Temporal-Consistent Video Restoration with Pre-trained Diffusion Models

Hengkang Wang[1*], **Yang Liu**[2], **Huidong Liu**[2], **Chien-Chih Wang**[2], **Yanhui Guo**[2]
Hongdong Li[2,3], **Bryan Wang**[2], **Ju Sun**[1]
[1]**Computer Science and Engineering, University of Minnesota** {wang9881,jusun}@umn.edu
[2]**Amazon.com, Inc.** {yliuu,liuhuido,ccwang,yanhuig,hongdli,brywan}@amazon.com [3]**ANU**

Abstract

Video restoration (VR) aims to recover high-quality videos from degraded ones. Although recent zero-shot VR methods using pre-trained diffusion models (DMs) show good promise, they suffer from approximation errors during reverse diffusion and insufficient temporal consistency. Moreover, dealing with 3D video data, VR is inherently computationally intensive. In this paper, we advocate viewing the reverse process in DMs as a function and present a novel Maximum a Posterior (MAP) framework that directly parameterizes video frames in the seed space of DMs, eliminating approximation errors. We also introduce strategies to promote bilevel temporal consistency: semantic consistency by leveraging clustering structures in the seed space, and pixel-level consistency by progressive warping with optical flow refinements. Extensive experiments on multiple virtual reality tasks demonstrate superior visual quality and temporal consistency achieved by our method compared to the state-of-the-art. Our project webpage with sample code is at: https://sun-umn.github.io/Temporal-Consistent-Video-Restoration/.

1 Introduction

Video restoration (VR) aims to recover high-quality (HQ) videos $\boldsymbol{X}$ from given low-quality (LQ) observations $\boldsymbol{Y} \approx \mathcal{A}(\boldsymbol{X})$, where $\mathcal{A}$ represents a spatial and/or temporal degradation. Typical VR tasks include super-resolution (Zhou et al. 2024; Chan et al. 2022; Wang et al. 2024-12-01; Liang et al. 2024), inpainting (Zhou et al. 2023; Lugmayr et al. 2022; Xu et al. 2019), and deblurring (Zhong et al. 2020; Nah et al. 2019). Modern VR methods rely on deep learning and fall into two main categories. **(1) The supervised learning approach** trains deep neural networks (DNNs) on LQ-HQ paired data, i.e., $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i)\}_{i=1}^K$, to learn direct mappings from $\boldsymbol{Y}$ to $\boldsymbol{X}$. Although conceptually simple, these methods require large-scale, high-quality paired datasets and significant computational resources (e.g. *32 A100-80G GPUs* for super-resolution (Zhou et al. 2024)). Moreover, they need to train new DNNs for different VR tasks, with limited task adaptability; **(2) The zero-shot paradigm** enabled by pre-trained deep generative models, especially diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020-10-02). Due to the lack of mature video DMs, recent research (Yeh et al. 2024-10-04; Cao et al. 2024; Kwon and Ye 2024-10-04,b) has adopted pre-trained image DMs for VR, achieving remarkable results without task-specific retraining.

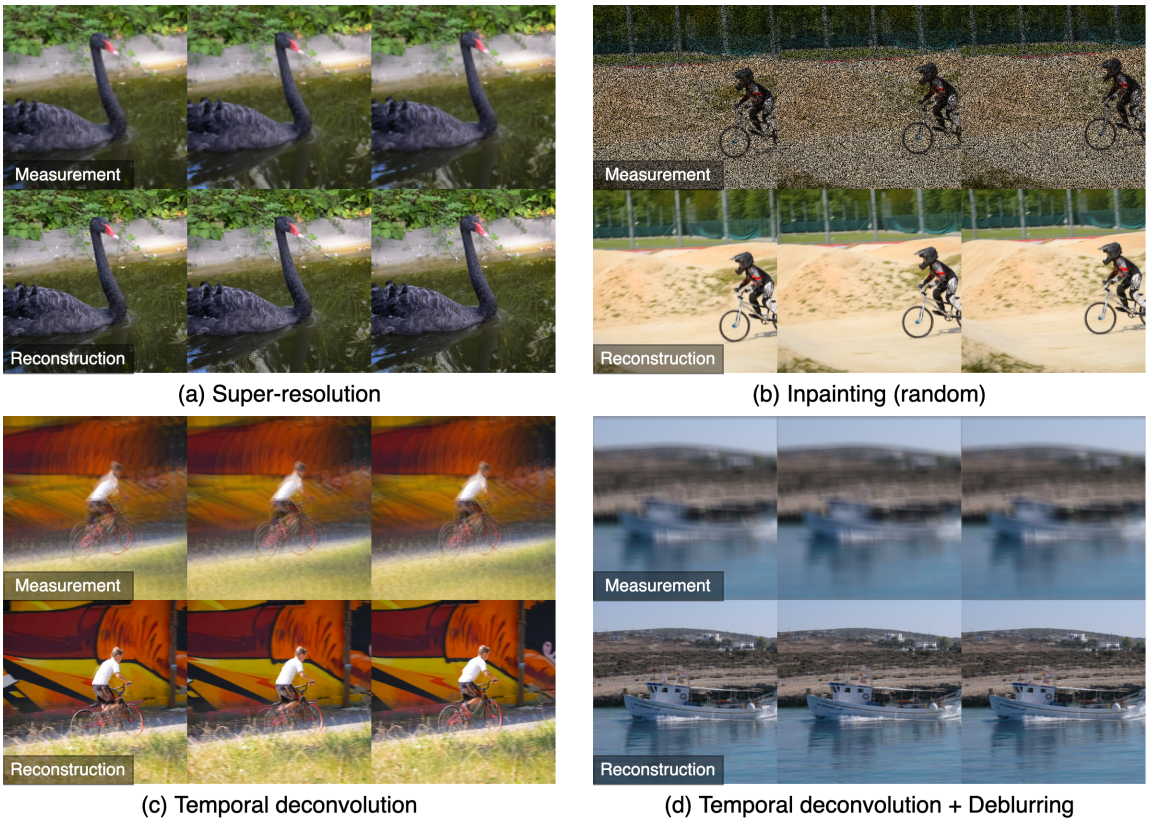
Most zero-shot DM-based VR methods (Kwon and Ye 2024-10-04; Cao et al. 2024; Kwon and Ye 2024b; Yeh et al. 2024-10-04) interleave reverse diffusion steps with iterative gradient updates or projections to approach the feasible set $\{\boldsymbol{X} \mid \boldsymbol{Y} \approx \mathcal{A}(\boldsymbol{X})\}$. However, these methods usually face three fundamental challenges. First, they suffer from **unavoidable approximation errors (Challenge 1)** when approximating an intractable likelihood (Chung et al. 2022-09-29), regardless of whether they employ gradient updates or projections. These errors accumulate throughout the reverse diffusion process, potentially degrading the reconstruction quality. Second, these zero-shot methods often struggle for satisfactory **temporal consistency (Challenge 2)** due to the difficulty in extracting accurate motion information from LQ measurements and the absence of explicit motion priors in the image DMs they leverage. Third, compared to image restoration, VR entails **significantly more computation and memory footprints (Challenge 3)**, creating substantial efficiency barriers that must be addressed for practical applications.

