a tunable guidance weight controlling the influence of the perceptual straightening penalty during sampling.

The details of PSG process is described in Algorithm 1.

Algorithm 1 Perceptual Straightness Refinement (PSG)

Require: Input frames $\bar{\mathbf{X}}_0$, number of iterations N_{refine}
Ensure: Perceptually refined frames $\tilde{\mathbf{X}}_0$

```
X̃_0 ← X̄_0.detach().clone().requires_grad_(True)
for i = 1 to N_refine do
    L_PS ← PerceptualStraightnessLoss(X̃_0)
    ∇L_PS ← autograd.grad(L_PS, X̃_0)
    X̃_0 ← X̃_0 − λ_PS · ∇L_PS
    if (i + 1) mod 2 = 0 then
        λ_PS ← λ_PS × 0.8
return X̃_0
```

This regularization steers the model toward generating temporally smoother aligned outputs without requiring motion supervision or reference video data.

V. Multi-Path Ensemble Sampling (MPES)

A. Motivation and Objective

The inherent stochasticity of diffusion sampling, where each denoising run can yield slightly different results even for the same input, introduces variations across different trajectories. Although individual predictions may be noisy, their average is likely to be more accurate statistically, and leads to better approximation of the ground truth [41, 42]. In the proposed

multi-path ensemble sampling. instead of relying on a single diffusion trajectory, we generate multiple sampling paths and fuse the resulting video samples to achieve better results.

B. Sources of Stochasticity

In Vision-XL (baseline), repeated runs with the same input can yield different results due to multiple sources of randomness. First, injected noise causes denoising paths to diverge, leading the U-Net to produce different predictions over time—even for identical inputs—resulting in multiple valid sampling trajectories. Second, since Vision-XL operates in the latent space via a pretrained VAE, small input differences introduce varying reconstruction errors. These errors can accumulate during iterative encode-decode cycles and drift the latent from the clean manifold. Lastly, the data consistency step involves conjugate gradient optimization, which can follow different paths depending on the input.

These trajectories explore different regions of the data distribution through diffusion prior. This diversity can be leveraged through ensembling to yield more robust and temporally consistent reconstructions.

C. Proposed Framework

We propose a multi-path ensemble sampling framework that generates multiple diffusion trajectories and fuses them to obtain robust video reconstruction. The key idea is to exploit the statistical advantages of ensemble methods while preserving the temporal consistency of video restoration results.

Let $\mathcal{S}_\theta$ denote a diffusion-based inverse problem solver (such as VISION-XL) that takes degraded measurements $\mathbf{Y}$ as input and generates a restored video sequence $\mathbf{X}$. Our multi-path ensemble approach, summarized in Algorithm 2, produces K independent sampling trajectories:

$$\mathbf{X}^{(k)} = \mathcal{S}_\theta(\mathbf{Y}, \boldsymbol{\xi}^{(k)}), \quad k = 1, 2, \ldots, K \tag{11}$$

where $\boldsymbol{\xi}^{(k)}$ denotes the random seed or stochastic components unique to the k-th path.

In this method, each sampling trajectory $\{\mathbf{z}_t^{(k)}\}_{t=0}^{\tau}$ creates its own independent path through the latent manifold. As a result, the paths can diverge significantly during the sampling process and may explore different regions of the posterior distribution $p(\mathbf{X}|\mathbf{Y})$ (Figure 3).

The final ensemble prediction is obtained by using a fusion function $\mathcal{F}$:

$$\mathbf{X}_{ensemble} = \mathcal{F}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(K)}) \tag{12}$$

Fusion Strategies To combine outputs from multiple diffusion sampling trajectories, we consider different representation space in which fusion is done.

*a) **Representation Spaces**:* Fusion can be applied at different stages of the generation process:

- Latent-Space Fusion: Predicted latents from each trajectory are fused in latent space.
- Pixel-Space Fusion: Latents are decoded into images first, and fusion is done in the pixel space.

In all the cases uniform averaging is used where all trajectories contribute equally to the final output.

Algorithm 2 Multi-Path Ensemble Sampling

Require: Measurements $\mathbf{Y}$, Number of paths K

for $k = 1, \ldots, K$ **do**
  $\mathbf{z}_\tau^{(k)} \leftarrow \text{BatchConsistentInversion}(\mathbf{Y}, \boldsymbol{\xi}^{(k)})$
  for $t = \tau, \tau-1, \ldots, 1$ **do**
    $\hat{\mathbf{z}}_0^{(k)} \leftarrow \text{TweedieDenoising}(\mathbf{z}_t^{(k)}, t)$
    $\hat{\mathbf{X}}_0^{(k)} \leftarrow \text{VAEDecode}(\hat{\mathbf{z}}_0^{(k)})$
    $\bar{\mathbf{X}}_0^{(k)} \leftarrow \text{DataConsistency}(\hat{\mathbf{X}}_0^{(k)}, \mathbf{Y})$
    $\mathbf{z}_{t-1}^{(k)} \leftarrow \text{Renoise\&Encode}(\bar{\mathbf{X}}_0^{(k)}, \boldsymbol{\epsilon}_t^{(k)})$
return $\mathcal{F}_{final}(\bar{\mathbf{X}}_0^{(1)}, \ldots, \bar{\mathbf{X}}_0^{(K)})$

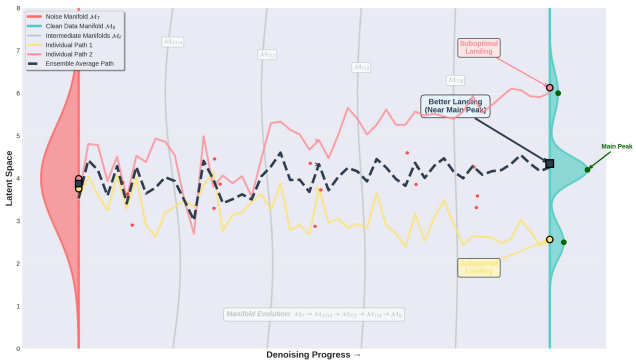

Fig. 3. Multi-Path Ensemble Sampling: Individual vs. average trajectories. The average leads to a better solution than individual ones.

D. Properties of Multi-Path Ensemble Sampling

Ensembling multiple diffusion trajectories offers the following theoretical advantages:

Manifold Regularization via Trajectory Averaging As illustrated in Figure 3, diffusion sampling can be considered as a trajectory across a sequence of latent manifolds $\mathcal{M}_t$, starting from pure noise $\mathcal{M}_T$ and moving toward the clean data manifold $\mathcal{M}_0$. Each sample path $\{\mathbf{z}_t^{(k)}\}_{t=0}^{T}$ may deviate from the related manifold because of the accumulated noise or optimization errors.

Averaging latent from multiple trajectories at each timestep (manifold) has a regularizing effect: when individual latents are near the manifold, their Euclidean average will also stay near the manifold.

Posterior Smoothing and MMSE Estimation As discussed in Appendix A, diffusion-based inverse solvers approximate sampling from the posterior distribution $p(\mathbf{X}|\mathbf{Y})$. Under squared loss, the optimal Bayesian estimator called the Minimum Mean Squared Error (MMSE) estimator is the posterior mean:

$$\mathbf{X}_{MMSE} = \mathbb{E}[\mathbf{X}|\mathbf{Y}] = \int \mathbf{X}\, p(\mathbf{X}|\mathbf{Y})\, d\mathbf{X} \tag{13}$$

This estimator achieves a reconstruction with the minimum expected squared difference from the ground truth. For high-dimensional distributions that the exact estimation of this expectation is not possible, it can be approximated using

Monte Carlo sampling [43]:

$$\mathbf{X}_{MMSE} \approx \frac{1}{K}\sum_{k=1}^{K}\mathbf{X}^{(k)}, \quad \mathbf{X}^{(k)} \sim p(\mathbf{X}|\mathbf{Y}) \tag{14}$$

According to the central limit theorem, the approximation error which is the standard deviation of $\mathbf{X}^{(k)}$ decreases as $O(1/\sqrt{K})$. This analysis offers a foundation for understanding why multi-path ensemble sampling improves quality of video restoration.

VI. Evaluation: Experimental Results

A. Evaluation Setup

Datasets We evaluate our methods on two benchmark datasets: DAVIS and REDS4. We use 73 sequences from DAVIS-2017 [44], resized to 768×1280, preserving their original landscape orientation. DAVIS includes high-quality videos with challenging scenarios, suitable for spatio-temporal restoration. REDS4 consists of four landscape-oriented clips from the REDS dataset [45], each with 100 frames at 720×1280 resolution. It offers diverse motion patterns and scene complexities, which is complementary to DAVIS.

Degradation Models We evaluate on five spatio-temporal inverse problems formulated as compositions of spatial and temporal operators. Spatial degradations include: Super-Resolution (4× downsampling) and Deblurring (Gaussian blur with $\sigma = 3.0$). Spatio-temporal degradations include: SR+ and Deblur+ (each adding 7-frame temporal averaging on top of their spatial counterpart) and Temporal Blur (13-frame averaging). The measurement model is defined as $\mathbf{y} = \mathcal{A}(\mathbf{x})$.

Evaluation Metrics To assess fidelity and perceptual quality, we use spatial distortion metrics (PSNR[46], SSIM [47]), spatial perceptual metrics (LPIPS[48], NIQE[49]), temporal metrics (OF-MSE, FVD [50], Perceptual Straightness) and spatio-temporal metrics (P-ST, and D-ST[16]).

Implementation Platform All experiments are conducted on a single NVIDIA V100 GPU with 32GB memory. Also, all inferences are done with mixed precision computation using half precision (torch.float16) for memory efficiency.

B. Effect of Perceptual Straightening Guidance (PSG)

We implemented a "Sequential Optimization" approach comprised of a two-stage optimization pipeline, where data consistency refinement precedes perceptual straightening optimization. In each sampling step of the VISION-XL pipeline, after decoding the estimated clean latent sequence to pixel-space, we first apply $N_{\text{DDS}} = 5$ conjugate gradient iterations for data consistency (DDS), then we compute perceptual straightening loss and apply $N_{\text{PS}} = 10$ gradient descent steps to enforce straightness in the perceptual space with the learning rate $\lambda_{\text{PS}} = 0.5$.

We also explored alternative PSG schemes, such as "Alternating Optimization," where we alternate single steps of data consistency and perceptual straightening, and "Joint Optimization," where we simultaneously optimize both objectives. However, neither scheme lead to better results. We observed that minimizing curvature when frames are still far from the natural-image manifold acts like a temporal low-pass filter to reduce frame-to-frame variation (i.e., introduces temporal blurring) instead of preserving motion-consistent changes. Overall, applying PSG optimization before the frames are sufficiently "clean" and aligned may not produce perceptually natural motion.

Results and Analysis

We report quantitative results for the proposed sequential optimization in Table I and II. Across DAVIS and REDS4 and for all degradations, the *Straightness* score improves compared to *Baseline+*, evidence that using Perceptual straightening guidance straightens the V1-trajectory for the restored videos. We observe that the spatial scores, PSNR, SSIM, and LPIPS do not improve but stay on par with Baseline+ indicating that perceptual straightening optimization is primarily a temporal regularizer and affects temporal scores.

Our analysis reveals a dependency between task complexity and effectiveness of perceptual straightening. When degradations contain temporal blur (**sr+**, **deblur+**, **temporal-blur**), improvements in the Straightness metric is more pronounced and Straightness correlates strongly with the FVD across both datasets. One explanation for this is that Straightness specifically measures the smoothness of frame-to-frame evolution in a perceptual space, and our perceptual straightening guidance primarily improves temporal coherence. In the motion-blur degradation, the dominant error is temporal (suppressed or smeared motion cues); correcting it both straightens the perceptual trajectory and reduces FVD. In contrast, for spatial-only degradations, temporal dynamics are largely preserved; gains in Straightness reflect minor flicker reduction but do not address the spatial realism that mainly determines FVD; then, the two metrics are not strongly correlated.

Since perceptual straightening acts as a temporal regularizer, we expect it to leave frame appearance largely unchanged while making motion look more natural by small edits. Hence, we demonstrate the visual effect of the PSG on three consecutive frames of a video to observe motion continuity. As Figures 4 and 5 depicts, perceptual straightening guidance on the temporal blur and SR+ tasks leads to reduced micro-wobble in structures like building edges, window frames, less boiling of repetitive textures and cleaner line persistence.

C. Effect of Multi-Path Ensemble Sampling (MPES)

To evaluate the effectiveness of our MPES framework introduced in Section V-C, we performed ablation studies to evaluate key aspects of MPES: (1) the ideal fusion representation space (latent vs pixel) and (2) the effect of ensemble size (K=2 vs K=3).

a) Experiment 1: Latent-Space Fusion (K=2)

In this setup, two independent samplings are run in parallel during the denoising. At the final step, the predicted latent representations are fused just before decoding. A refinement step is used after fusion.

b) Experiment 2: Pixel-Space Fusion (K=2) or (K=3)

In this case, the fusion is applied in the pixel space after decoding K=2 or K=3 samples. A refinement step is again used

TABLE I
QUANTITATIVE EVALUATION ON **DAVIS** DATASET COMPARING **PERCEPTUAL STRAIGHTENING** OPTIMIZATION VERSUS **BASELINE+** ACROSS VARIOUS DEGRADATION TASKS. BEST VALUES ARE **BOLDED**.