In this paper, we focus on solving VR problems using pre-trained image DMs, while addressing the three fundamental challenges faced by state-of-the-art (SOTA) methods. We specifically choose **latent diffusion models (LDMs)** as our backbone DMs due to their superior generation quality, computational efficiency, and widespread adoption. **(Tackling Challenge 1)** Our method builds upon the classical Maximum a Posterior (MAP) framework (Ulyanov, Vedaldi, and Lempitsky 2018; Pan et al. 2022-11; Zhuang et al. 2024-02-01; Li et al. 2023a,b; Wang et al. 2024):

$$\min_{\boldsymbol{X}} \underbrace{\ell(\boldsymbol{Y}, \mathcal{A}(\boldsymbol{X}))}_{\text{data consistency}} + \lambda \underbrace{\Omega(\boldsymbol{X})}_{\text{regularization}} . \tag{1}$$

We take a novel perspective that views the entire reverse diffusion process as a function $\mathcal{R}$ which, when composed with the pre-trained decoder $\mathcal{D}$ in LDMs, maps from the seed space directly to the image manifold $\mathcal{M}$. This allows us to naturally reparameterize the video frame-by-frame as $\boldsymbol{X} = [\mathcal{D} \circ \mathcal{R}(\boldsymbol{z}_1), \ldots, \mathcal{D} \circ \mathcal{R}(\boldsymbol{z}_N)]$; plugging this into Eq. (1)

[*]This work of Wang H. was partially done while interning at Amazon.com, Inc.



(a) Super-resolution (b) Inpainting (random)

(c) Temporal deconvolution (d) Temporal deconvolution + Deblurring

Figure 1: Visualization results by our method: (a) Super-resolution $\times 4$; (b) Inpainting with 50% random pixel masking; (c) Temporal deconvolution using uniform PSF with kernel width $k = 7$; and (d) Temporal deconvolution with motion deblurring.

leads to a unified optimization formulation **with respect to the seeds** $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$. This reparametrization approach effectively leverages the powerful LDM priors for VR while avoiding the approximation errors inherent in SOTA methods. To **address Challenge 2**, we design a hierarchical framework to promote bilevel temporal consistency. For **semantic-level temporal consistency**, we first explore the seed space and observe an intriguing *clustering phenomenon*: input seeds of frames from different videos are scattered, while those of frames within the same video are clustered. Motivated by this, we construct a noise prior by hypothesizing that consecutive frames share a common seed with only minor frame-specific variations. To enhance **pixel-level temporal consistency**, we implement a progressive warping mechanism that combines image warping with incremental optical flow (OF) refinements. Our ablation study in Table 7 underscores the effectiveness of both components in enhancing bilevel temporal consistency. To **deal with Challenge 3**, we introduce an efficient diffusion sampling strategy using the DDIM sampler. Notably, our ablation study (Section 3.3) reveals that 4 reverse steps in $\mathcal{R}$ are sufficient to achieve the SOTA performance. Moreover, by leveraging the multivariate mean value theorem, we significantly reduce the computational complexity of the proposed method. To further boost efficiency, we make the trainable residuals for each frame more lightweight through low-rank approximations.

Our contributions can be summarized as follows:

- We propose a MAP-based framework for VR that harnesses pre-trained image DMs by reparameterizing frames via the entire reverse diffusion process, eliminating the approximation errors that have plagued SOTA methods.
- We devise a compelling hierarchical approach for bilevel temporal consistency that unites semantic-level coherence (through our key discovery of clustering patterns in the seed space) with pixel-level precision (via dynamic progressive warping with optical flow refinements).
- We design our method with exceptional computational efficiency through three innovations: an optimized DDIM sampling strategy that requires only 4 steps, an approximate reformulation using the multivariate mean value theorem, and memory-efficient trainable residuals with low-rank approximations.
- Our comprehensive experiments on challenging VR tasks demonstrate that our method leads to substantial performance improvements over SOTA methods, providing exceptional visual quality and temporal consistency.