Degradation	Method	Spatial				Temporal				Spatio-Temporal	
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	flow_MSE↓	fvd-videogpt↓	fvd-styleganv↓	Straightness↑	D-ST↓	P-ST↓
sr	PS	29.6799	0.8271	**6.0284**	**0.2223**	**0.0142**	85.9838	83.3062	**95.760**	159.9752	**0.1330**
	Baseline+	**29.7367**	**0.8294**	6.1289	0.2249	0.0145	**83.8501**	**82.2478**	95.454	**159.4358**	0.1350
+sr	PS	28.5560	0.7871	**6.2095**	**0.2861**	0.0151	**117.1282**	**116.1784**	**92.790**	178.4411	**0.1766**
	Baseline+	**28.5687**	**0.7878**	6.2499	0.2869	**0.0151**	122.4390	120.0357	92.322	**178.1407**	0.1781
deblur	PS	31.1147	0.8525	6.6008	0.2303	0.0126	70.8423	70.7495	**95.940**	130.8105	**0.1376**
	Baseline+	**31.1342**	**0.8529**	**6.5927**	**0.2302**	**0.0123**	**69.1178**	**68.6272**	95.616	**130.3061**	0.1379
+deblur	PS	**28.1387**	**0.7736**	6.6618	**0.3471**	0.0159	**164.5605**	**164.5999**	**93.096**	189.0676	**0.2136**
	Baseline+	28.1199	0.7735	**6.6476**	0.3473	**0.0159**	169.0877	167.1298	92.25	**189.5531**	0.2157
temporal blur	PS	30.4432	0.7938	**4.5684**	0.2151	**0.0127**	**49.5639**	**48.8714**	**94.410**	**159.1802**	**0.1305**
	Baseline+	**30.4446**	**0.7938**	4.5720	**0.2149**	0.0127	51.4089	50.0101	94.194	159.1931	0.1308

TABLE II
QUANTITATIVE EVALUATION ON **REDS4** DATASET COMPARING **PERCEPTUAL STRAIGHTENING** OPTIMIZATION VERSUS **BASELINE+** ACROSS VARIOUS DEGRADATION TASKS. BEST VALUES ARE **BOLDED**.

Degradation	Method	Spatial				Temporal				Spatio-Temporal	
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	flow_MSE↓	fvd-videogpt↓	fvd-styleganv↓	Straightness↑	D-ST↓	P-ST↓
sr	PS	26.2507	0.7413	**5.2627**	**0.3130**	**0.0131**	54.0367	52.8803	**108.644**	178.3603	**0.1730**
	Baseline+	**26.2763**	**0.7431**	5.3510	0.3195	0.0135	**51.6672**	**51.8888**	103.626	**177.7404**	0.1767
+sr	PS	25.1730	0.6873	**5.8042**	**0.3737**	**0.0147**	**140.0593**	**139.8951**	**98.298**	226.1452	**0.2178**
	Baseline+	**25.2008**	**0.6887**	5.8415	0.3758	0.0149	141.7857	140.7089	98.082	**225.0139**	0.2195
deblur	PS	27.6859	0.7906	6.6324	0.2998	0.0121	50.6460	50.3979	**103.968**	133.2403	0.1652
	Baseline+	**27.7020**	**0.7913**	**6.6048**	**0.2996**	**0.0117**	**46.7897**	**46.8866**	103.932	**132.4128**	**0.1651**
+deblur	PS	25.1441	0.6730	6.2271	0.4246	0.0157	**147.1632**	**147.1276**	**98.856**	230.0379	0.2461
	Baseline+	**25.1720**	**0.6741**	**6.2111**	**0.4235**	**0.0153**	147.5224	147.4916	98.658	**228.1795**	**0.2460**
temporal blur	PS	26.6410	0.7321	3.6202	0.1638	0.0135	**195.7243**	**194.9824**	**98.514**	173.3719	0.0953
	Baseline+	**26.6432**	**0.7323**	**3.6108**	**0.1635**	**0.0133**	207.6850	207.3517	98.424	**173.1191**	**0.0952**

Fig. 4. Visual comparison of PSG vs. Baseline+ on three successive frames on a sequence from **REDS** dataset for the Temporal_blur task: (A) Measurement, (B) Baseline+, (C) PSG, (D) Ground-Truth.

Fig. 5. Visual comparison of PSG vs. Baseline+ on three successive frames on a sequence from **REDS** dataset for the +SR task: (A) Measurement, (B) Baseline+, (C) PSG, (D) Ground-Truth.

after fusion. The experiment with K=3 investigates whether increasing the number of trajectories leads to improved robustness based on the intuition that multiple diverse reconstructions can reduce artifacts through aggregation.

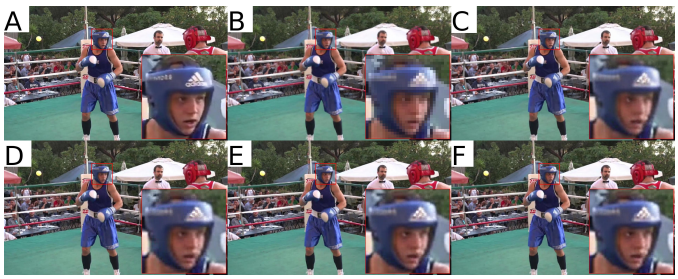

Fig. 6. Visual comparison of **Ensemble sampling** vs. Baseline+ for the **SR** task on **DAVIS** dataset: (A) Ground-Truth, (B) Measurement, (C) Baseline+, (D) Latent Fusion, K=2, (E) Pixel Fusion, K=2, (F) Pixel Fusion, K=3.

Fig. 7. Visual comparison of **Ensemble sampling** vs. Baseline+ for the **+SR** task on **DAVIS** dataset: (A) Ground-Truth, (B) Measurement, (C) Baseline+, (D) Latent Fusion, K=2, (E) Pixel Fusion, K=2, (F) Pixel Fusion, K=3.

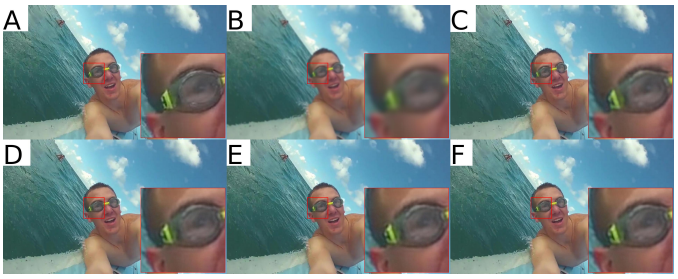

Fig. 8. Visual comparison of **Ensemble sampling** vs. Baseline+ for the **Deblur** task on **DAVIS** dataset: (A) Ground-Truth, (B) Measurement, (C) Baseline+, (D) Latent Fusion, K=2, (E) Pixel Fusion, K=2, (F) Pixel Fusion, K=3.

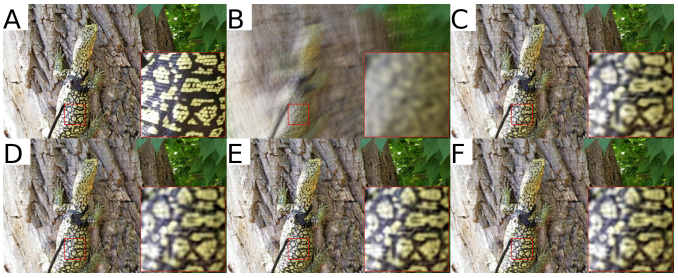

Fig. 9. Visual comparison of **Ensemble sampling** vs. Baseline+ for the **+Deblur** task on **DAVIS** dataset: (A) Ground-Truth, (B) Measurement, (C) Baseline+, (D) Latent Fusion, K=2, (E) Pixel Fusion, K=2, (F) Pixel Fusion, K=3.

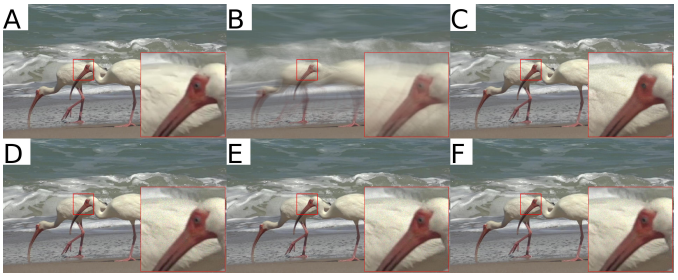

Fig. 10. Visual comparison of **Ensemble sampling** vs. Baseline+ for the **Temporal Blur** task on **DAVIS** dataset: (A) Ground-Truth, (B) Measurement, (C) Baseline+, (D) Latent Fusion, K=2, (E) Pixel Fusion, K=2, (F) Pixel Fusion, K=3.

Results and Analysis

Tables III and IV illustrate the performance comparison of ensemble sampling strategies for DAVIS and REDS4 datasets respectively, displaying consistent patterns and also dataset-specific characteristics that we will go over them. Also, The visual results comparing different ensemble sampling strategies with the baseline across multiple degradation types are presented in Figures 6–10.

a) Latent vs Pixel Space Fusion: Pixel-space fusion is generally better than latent-space fusion over different tasks and datasets. On DAVIS inpainting task, for example, Experiment 3 achieves 32.28 dB PSNR versus 31.17 dB in Experiment 2. This superiority potentially occurs because latent-space fusion loses high-frequency information during the single decoding step of averaged latent representations.

b) Ensemble Size Effects: Increasing the number of trajectories from K=2 to K=3 slightly improves the results across both datasets specially for spatial fidelity metrics. DAVIS shows gains of 0.4-0.5 dB PSNR (temporal blur: 34.72→35.17 dB) and REDS4 exhibits similar improvements (30.53→30.62 dB). As discussed in sectionV-D, these improvements specially in distortion metrics align with ensemble theory that states averaging independent estimators reduces variance. However, the computational cost increases linearly making K=2 more proper in practice.

c) Cross-Dataset evaluation: Performance is consistent for both DAVIS and REDS4, but REDS4 shows 3-4 dB lower PSNR values because of the complexity of motion patterns in it. In general, pixel-space fusion with K=3 shows better metrics on both datasets.

d) Perception-Distortion Trade-offs: To find the method with the best perception-distortion (P-D) balance, Figure 11 depicts the perception-distortion trade-off based on P_{ST} (spatio-temporal perceptual quality) and D_{ST} (spatio-temporal distortion) measures proposed in [16]. It can be seen that the proposed multi-path ensembling strategy offers the best P-D trade-off in all cases (closest to the lower left corner).

Ensemble Method vs. Baseline Comparison

We compare all three ensembling strategies against the baseline Vision-XL initialized with frame-based DDIM inversion (Baseline+) since the same initialization strategy is used in all Ensemble experiments. Table III and Table IV present comparison results for DAVIS and REDS4 datasets.

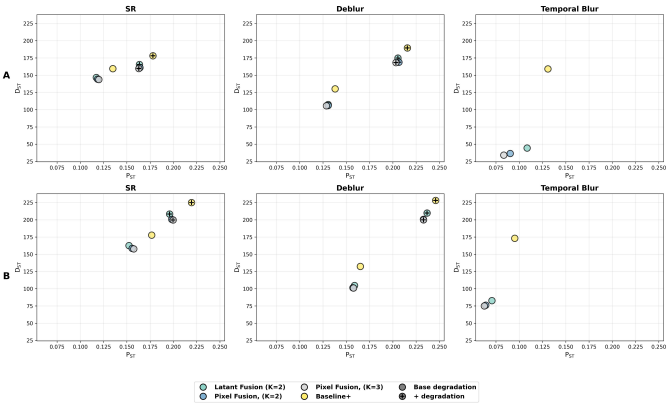

Fig. 11. Perception-Distortion trade-off for different Ensemble methods and Baseline+ on (A) **DAVIS** and (B) **REDS4** datasets.

TABLE III

QUANTITATIVE EVALUATION ON **DAVIS** DATASET COMPARING DIFFERENT **ENSEMBLE** ABLATION EXPERIMENTS ACROSS VARIOUS DEGRADATION TASKS. BEST VALUES ARE **BOLDED** AND BLUE VALUES REPRESENT THE SECOND-BEST SCORES