2 Background and related work

Diffusion models (DMs) Recently, DMs have dominated generative models, capable of producing high-quality objects. The early denoising diffusion probabilistic model (**DDPM**) (Ho, Jain, and Abbeel 2020) involves two processes:

a *forward diffusion* process that transforms any clean data sample $\boldsymbol{x}_0 \sim p_{\text{data}}$ into pure noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ by sequential noise injection, governed by the stochastic differential equation (SDE): $d\boldsymbol{x} = -\beta_t/2 \cdot \boldsymbol{x}dt + \sqrt{\beta_t}d\boldsymbol{w}$, where β_t represents the noise schedule, and $\boldsymbol{w}$ denotes the standard Wiener process; a *reverse diffusion* process that performs sequential denoising, turning any seed noise into a useful data sample—hence responsible for data generation, and follows:

$$d\boldsymbol{x} = -\beta_t \left[\boldsymbol{x}/2 + \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] dt + \sqrt{\beta_t} d\overline{\boldsymbol{w}}, \quad (2)$$

where $\overline{\boldsymbol{w}}$ is the time-reversed Brownian motion, and the term $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$, known as the score function, represents the gradient of the log-likelihood. To train an DM, this score function is approximated using a DNN, $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x})$, trained via score matching techniques (Hyvärinen 2005; Song and Ermon 2019). In practice, the diffusion process is discretized in T time steps, using a predefined variance schedule $\beta_1, \ldots, \beta_T$. Defining $\alpha_t \doteq 1 - \beta_t$ with $\alpha_T \to 0$ and $\bar{\alpha}_t \doteq \prod_{s=1}^{t} \alpha_s$, the forward steps are written as: $\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The reverse steps run as $\boldsymbol{x}_{t-1} = 1/\sqrt{\alpha_t} \cdot \left(\boldsymbol{x}_t - \beta_t \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}_t)/\sqrt{1-\bar{\alpha}_t}\right) + \sqrt{\beta_t}\boldsymbol{z}$; this iterative sampling process usually requires a large number of steps to achieve high-quality generation, leading to slow inference. To address this slowness, the denoising diffusion implicit model (**DDIM**) (Song, Meng, and Ermon 2020-10-02) introduces a non-Markovian relaxation of the forward process, allowing each step $\boldsymbol{x}_t$ to depend not only on $\boldsymbol{x}_{t-1}$ but also directly on $\boldsymbol{x}_0$. This relaxation enables sampling with fewer steps while maintaining generation quality. The reverse step in DDIM is

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}_t), \quad (3)$$

where $\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) \doteq (\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x}_t))/\sqrt{\bar{\alpha}_t}$ estimates the clean image $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$.

Despite DDIM's speedup, training diffusion models in high-resolution pixel spaces remains computationally expensive. Latent diffusion models (**LDMs**) (Rombach et al. 2022) overcome this by performing both training and inference in low-dimensional latent spaces, wrapped into pre-trained encoder-decoder models. The LDM framework has become dominant in SOTA visual generative models (Rombach et al. 2022; Podell et al. 2023-10-13; Peebles and Xie 2023).

DMs for video restoration Approaches to VR using DMs can be categorized into two main classes: supervised (Daras et al. 2024; Zhou et al. 2024) and zero-shot (Kwon and Ye 2024-10-04,b; Cao et al. 2024; Yeh et al. 2024-10-04). Supervised methods train DM-based models on paired data or correlated noise, which is outside our focus. Zero-shot methods largely inherit ideas from zero-shot image restoration with pre-trained DMs and most of them focus on directly modeling the conditional distribution $p_t(\boldsymbol{x}|\boldsymbol{y})$ and substitute the unconditional score function $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ in Eq. (2) with the conditional score function $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}|\boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{y}|\boldsymbol{x})$, i.e.,

$$d\boldsymbol{x} = -\beta_t \left[\boldsymbol{x}/2 + (\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{y}|\boldsymbol{x}))\right] dt + \sqrt{\beta_t} d\overline{\boldsymbol{w}}. \quad (4)$$

Here, while $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ can be approximated using the pre-trained score function $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{x})$, the term $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{y}|\boldsymbol{x})$ remains intractable because $\boldsymbol{y}$ does not depend directly on $\boldsymbol{x}(t)$. To circumvent this, one line of research directly approximates $p_t(\boldsymbol{y}|\boldsymbol{x}(t))$ (Chung et al. 2022-09-29; Fei et al. 2023), while the other interleaves diffusion reverse steps from Eq. (3) with projections (or gradient updates) (Chung, Lee, and Ye 2023-10-13) to guide the process toward the feasible set $\{\boldsymbol{x}|\boldsymbol{y} \approx \mathcal{A}(\boldsymbol{x})\}$. Unfortunately, both strategies introduce **unavoidable approximation errors (Challenge 1)** that can accumulate during the reverse diffusion process (RDP), ultimately limiting their practical performance. For VR, (Cao et al. 2024) follows the generative diffusion prior (GDP) framework (Fei et al. 2023) using gradient steps, while (Kwon and Ye 2024-10-04,b) implement the decomposed diffusion sampling (DDS) (Chung, Lee, and Ye 2023-10-13) approach with several conjugate gradient (CG) update steps. However, *CG requires the degradation operator $\mathcal{A}$ to be known, symmetric and positive-definite* (Shewchuk 1994), restricting its applicability for VR.

VR also needs good temporal consistencies. Existing zero-shot methods use batch-consistent sampling strategies (Kwon and Ye 2024-10-04,b) and leverage optical flow (OF) guidance. However, accurate OF estimation from LQ videos is inherently challenging. Different approaches attempt to overcome this limitation: (Yeh et al. 2024-10-04) directly obtains OF from LQ videos and proposes a hierarchical latent warping technique, while (Cao et al. 2024) acquires OFs during intermediate stages of the RDP—when results may still contain noise—and applies these OFs for warping in the image space. In summary, the SOTA methods achieve only semantic-level alignment (Kwon and Ye 2024-10-04,b; Yeh et al. 2024-10-04) or attempt pixel-level alignment with sub-optimal OFs (Cao et al. 2024), leaving **temporal consistency** in VR a critical standing challenge **(Challenge 2)**.