Degradation	Method	Spatial				Temporal				Spatio-Temporal	
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	flow_MSE↓	fvd-videogpt↓	fvd-styleganv↓	Straightness↑	D-ST↓	P-ST↓
sr	Latent Fusion (K=2)	30.2939	0.8520	**5.6751**	**0.1963**	0.0137	70.5963	70.3234	96.048	146.6228	**0.1171**
	Pixel Fusion, (K=2)	30.4857	0.8558	5.7019	0.1988	0.0134	**67.5485**	**67.6531**	96.084	144.3505	0.1185
	Pixel Fusion, (K=3)	**30.5181**	**0.8566**	5.7677	0.2011	**0.0134**	69.4153	68.6170	**96.086**	**143.7916**	0.1199
	Baseline+	29.7367	0.8294	6.1289	0.2249	0.0145	83.8501	82.2478	95.454	159.4358	0.1350
+sr	Latent Fusion (K=2)	28.9560	0.8003	**5.4961**	**0.2633**	0.0148	99.8820	98.6967	92.250	165.5139	0.1635
	Pixel Fusion, (K=2)	29.2106	0.8076	5.6564	0.2659	0.0144	92.6848	91.0386	92.844	160.9432	0.1641
	Pixel Fusion, (K=3)	**29.2684**	**0.8114**	5.7581	0.2643	**0.0142**	**91.3272**	**90.1016**	**92.952**	**159.7503**	**0.1629**
	Baseline+	28.5687	0.7878	6.2499	0.2869	0.0151	122.4390	120.0357	92.322	178.1407	0.1781
deblur	Latent Fusion (K=2)	32.4332	0.8780	**6.1444**	0.2196	**0.0108**	64.6716	64.7970	96.390	107.5168	0.1305
	Pixel Fusion, (K=2)	32.5858	0.8807	6.1676	0.2196	0.0111	**63.7341**	**64.3902**	96.390	106.0960	0.1305
	Pixel Fusion, (K=3)	**32.6534**	**0.8834**	6.1907	**0.2163**	0.0111	64.4063	64.5867	**96.392**	**105.7445**	**0.1286**
	Baseline+	31.1342	0.8529	6.5927	0.2302	0.0123	69.1178	68.6272	95.616	130.3061	0.1379
+deblur	Latent Fusion (K=2)	28.5169	0.7770	**6.3212**	0.3312	0.0150	147.8147	146.2795	92.340	174.9943	0.2055
	Pixel Fusion, (K=2)	28.7287	0.7768	6.3384	0.3358	**0.0146**	142.6238	139.6994	93.150	168.7994	0.2065
	Pixel Fusion, (K=3)	**28.7903**	**0.7818**	6.3522	**0.3309**	0.0149	**142.5426**	**139.6423**	**93.222**	**168.2588**	**0.2034**
	Baseline+	28.1199	0.7735	6.6476	0.3473	0.0159	169.0877	167.1298	92.250	189.5531	0.2157
temporal blur	Latent Fusion (K=2)	33.5976	0.8168	5.4338	0.1772	0.0115	17.2757	16.9627	93.564	44.3696	0.1085
	Pixel Fusion, (K=2)	34.7168	0.8469	5.2299	0.1481	**0.0109**	12.0802	12.5348	93.978	36.4243	0.0903
	Pixel Fusion, (K=3)	**35.1709**	**0.8583**	5.1405	**0.1369**	0.0111	**10.4322**	**10.6899**	94.068	**34.0602**	**0.0834**
	Baseline+	30.4446	0.7938	**4.5720**	0.2149	0.0127	51.4089	50.0101	**94.194**	159.1931	0.1308

TABLE IV

QUANTITATIVE EVALUATION ON **REDS4** DATASET COMPARING DIFFERENT **ENSEMBLE** ABLATION EXPERIMENTS ACROSS VARIOUS DEGRADATION TASKS. BEST VALUES ARE **BOLDED**.

Degradation	Method	Spatial				Temporal				Spatio-Temporal	
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	flow_MSE↓	fvd-videogpt↓	fvd-styleganv↓	Straightness↑	D-ST↓	P-ST↓
sr	Latent Fusion (K=2)	26.6506	0.7697	5.0066	**0.2774**	0.0126	55.8290	54.7760	104.400	162.5172	**0.1523**
	Pixel Fusion, (K=2)	26.7709	0.7747	**4.9310**	0.2839	0.0124	50.7844	51.1504	104.454	158.5174	0.1557
	Pixel Fusion, (K=3)	**26.7912**	**0.7756**	5.0098	0.2870	**0.0124**	**50.1927**	**50.8866**	**104.472**	**157.7888**	0.1574
	Baseline+	26.2763	0.7431	5.3510	0.3195	0.0135	51.6672	51.8888	103.626	177.7404	0.1767
+sr	Latent Fusion (K=2)	25.5295	0.7063	**5.3644**	**0.3366**	0.0143	135.6013	135.3260	98.424	208.2861	**0.1960**
	Pixel Fusion, (K=2)	25.7114	0.7158	5.3732	0.3421	0.0145	**123.7890**	**123.7817**	98.910	**200.6943**	0.1982
	Pixel Fusion, (K=3)	**25.7314**	**0.7175**	5.4540	0.3450	**0.0143**	124.5278	124.0970	**98.964**	199.7032	0.1997
	Baseline+	25.2008	0.6887	5.8415	0.3758	0.0149	141.7857	140.7089	98.082	225.0139	0.2195
deblur	Latent Fusion (K=2)	28.7609	0.8308	5.9514	0.2912	0.0112	54.5014	52.9646	105.029	104.6369	0.1588
	Pixel Fusion, (K=2)	28.9183	0.8361	**5.9123**	**0.2880**	0.0111	51.8873	53.1965	105.029	101.4920	**0.1571**
	Pixel Fusion, (K=3)	**28.9350**	**0.8378**	5.9346	0.2893	**0.0111**	52.5554	52.6310	**105.030**	**101.1511**	0.1578
	Baseline+	27.7020	0.7913	6.6048	0.2996	0.0117	**46.7897**	**46.8866**	103.932	132.4128	0.1651
+deblur	Latent Fusion (K=2)	25.5242	0.6810	6.0540	0.4088	**0.0142**	134.2759	133.3326	98.928	209.8171	0.2368
	Pixel Fusion, (K=2)	25.7344	0.6873	**6.0460**	**0.4039**	0.0146	126.7702	125.8548	99.414	200.6415	**0.2328**
	Pixel Fusion, (K=3)	**25.7601**	**0.6910**	6.0617	0.4045	0.0148	**126.4900**	**125.6219**	**99.486**	**199.9967**	0.2330
	Baseline+	25.1720	0.6741	6.2111	0.4235	0.0153	147.5224	147.4916	98.658	228.1795	0.2460
temporal blur	Latent Fusion (K=2)	30.1218	0.8552	**2.9064**	0.1215	0.0136	143.9532	143.6348	98.640	82.7805	0.0706
	Pixel Fusion, (K=2)	30.5346	0.8660	2.9575	0.1110	**0.0127**	**138.6359**	**138.4109**	99.414	76.1517	0.0639
	Pixel Fusion, (K=3)	30.6237	**0.8699**	2.9538	**0.1085**	0.0128	143.9935	142.2372	**99.504**	**75.0959**	**0.0625**
	Baseline+	26.6432	0.7323	3.6108	0.1635	0.0133	207.6850	207.3517	98.424	173.1191	0.0952

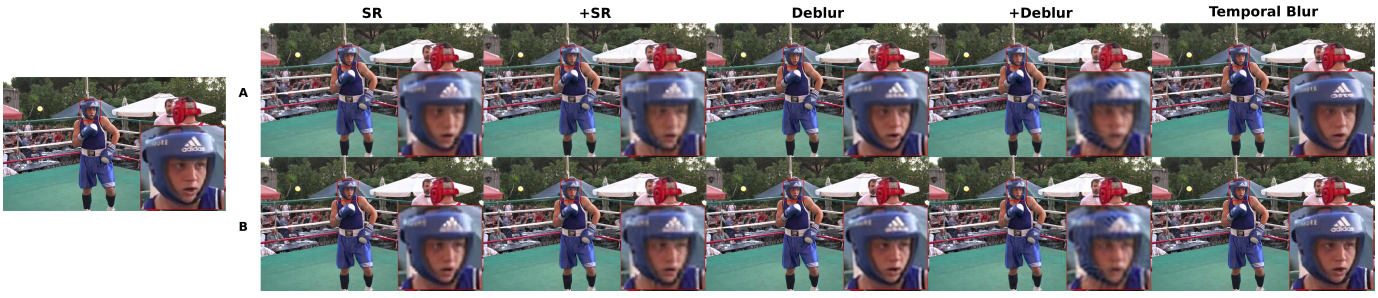

Fig. 12. Visual comparison **(A) Baseline+** versus **(B) pixel fusion (K=2)** on a video from the **DAVIS** dataset for different degradations.

a) Universal Ensemble Improvements: All ensemble methods outperform Baseline+ across almost all restoration tasks and metrics. For example, on DAVIS deblurring, all independent methods considerably outperform Baseline+: Latent-space K=2 (32.43 dB), Pixel-space K=2 (32.59 dB) and K=3 (32.65 dB) all surpass Baseline+ (31.13 dB) by 1.3-1.5 dB. This consistent improvement indicates that trajectory diversity substantially improves reconstruction quality.

b) Task-Dependent Performance Improvement: By investigating the results, we can see that the benefit of ensemble sampling is correlated with task difficulty. Simple tasks like SR show moderate 0.5-0.8 dB improvement for PSNR across all ensemble methods, while challenging tasks show huge gains: deblurring improvements range from 1.3-1.5 dB, and temporal degradations achieve 3.2-4.7 dB gains. This observation shows

the importance of ensemble diversity as inverse problems approach face limitations with single path.

Figure 12 shows the comparison results of pixel-space fusion (k=2) versus Baseline+ for all restoration tasks. As it is shown from the results, ensemble method improved the restoration quality for all tasks.

c) Ensemble Method Generalizability: In spite of different content characteristics of DAVIS and REDS4 and different degradation types, ensemble methods show consistent improvement over Baseline+. This observation strongly validate the generalization ability of our ensemble sampling framework.

d) Computational Efficiency: Independent pixel-fusion with K=2 provides the best balance between performance (across both datasets) and computational efficiency. It achieves similar performance as Independent pixel-fusion with K=3, while requiring only about double (rather than triple) of the computational load of Baseline+.

The results indicate that independent pixel fusion ensemble sampling has significant and consistent improvements over single-path Vision-XL baseline. Specifically for challenging inverse problems, where single path sampling may lead to suboptimal solutions, the role of trajectory diversity becomes more crucial and it is worth the additional computation.

VII. Conclusion and Future Directions

We propose two complementary inference-time strategies: i) Perceptual Straightening Guidance (PSG) and ii) Multi-Path Ensemble Sampling (MPES) to improve temporal consistency and fidelity in zero-shot diffusion-based video restoration.

The proposed PSG leverages a neuroscience-inspired hypothesis by introducing a curvature penalty in the retinal–V1 space. It is used both as a perceptual metric and as a guidance term to refine denoising trajectories. In particular, when degradations contain temporal blur (`sr+`, `deblur+`, `temporal-blur`), the improvement in Straightness is more pronounced and aligns better with FVD. One explanation is that perceptual straightening improves frame-to-frame smoothness, correcting temporal artifacts such as motion smearing and micro-wobble in structures like edges and textures.

The proposed MPES strategy shows that averaging multiple diffusion trajectories leads to consistent gains in both spatial fidelity, perceptual quality, and temporal coherence across diverse restoration tasks. Ensemble methods help reduce stochastic artifacts and stabilize frame evolution without requiring retraining. Despite the theoretical risk of blurring due to MMSE approximation, our analysis shows that correlation among trajectories and concentration near the natural manifold preserves structure and enhances stability.

These findings highlight that simple, training-free strategies like PSG and MPES can significantly improve zero-shot diffusion-based video restoration. There are several potential directions for future extensions of our work:

- **Better Perceptual Embedders:** Future work can adopt feature spaces that were shown to separate natural and AI-generated video trajectories, such as those used in perceptual-straightening-based detectors [30], for both the metric and the guidance.
- **Partial-path Guidance:** Perceptual-straightening guidance can be applied only in the mid/late denoising steps or on shorter frame sequences to avoid over-smoothing.
- **Improved Curvature Penalty:** One can adopt margin-based (hinge/Huber) curvature penalties with adaptive thresholds and a scheduler that increases the PS update weight as samples approach the image manifold.
- **Advanced Fusion Strategies:** Developing adaptive fusion mechanisms, such as attention-based or quality-aware weighting mechanisms, that dynamically combine results based on reconstruction confidence.
- **Generalization Studies:** Testing multi-path ensemble sampling using different diffusion architectures (e.g., Stable Diffusion 1.5 [51], pix-art [52], Consistency Models) and broader domains.
- **Partial-Path Ensembling:** Instead of running K full trajectories, we could run K trajectories only for initial denoising steps (e.g., 30–50% of the denoising process), then fuse them into a single path for the remaining steps. This reduces computational cost while still using early-stage diversity to stabilize reconstruction.

Appendix

A. Diffusion Models for Inverse Problems

Denoising Diffusion Probabilistic Models (DDPMs) [17], employ a two-step process: a forward diffusion process that adds noise to an image and a learned reverse process that gradually removes the noise. Latent Diffusion Models (LDMs) [53] perform diffusion in a lower-dimensional latent space $\mathcal{Z}$ to improve the computational efficiency.

Let $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times C}$ be the clean image and (E_θ, D_θ) be a pretrained VAE encoder–decoder pair, then:

$$\mathbf{z}_0 = E_\theta(\mathbf{x}_0), \quad \hat{\mathbf{x}}_0 = D_\theta(\mathbf{z}_0), \tag{15}$$

where $\mathbf{z}_0 \in \mathbb{R}^{h\times w\times c}$, $h \ll H$, $w \ll W$. The forward process over T timesteps is defined in $\mathcal{Z}$ as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) := \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}), \tag{16}$$

with the closed-form:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{17}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$.