3 Our method

In this paper, we employ **pre-trained image LDMs** as priors to solve video restoration (VR) problems due to their superior generation quality, computational efficiency, and widespread adoption. In Section 3.1, we propose a novel LDM-based formulation that overcomes the limitations of SOTA methods, addressing **challenge 1**. In Section 3.2, we introduce two effective components for bilevel temporal consistency, tackling **challenge 2**. Finally, in Section 3.3, we present several essential techniques to ensure computational and memory efficiency, targeting **challenge 3**.

3.1 Our basic formulation

To mitigate approximation errors in existing interleaving methods, we introduce a novel and principled formulation for solving VR problems following the classical maximum a posteriori (MAP) principle. Our goal is to recover a high-quality video $\boldsymbol{X}$ that not only satisfies the measurement constraint $\boldsymbol{Y} \approx \mathcal{A}(\boldsymbol{X})$, but also resides near the realistic video manifold $\mathcal{M}$: $\min_{\boldsymbol{X} \in \mathcal{M}} \ \ell(\boldsymbol{Y}, \mathcal{A}(\boldsymbol{X}))$. Inspired by (Wang et al. 2024), we propose viewing the entire RDP as a function $\mathcal{R}$, which, when composed with the pre-trained decoder $\mathcal{D}$ in LDMs,

maps the seed space to the video space. This fresh perspective enables us to reparameterize the video as $\boldsymbol{X} = \mathcal{D} \circ \mathcal{R}(\boldsymbol{Z})$, which can then be plugged into the MAP framework Eq. (1), leading to a unified formulation:

$$\boldsymbol{Z}^* \in \min_{\boldsymbol{Z}} \ \ell(\boldsymbol{Y}, \mathcal{A}(\mathcal{D} \circ \mathcal{R}(\boldsymbol{Z}))), \tag{5}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]$, and $\mathcal{D}$ and $\mathcal{R}$ are applied framewise. The final reconstruction can be obtained as $\boldsymbol{X}^* = \mathcal{D} \circ \mathcal{R}(\boldsymbol{Z}^*)$. Note that the RDP consists of multiple iterative steps. Mathematically, a single reverse step can be expressed as a function g that depends on $\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(i)}$, i.e., $g_{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(i)}}$, representing the i-th reverse step that maps the latent variable $\boldsymbol{z}_{i+1}$ to $\boldsymbol{z}_i$. The full RDP can then be written as

$$\mathcal{R} \doteq g_{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(0)}} \circ g_{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(1)}} \circ \cdots \circ g_{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(T-2)}} \circ g_{\boldsymbol{\varepsilon}_{\boldsymbol{\theta}}^{(T-1)}}. \tag{6}$$

3.2 Promoting bilevel temporal consistency

Although Eq. (5) can address VR tasks leveraging pre-trained image LDMs, the reconstructed videos still may not have good temporal consistency (see "Base" in Table 7). This is not surprising, as it tries to recover individual frames separately and fails to capture inter-frame dependencies—a critical issue to be addressed by all methods relying on image DMs (Cao et al. 2024; Kwon and Ye 2024-10-04,b; Yeh et al. 2024-10-04). In this section, we address this critical issue **(Tackling Challenge 2)** by introducing two distinct components that target temporal consistency at two different levels.

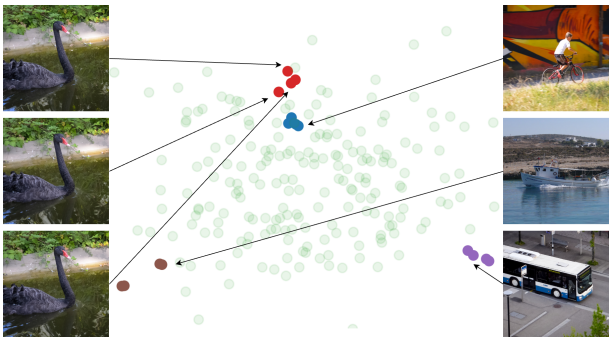


Figure 2: T-SNE (Maaten and Hinton 2008) visualization of seeds extracted for video frames. Green dots are i.i.d. Gaussian noise. Seeds for the same video form clusters, while those for different videos are scattered.

Noise prior for semantic-level consistency

An intriguing clustering phenomenon To strengthen temporal consistency, we begin by exploring potential structures in the seed space. We obtain seeds $\boldsymbol{Z}$ for randomly sampled frames from videos by solving the regression problem: $\min_{\boldsymbol{Z}} \ \ell(\boldsymbol{X}, \mathcal{D} \circ \mathcal{R}(\boldsymbol{Z}))$. From Fig. 2, we observe an interesting pattern: seeds for frames of different videos are widely scattered without apparent correlation, whereas **seeds for frames of the same video tend to cluster together**—a pattern we can potentially leverage to improve temporal consistency across frames.

To this end, we hypothesize that consecutive frames share a common seed but have minor frame-specific deviations. Accordingly, we decompose the seed matrix as $\boldsymbol{Z} = \boldsymbol{z}_{shared}\mathbb{1}^\intercal + \boldsymbol{R}$, where $\boldsymbol{z}_{shared}$ represents the shared seed, $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N]$ captures frame-specific residuals, and $\mathbb{1}$ is an all-one vector. Moreover, we need to ensure that the residuals are small by constraining $\|\boldsymbol{r}_i\|_2 \leq \sigma$ for a small constant $\sigma > 0$. Our new formulation built on Eq. (5) is

$$\begin{aligned} \min_{\boldsymbol{z}_{shared}, \boldsymbol{R}} \quad & \ell(\boldsymbol{Y}, \mathcal{A}(\mathcal{D} \circ \mathcal{R}(\boldsymbol{z}_{shared}\mathbb{1}^\intercal + \boldsymbol{R}))), \\ \text{s.t.} \quad & \|\boldsymbol{r}_i\|_2 \leq \sigma, \quad \forall i \in \{1, \ldots, N\}, \end{aligned} \tag{7}$$

where $\mathcal{D}$ and $\mathcal{R}$ are applied framewise. To solve this constrained optimization problem, we choose the Projected Gradient Descent (PGD) method. Our algorithm alternates between gradient descent and projection:

$$\begin{aligned} (\boldsymbol{z}_{shared}, \boldsymbol{R}) &\leftarrow (\boldsymbol{z}_{shared}, \boldsymbol{R}) - \eta \nabla_{(\boldsymbol{z}_{shared}, \boldsymbol{R})} \ell, \\ \boldsymbol{r}_i &\leftarrow \Pi_{\|\boldsymbol{r}_i\|_2 \leq \sigma}(\boldsymbol{r}_i), \quad \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{8}$$

Here, $\Pi_{\|\boldsymbol{r}_i\|_2 \leq \sigma}$ denotes the projection operator that maps $\boldsymbol{r}_i$ to the closest point within the ℓ_2 norm ball of radius σ.

Progressive warping for pixel-level consistency As shown in Table 7, incorporating the noise prior improves both data fidelity and temporal consistency. However, fine-grained consistency across frames remains weak. To enhance it, we introduce a progressive warping loss:

$$\begin{aligned} \mathcal{L}_{\text{warp}} &= \sum\nolimits_{n=1}^{N-1} \ell\left(\boldsymbol{M} \odot \boldsymbol{x}'_n, \boldsymbol{M} \odot \mathcal{W}\left(\boldsymbol{x}'_{n+1}, \boldsymbol{f}'_{n \to n+1}\right)\right), \\ \boldsymbol{f}'_{n \to n+1} &= \text{stopgrad}\left(\text{RAFT}\left(\boldsymbol{x}'_n, \boldsymbol{x}'_{n+1}\right)\right), \end{aligned} \tag{9}$$

where $\mathcal{W}$ denotes backward warping (Sun et al. 2018), $\boldsymbol{M}$ is the estimated non-occlusion mask obtained via forward-backward consistency checks (Meister, Hur, and Roth 2018-04-27), and $\boldsymbol{f}'_{n \to n+1}$ represents the optical flow (OF) from frame n to $n+1$ and is treated as constant during backpropagation. The frame reconstructed at the time step n is denoted as $\boldsymbol{x}'_n$. We compute the OF using the pre-trained estimator RAFT (Teed and Deng 2020). This warping loss explicitly penalizes pixel-level changes between motion-compensated consecutive frames, thereby enhancing temporal consistency at the pixel level.

In practice, we introduce the warping loss $\mathcal{L}_{\text{warp}}$ only after the estimated frames $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_N$ attain sufficient quality, as early iterations often yield noisy results. The evolution of performance over iterations is illustrated in Fig. 3; it shows an evident performance boost in both data fidelity and temporal consistency after incorporating the proposed warping loss. To reduce computation, we update the OFs every P iterations instead of each. We term our approach **progressive warping** because the estimated OFs themselves are progressively refined during iterations. However, these estimated OFs can exhibit slight fluctuations, which can hinder stable training. To address this, we apply an exponential moving average (EMA) to smooth out the OFs:

$$\boldsymbol{f}^{(t)}_{n \to n+1} = \beta \boldsymbol{f}^{(t-1)}_{n \to n+1} + (1 - \beta) \boldsymbol{f}'^{(t)}_{n \to n+1}, \tag{10}$$

where $\boldsymbol{f}^{(t)}_{n \to n+1}$ is the stabilized flow at iteration t, $\boldsymbol{f}'^{(t)}_{n \to n+1}$ is the newly estimated flow, and $\beta \in [0, 1)$ is the weighted

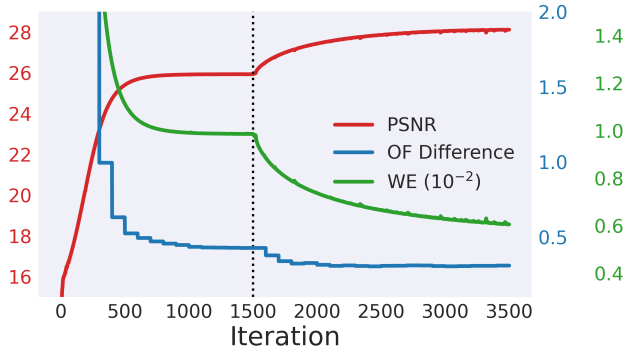


Figure 3: Evolution of key metrics during the iterative VR process: PSNR, OF difference (measured against ground truth), and warping error (WE). All metrics improve in early iterations. After OF difference stabilizes (vertical dotted line), progressive warping is activated, further enhancing PSNR and reducing WE, demonstrating our framework's effectiveness for temporal consistency.

averaging coefficient. This smoothing mechanism not only stabilizes the training process, but also enhances the temporal consistency of the OFs.

3.3 Boosting computational and memory efficiency

Efficient diffusion sampling DDPMs typically require dozens or hundreds of sampling steps, inducing significant computational and memory burdens for our approach in Eq. (7). To address this challenge, we implement the DDIM sampler for the reverse process $\mathcal{R}$, which enables us to skip intermediate steps while preserving generation quality. Surprisingly, we find that 4 reverse steps are sufficient for our method to outperform the SOTA, as shown in Section 3.3. Adding more steps does not provide substantial benefits and can even slightly degrade the results, possibly due to numerical issues from vanishing gradients. Hence, we take 4 as the default number of reverse steps. We also implement gradient checkpointing techniques to further reduce memory costs.

Steps	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
SOTA (Kwon and Ye 2024b)	26.03	0.717	0.339	1.411
2	27.87	0.785	0.324	0.742
4	**27.95**	**0.790**	**0.321**	**0.725**
10	27.70	0.777	0.347	0.746

Table 1: Ablation study on performance vs. the number of reverse steps in diffusion process $\mathcal{R}$, performed on the DAVIS dataset for video super-resolution $\times 4$.