The reverse process is modeled as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) := \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \tag{18}$$

Experimentally, [33] chose to set $\Sigma_\theta(\mathbf{z}_t, t) = \beta_t\mathbf{I}$ while the mean $\boldsymbol{\mu}_\theta$ is obtained by predicting the noise $\boldsymbol{\epsilon}$ added in the forward process. Diffusion models are trained by optimizing the simple mean-squared error (MSE) loss between the true and predicted noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right]. \tag{19}$$

Bayesian Posterior Sampling: Restoration is interpreted as sampling from the posterior:

$$p_\theta(\mathbf{x}_0|\mathbf{y}) \propto p_\theta(\mathbf{x}_0) \cdot p(\mathbf{y}|\mathbf{x}_0), \tag{20}$$

where we aim to estimate an unknown signal $\mathbf{x}_0$ from degraded observations $\mathbf{y} = A(\mathbf{x}_0)$, with A being the corruption process, $p_\theta(\mathbf{x}_0)$ is the learned prior provided by pre-trained diffusion model and $p(\mathbf{y}|\mathbf{x}_0)$ is the likelihood eflecting how well a reconstructed image is consistent with observed measurement. In diffusion-based inverse problem solvers the main problem is how to inject the conditioning information from $p(\mathbf{y}|\mathbf{x}_0)$ into this generative process.

To do so, one strategy is guiding the reverse process using intermediate denoised estimates $\hat{\mathbf{z}}_0^{(t)}$, computed via Tweedie's formula [41]:

$$\hat{\mathbf{z}}_0^{(t)} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) \right), \tag{21}$$

and then decoding the estimated clean latent to pixel space and refining it by minimizing a data-consistency loss:

$$\mathcal{L}_{\text{data}}(\hat{\mathbf{x}}_0^{(t)}) = \|A(\hat{\mathbf{x}}_0^{(t)}) - \mathbf{y}\|^2. \tag{22}$$

The refined sample is re-noised and the reverse process continues. This iterative refinement effectively integrates measurement constraints into the generative framework, making diffusion models suitable for a wide range of inverse problems.

B. Perceptual Representation in PSH

The perceptual representation for a video is computed through a two-stage nonlinear model suggested by [1] that simulates the early processing of human visual system illustrated in Figure 13. The two-stage perceptual encoder structure is illustrated in Figure 14.

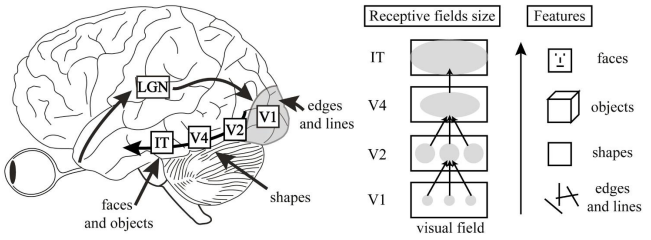

Fig. 13. Early processing of human visual system [54]

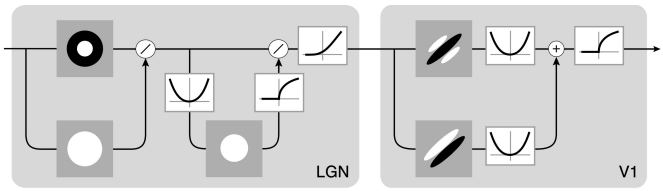

Fig. 14. Two-stage cascade model describing computations in the retina, lateral geniculate nucleus (LGN) and V1 [1]

a) Stage1: RetinaND: *RetinaND* module is designed to mimic the early visual transformations in the retina and lateral geniculate nucleus (LGN) of the human's visual system. It converts the video input into a form that is more aligned with human vision. It normalizes the lighting conditions, emphasizes edges and motion transitions to provide a robust representation for subsequent spatiotemporal modeling. The RetinaND

Fig. 15. Output of RetinaND module

module, implemented in the official codebase [1], consists of three convolutional layers, corresponding to:

1) **Center-surround filtering**: enhances local contrast and edges via a Difference-of-Gaussians kernel, implemented as a linear convolution:

$$y = k_{\text{lin}} * x \tag{23}$$

where x is the input frame, k_{lin} is the center-surround filter, and $*$ shows 2D convolution.

2) **Luminance gain control**: normalizes the filtered response using a smoothed estimate of local brightness:

$$l = k_{\text{lum}} * x, \qquad y \leftarrow \frac{y}{1 + \alpha \cdot l} \tag{24}$$

where k_{lum} is the luminance smoothing kernel, and α is a fixed scalar gain parameter.

3) **Contrast gain control**: normalizes the signal based on a local measure of contrast energy:

$$c = \sqrt{k_{\text{con}} * y^2 + \varepsilon}, \qquad y \leftarrow \frac{y}{1 + \beta \cdot c} \tag{25}$$

where k_{con} is a contrast kernel, β is another gain parameter, and ε is a small constant for numerical stability.

4) **Nonlinear activation**, applies Softplus nonlinearity to enhance robustness and biological plausibility:

$$y \leftarrow \log\left(1 + e^y\right) \tag{26}$$

All filters and normalization constants are designed analytically using known biological response functions. The input and output of RetinaND module for a sample image are depicted in Figure 15.

b) Stage 2: V1 Feature Extraction: This stage of the perceptual model mimics the primary visual cortex (V1) whose responses are known to be tuned to specific spatial orientations and scales. A *Steerable Pyramid* decomposition is used to extract orientation-, scale-, and frequency-selective responses from the output of the RetinaND module.

The Steerable Pyramid is a linear, multiscale, and orientation-selective transform that uses a set of separable bandpass and lowpass filters in the Fourier domain to extract multi-scale, orientation-selective features through the steps:

1) The input y is transformed to the frequency domain:

$$\hat{y}^{(0)} = \mathcal{F}\{y\} \tag{27}$$

2) At each scale $s = 1, \ldots, S$, bandpass filters $B_k^{(s)}$ extract orientation-selective responses for $k = 1, \ldots, K$

directions, while a lowpass filter $L^{(s)}$ generates the coarser scale:

$$\hat{z}_k^{(s)} = B_k^{(s)} \cdot \hat{y}^{(s-1)}, \qquad \hat{y}^{(s)} = L^{(s)} \cdot \hat{y}^{(s-1)} \tag{28}$$

3) Inverse Fourier transform reconstructs spatial responses:

$$z_k^{(s)} = \mathcal{F}^{-1}\{\hat{z}_k^{(s)}\} \tag{29}$$

4) After the pyramid decomposition, both the high-pass (finest scale) and low-pass (coarsest scale) bands are removed. The remaining complex-valued orientation responses are converted to magnitude energy maps:

$$e_k^{(s)} = |z_k^{(s)}|^2 = \Re(z_k^{(s)})^2 + \Im(z_k^{(s)})^2 \tag{30}$$

5) At the last step each energy map $e_k^{(s)}$ is flattened and concatenated into a single feature vector:

$$\mathbf{v} = \texttt{concat}\left(\texttt{vec}(e_k^{(s)}) \ \forall \ s, k\right) \tag{31}$$

This representation, which is used in the trajectory straightness analysis, encodes multi-scale, orientation-specific contrast energy, reflecting the known response of V1 simple cells.