Efficient reformulation via mean value theorem (MVT) There is a potential computational bottleneck in Eq. (7): When performing backpropagation with respect to $\boldsymbol{R}$, the term $\mathcal{R}(\boldsymbol{z}_{shared}\mathbb{1}^\intercal + \boldsymbol{R})$ requires N separate forward passes through $\mathcal{R}$, expensive both in computation and in memory when N is large. To address this, we assume that $\mathcal{R}: \mathbb{R}^d \to \mathbb{R}^d$ is continuously differentiable, with the operator norm (i.e., the largest singular value) of the Jacobian uniformly bounded by L. Then, for all i, $\|\mathcal{R}(\boldsymbol{z}_{shared} + \boldsymbol{r}_i) - \mathcal{R}(\boldsymbol{z}_{shared})\| \leq L\|\boldsymbol{r}_i\|_2 \leq L\sigma$ by the multivariate mean-value theorem (Coleman 2012). So we consider

$$\min_{\boldsymbol{z}_{shared}, \boldsymbol{R}} \ell(\boldsymbol{Y}, \mathcal{A}(\mathcal{D}(\mathcal{R}(\boldsymbol{z}_{shared}) + \boldsymbol{R}))), \quad \text{s.t. } \|\boldsymbol{r}_i\|_2 \leq L\sigma, \quad \forall i \in \{1, \ldots, N\}. \tag{11}$$

Since $\boldsymbol{R}$ is now outside the RDP $\mathcal{R}$, backpropagation through $\mathcal{R}$ is only needed for $\boldsymbol{z}_{shared}$, reducing both memory and computation by a factor of N. To further reduce the cost due to backpropagation, we repeat the above idea by decomposing the pre-trained decoder $\mathcal{D}$ as $\mathcal{D} = \mathcal{D}_1 \circ \mathcal{D}_2$ and putting the learnable residuals as input to $\mathcal{D}_1$, leading to

$$\boxed{\min_{\boldsymbol{z}_{shared}, \boldsymbol{R}} \ell(\boldsymbol{Y}, \mathcal{A}(\mathcal{D}_1(\mathcal{D}_2 \circ \mathcal{R}(\boldsymbol{z}_{shared}) + \boldsymbol{R}))), \quad \text{s.t. } \|\boldsymbol{r}_i\|_2 \leq L'\sigma, \quad \forall i \in \{1, \ldots, N\}.} \tag{12}$$

Here, L' depends on both L and also the maximum operator norm of the Jacobian of $\mathcal{D}_2$, and accounts for the maximum amplification of perturbations through $\mathcal{D}_2 \circ \mathcal{R}$, ensuring rigorous bounds and applicability of our formulation to long video sequences. In practice, we place the trainable $\boldsymbol{R}$ directly on the last layer of $\mathcal{D}$. We set $L'\sigma$ as $C\sqrt{\text{dimension}(\boldsymbol{r}_i)}$ for all i, where C is a tunable hyperparameter.

Lightweight low-rank residual parameterization For VR, the trainable residuals $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$ are high-dimensional tensors. In our implementation with Stable Diffusion, for example, the last layer of $\mathcal{D}$ has dimension $(1, 128, 512, 512)$ for each residual. This leads to substantial memory and computation burdens. We address this by enforcing low-rank structures on the residuals. Specifically, we decompose each residual along its spatial dimensions as: $\boldsymbol{r}_n = \boldsymbol{A}_n\boldsymbol{B}_n$, where $\boldsymbol{A}_n \in \mathbb{R}^{1\times 128\times 512\times k}$ and $\boldsymbol{B}_n \in \mathbb{R}^{1\times 128\times k\times 512}$ with $k \ll 512$. This reduces the total parameter count for $\boldsymbol{R}$ from $O(N \cdot 128 \cdot 512 \cdot 512)$ to $O(N \cdot 128 \cdot 512 \cdot 2k)$, resulting in significant computational savings. To make the low-rank parameterization compatible with the PGD algorithm to solve Eq. (12), we need to ensure $\|\boldsymbol{A}_n\boldsymbol{B}_n\| \leq L'\sigma$ for all i. For this, we take the heuristic projection

$$(\boldsymbol{A}'_n, \boldsymbol{B}'_n) \leftarrow (\boldsymbol{A}_n, \boldsymbol{B}_n)/\sqrt{\|\boldsymbol{A}_n\boldsymbol{B}_n\|/L'\sigma}, \tag{13}$$

after each gradient step on $(\boldsymbol{A}_n, \boldsymbol{B}_n)$. This produces a feasible $\boldsymbol{r}_n$, as $\|\boldsymbol{A}'_n\boldsymbol{B}'_n\| = \|\boldsymbol{A}_n\boldsymbol{B}_n\|/(\|\boldsymbol{A}_n\boldsymbol{B}_n\|/L'\sigma) = L'\sigma$. With extra analytical and computational efforts, it is possible to develop a rigorous orthogonal projector here. But, we stick to this simple one, as it is easy to implement and effective.

4 Experiments

Experimental setup We conduct comprehensive experiments on five VR tasks that involve various spatial and temporal degradations, following the protocols in (Kwon and Ye 2024-10-04,b; Cao et al. 2024). The first three tasks involve spatial degradation only: (1) $4\times$ **super-resolution**, where low-resolution capture is simulated by applying $4\times$ average pooling to high-resolution videos; (2) **inpainting** with random masking at a missing rate of $r = 0.5$; and (3)

Algorithm 1: Our video restoration framework

Input: Epochs E, transition E_T, diffusion steps T, $\boldsymbol{Y}$, $\mathcal{A}$

Initialize $\boldsymbol{z}^0_{shared} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and residuals $\boldsymbol{r}^0_1, \cdots \boldsymbol{r}^0_N$
for $e = 0$ to $E-1$ **do**
  for $i = T-1$ to 0 **do** } $\mathcal{R}$
    $\hat{\boldsymbol{s}} \leftarrow \boldsymbol{\varepsilon}^{(i)}_{\boldsymbol{\theta}}(\boldsymbol{z}^e_i)$
    $\hat{\boldsymbol{z}}^e_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_i}}\left(\boldsymbol{z}^e_i - \sqrt{1-\bar{\alpha}_i}\hat{\boldsymbol{s}}\right)$
    $\boldsymbol{z}^e_{i-1} \leftarrow$ DDIM reverse with $\hat{\boldsymbol{z}}^e_0, \hat{\boldsymbol{s}}$
  end for
  ## Current reconstruction for $n-$th frame
  $\boldsymbol{x}'_n = \mathcal{D}_1(\mathcal{D}_2 \circ \mathcal{R}(\boldsymbol{z}^e_{shared}) + \boldsymbol{r}^e_n)$
  if $e >= E_T$ **then**
    $\boldsymbol{f}'^{(e)}_{n\to n+1} \leftarrow \text{stopgrad}\left(\textbf{RAFT}(\boldsymbol{x}'_n, \boldsymbol{x}'_{n+1})\right)$
    $\boldsymbol{f}^{(e)}_{n\to n+1} = \beta \boldsymbol{f}^{(e-1)}_{n\to n+1} + (1-\beta)\boldsymbol{f}'^{(e)}_{n\to n+1}$
    Calculate $\mathcal{L}_{\text{warp}}$ via Eq. (9)
  else
    $\mathcal{L}_{\text{warp}} = 0$
  end if
  Update $\boldsymbol{z}^{e+1}_{shared}, \boldsymbol{r}^{e+1}_1, \cdots, \boldsymbol{r}^{e+1}_N$ via Eq. (12) with $\mathcal{L}_{\text{warp}}$
  Project each $\boldsymbol{r}^{e+1}_n$ onto ℓ_2 norm ball
end for

Output: Recovered video $[\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_N]$, where $\boldsymbol{x}'_n = \mathcal{D}_1\left(\mathcal{D}_2 \circ \mathcal{R}\left(\boldsymbol{z}^E_{shared}\right) + \boldsymbol{r}^E_n\right)$

motion deblurring, where blurry videos are simulated by applying a 33×33 motion blur kernel of strength 0.5 to clean videos. In addition, we examine (4) **temporal deconvolution**, where degraded videos are generated by applying a uniform point-spread-function (PSF) convolution of width 7 along the temporal dimension, simulating the common artifact that multiple frames blend together in time-varying video capturing. The last task (5) **temporal deconvolution with spatial deblurring** combines (4) temporal deconvolution and (3) spatial motion deblurring, representing complex scenarios.

Competing methods While our work focuses on zero-shot methods for VR using pre-trained image DMs, we benchmark our proposed method against both zero-shot and supervised methods. To ensure fair comparison, we select methods with publicly available implementations and evaluate them using their default settings wherever possible. For zero-shot methods, we compare against SVI (Kwon and Ye 2024-10-04), VISION-XL (with SDXL) (Kwon and Ye 2024b), VISION-base (with SD-base) (Kwon and Ye 2024b), and DiffIR2VR (for super-resolution only) (Yeh et al. 2024-10-04). For supervised methods, our benchmarks include the following: SD $\times 4$ (Rombach et al. 2022), VRT (Liang et al. 2024), RealBasicVSR (Chan et al. 2022), StableSR (Wang et al. 2024-12-01), and Upscale-A-Video (UAV) (Zhou et al. 2024) for super-resolution; SD Inpainting (Rombach et al. 2022) and ProPainter (Zhou et al. 2023) for inpainting; and VRT (Liang et al. 2024), DeBlurGANv2 (Kupyn et al. 2019), Stripformer (Tsai et al. 2022), and ID-Blau (Wu et al. 2024) for motion deblurring and/or temporal deconvolution. **In all tables included in Section 4.1, supervised and zero-shot methods are above and below the dotted line, respectively**. We leave implementation details in (Wang et al. 2025).

4.1 Results

Methods	DAVIS				REDS			
	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
SD $\times 4$ (Rombach et al. 2022)	24.33	0.615	0.358	1.523	23.57	0.619	0.371	1.769
VRT (Liang et al. 2024)	25.97	0.780	0.295	1.063	24.54	0.756	0.305	1.266
RealBasicVSR (Chan et al. 2022)	26.34	0.734	**0.294**	0.962	26.31	0.759	**0.260**	1.019
StableSR (Wang et al. 2024-12-01)	22.56	0.590	0.339	1.977	21.27	0.578	0.342	2.535
UAV (Zhou et al. 2024)	23.65	0.589	0.397	1.623	23.02	0.593	0.422	1.920
DiffIR2VR (Yeh et al. 2024-10-04)	25.01	0.637	0.337	1.321	24.14	0.633	0.328	1.592
SVI (Kwon and Ye 2024-10-04)	23.19	0.562	0.457	1.796	21.93	0.534	0.496	2.322
VISION-XL (Kwon and Ye 2024b)	26.95	0.749	0.349	0.981	25.71	0.725	0.377	1.215
VISION-base (Kwon and Ye 2024b)	26.17	0.712	0.338	1.137	24.95	0.687	0.376	1.406
Ours	**28.21**	**0.799**	0.315	**0.698**	**27.40**	**0.792**	0.339	**0.824**
Ours vs. Best compe.	+1.26	+0.019	+0.021	-0.264	+1.09	+0.033	+0.079	-0.195

Table 2: (Spatial task) Quantitative comparisons for video **super-resolution** $\times 4$ (**Bold**: best, under: second best, green: performance increase, red: performance decrease)