REFERENCES

[1] O. J. Hénaff, R. L. Goris, and E. P. Simoncelli, "Perceptual straightening of natural videos," *Nature Neuroscience*, vol. 22, no. 6, pp. 984–991, 2019.

[2] T. Kwon and J. C. Ye, "VISION-XL: High definition video inverse problem solver using latent image diffusion models," *Int. Conf. Comp. Vision (ICCV)*, 2025.

[3] W. Zhao and H. S. Sawhney, "Is super-resolution with optical flow feasible?," in *European Conf. on Computer Vision*, pp. 599–613, 2002.

[4] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[5] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Trans. on Image Processing*, vol. 18, no. 9, pp. 1958–1975, 2009.

[6] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. on Image Processing*, vol. 18, no. 1, pp. 36–51, 2008.

[7] P. Vandewalle, K. Krichane, D. Alleysson, and S. Süsstrunk, "Joint demosaicing and super-resolution imaging from a set of unregistered aliased images," in *Digital Photography III*, vol. 6502, pp. 78–89, SPIE, 2007.

[8] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[9] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 4947–4956, 2021.

[10] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 5972–5981, 2022.

[11] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018.

[12] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. Van Gool, "Vrt: A video restoration transformer," *IEEE Trans. on Image Processing*, vol. 33, pp. 2171–2182, 2024.

[13] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool, "Recurrent video restoration transformer with guided deformable attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 378–393, 2022.

[14] J. Wang, G. Teng, and P. An, "Video super-resolution based on generative adversarial network and edge enhancement," *Electronics*, vol. 10, no. 4, p. 459, 2021.

[15] M. Chu, Y. Xie, L. Leal-Taixé, and N. Thuerey, "Temporally coherent GANs for video super-resolution (tecogan)," *arXiv preprint arXiv:1811.09393*, vol. 1, no. 2, p. 3, 2018.

[16] N. Rahimi and A. M. Tekalp, "Spatio-temporal perception-distortion trade-off in learned video SR," in *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 1400–1404, IEEE, 2023.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] S. Zhou, P. Yang, J. Wang, Y. Luo, and C. C. Loy, "Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution," in *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 2535–2545, 2024.

[19] X. Li, Y. Liu, S. Cao, Z. Chen, S. Zhuang, X. Chen, Y. He, Y. Wang, and Y. Qiao, "Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency," *arXiv e-prints*, pp. arXiv–2501, 2025.

[20] C. Rota, M. Buzzelli, and J. van de Weijer, "Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models," in *European Conf. on Computer Vision*, pp. 36–53, Springer, 2024.

[21] Z. Zou, J. Liu, S. Shoushtari, Y. Wang, and U. S. Kamilov, "Flair: A conditional diffusion framework with applications to face video restoration," in *IEEE/CVF Winter Conf. on App. of Computer Vision (WACV)*, pp. 5228–5238, IEEE, 2025.

[22] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in neural information processing systems*, vol. 35, pp. 8633–8646, 2022.

[23] Y. Bao, D. Qiu, G. Kang, B. Zhang, B. Jin, K. Wang, and P. Yan, "Latentwarp: Consistent diffusion latents for zero-shot video-to-video translation," *arXiv preprint arXiv:2311.00353*, 2023.

[24] G. Daras, W. Nie, K. Kreis, A. Dimakis, M. Mardani, N. Kovachki, and A. Vahdat, "Warped diffusion: Solving video inverse problems with image diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 101116–101143, 2024.

[25] T. Kwon and J. C. Ye, "Solving video inverse problems using image diffusion models," *Int. Conf. on Learning Representations (ICLR)*, 2025.

[26] C. Cao, H. Yue, X. Liu, and J. Yang, "Zero-shot video restoration and enhancement using pre-trained image diffusion model," in *AAAI Conf. on Artificial Intelligence*, vol. 39, pp. 1935–1943, 2025.

[27] C.-H. Yeh, C.-Y. Lin, Z. Wang, C.-W. Hsiao, T.-H. Chen, H.-S. Shiu, and Y.-L. Liu, "Diffir2vr-zero: Zero-shot video restoration with diffusion-based image restoration models," *preprint arXiv:2407.01519*, 2024.

[28] P. Kancharla and S. S. Channappayya, "Improving the visual quality of video frame prediction models using the perceptual straightening hypothesis," *IEEE Signal Proc. Letters*, vol. 28, pp. 2167–2171, 2021.

[29] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Trans. on Image Processing*, vol. 30, pp. 4098–4111, 2021.

[30] C. Internò, R. Geirhos, M. Olhofer, S. Liu, B. Hammer, and D. Klindt, "AI-generated video detection via perceptual straightening," *arXiv preprint arXiv:2507.00583*, 2025.

[31] J. Zhang, C. Tao, Z. Xu, Q. Xie, W. Chen, and R. Yan, "Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation," in *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 435–444, 2019.

[32] G. Mordido, H. Yang, and C. Meinel, "Dropout-gan: Learning from a dynamic ensemble of discriminators," *preprint arXiv:1807.11346*, 2018.

[33] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[34] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[35] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[36] J. Zhang, Y. Zhou, J. Gu, C. Wigington, T. Yu, Y. Chen, T. Sun, and R. Zhang, "Artist: Improving the generation of text-rich images with disentangled diffusion models and large language models," in *IEEE/CVF Winter Conf. on App. of Computer Vision (WACV)*, pp. 1268–1278, IEEE, 2025.

[37] C. Wang, K. Tian, Y. Guan, F. Shen, Z. Jiang, Q. Gu, and J. Zhang, "Ensembling diffusion models via adaptive feature aggregation," in *Int. Conf. on Learning Representations*, 2025.

[38] C. Korkmaz, A. M. Tekalp, and Z. Doğan, "Leveraging vision-language models to select trustworthy super-resolution samples generated by diffusion models," *IEEE Trans. on Circuits and Systems for Video Technology*, 2025.

[39] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," *Int. Conf. on Learning Representations (ICLR)*, 2024.

[40] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *Int. Conf. on Learning Representations (ICLR)*, 2021.

[41] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011.

[42] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[43] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[44] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675*, 2017.

[45] S. Nah *et al.*, "NTIRE 2019 challenge on video super-resolution: Methods and results," in *IEEE Conf. on Comp. Vision and Pattern Recog. Workshops*, 2019.

[46] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Int. Conf. on Pattern Recognition (ICPR)*, pp. 2366–2369, IEEE, 2010.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pp. 586–595, 2018.

[49] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[50] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Fvd: A new metric for video generation," 2019.

[51] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[52] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," 2023.

[53] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conf. on Comp. Vis. Patt. Recog. (CVPR)*, pp. 10684–10695, 2022.

[54] M. H. Herzog and A. M. Clarke, "Why vision is not both hierarchical and feedforward," *Frontiers in Computational Neuroscience*, vol. 8, p. 135, 2014.