Methods	DAVIS				REDS			
	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
SD Inpainting (Rombach et al. 2022)	16.98	0.258	0.679	6.776	15.77	0.238	0.677	9.253
ProPainter (Zhou et al. 2023)	28.60	0.823	0.281	0.655	28.05	0.827	0.273	0.733
SVI (Kwon and Ye 2024-10-04)	25.80	0.699	0.318	1.134	24.61	0.679	0.323	1.389
VISION-XL (Kwon and Ye 2024b)	29.93	0.862	0.186	0.612	28.66	0.840	0.207	0.745
VISION-base (Kwon and Ye 2024b)	26.54	0.732	0.286	1.033	25.30	0.711	0.293	1.268
Ours	**34.27**	**0.947**	**0.124**	**0.273**	**33.19**	**0.942**	**0.125**	**0.347**
Ours vs. Best compe.	+4.34	+0.085	-0.062	-0.339	+4.53	+0.102	-0.082	-0.386

Table 3: (Spatial task) Quantitative comparisons for video **inpainting** with random masking 50% pixels (**Bold**: best, under: second best, green: performance increase, red: performance decrease)

Methods	DAVIS				REDS			
	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
VRT (Liang et al. 2024)	22.98	0.576	0.459	2.35	22.56	0.593	0.465	2.392
DeBlurGANv2 (Kupyn et al. 2019)	24.32	0.649	0.371	1.734	24.11	0.677	0.368	1.722
Stripformer (Tsai et al. 2022)	24.07	0.612	0.381	2.218	24.72	0.699	0.343	1.797
ID-Blau (Wu et al. 2024)	23.07	0.589	0.397	2.623	23.84	0.654	0.359	2.198
SVI (Kwon and Ye 2024-10-04)	12.57	0.259	0.672	26.366	13.61	0.287	0.656	19.592
VISION-XL (Kwon and Ye 2024b)	19.08	0.454	0.539	8.421	17.27	0.410	0.567	9.157
VISION-base (Kwon and Ye 2024b)	12.81	0.290	0.699	26.359	13.98	0.320	0.684	21.882
Ours	**31.70**	**0.889**	**0.211**	**0.443**	**30.33**	**0.873**	**0.250**	**0.543**
Ours vs. Best compe.	+7.38	+0.240	-0.160	-1.291	+5.61	+0.174	-0.093	-1.179

Table 4: (Spatial task) Quantitative comparisons for video **motion debluring** (**Bold**: best, under: second best, green: performance increase, red: performance decrease)

Methods	DAVIS				REDS			
	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
VRT (Liang et al. 2024)	21.07	0.594	0.418	3.383	19.80	0.552	0.455	4.519
DeBlurGANv2 (Kupyn et al. 2019)	20.89	0.586	0.414	3.514	19.58	0.544	0.446	4.734
Stripformer (Tsai et al. 2022)	20.75	0.565	0.411	3.638	19.49	0.524	0.453	4.893
ID-Blau (Wu et al. 2024)	18.42	0.491	0.476	6.128	17.73	0.467	0.515	6.968
SVI (Kwon and Ye 2024-10-04)	27.63	0.766	0.151	0.901	26.14	0.740	0.162	1.119
VISION-XL (Kwon and Ye 2024b)	31.79	0.908	0.125	0.501	30.21	0.889	0.136	0.607
VISION-base (Kwon and Ye 2024b)	27.66	0.767	0.151	0.897	26.17	0.741	0.162	1.115
Ours	**32.965**	**0.948**	**0.104**	**0.317**	**33.307**	**0.952**	**0.099**	**0.362**
Ours vs. Best compe.	+1.17	+0.040	-0.021	-0.184	+3.09	+0.063	-0.037	-0.245

Table 5: (Temporal task) Quantitative comparisons for video **temporal deconvolution** (**Bold**: best, under: second best, green: performance increase, red: performance decrease)

As shown in Tables 2 to 6, our proposed method consistently outperforms existing ones across all VR tasks for almost all metrics on both the DAVIS and REDS datasets. We achieve consistent PSNR gains, ranging from 1–1.3dB in super-resolution, to the remarkable 6–8dB in combined degradation tasks, confirming our method's effectiveness across levels of degradation complexity. Most notably, our method

Methods	DAVIS				REDS			
	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
VRT (Liang et al. 2024)	19.76	0.48	0.593	4.044	18.71	0.445	0.638	5.325
DeBlurGANv2 (Kupyn et al. 2019)	19.91	0.496	0.556	3.921	18.89	0.460	0.600	5.219
Stripformer (Tsai et al. 2022)	20.00	0.503	0.548	3.879	18.88	0.458	0.588	5.224
ID-Blau (Wu et al. 2024)	19.76	0.489	0.554	4.060	18.73	0.450	0.593	5.418
SVI (Kwon and Ye 2024-10-04)	12.21	0.260	0.705	27.955	12.27	0.255	0.711	28.375
VISION-XL (Kwon and Ye 2024b)	15.94	0.346	0.627	15.220	16.70	0.393	0.637	12.675
VISION-base (Kwon and Ye 2024b)	12.17	0.262	0.727	30.325	11.92	0.269	0.732	30.436
Ours	**26.97**	**0.778**	**0.343**	**0.852**	**26.61**	**0.763**	**0.381**	**0.982**
Ours vs. Best compe.	+6.97	+0.275	-0.205	-3.027	+7.72	+0.303	-0.207	-4.237

Table 6: (Spatio-temporal task) Quantitative comparisons for video **temporal deconvolution with spatial deblurring** (**Bold**: best, under: second best, green: performance increase, red: performance decrease)

delivers substantial improvements in temporal consistency, with Warping Error (WE) reductions ranging from 0.18–0.26 in simpler tasks and dramatic 3–4.2 in combined temporal deconvolution with spatial deblurring. Similar performance trends are also observed in the SPMCS and UDM10 datasets, with detailed results provided in (Wang et al. 2025).

In particular, here, the test distribution may differ from the original training distribution for both supervised and zero-shot methods, potentially explaining the noticeable performance degradation compared to their originally reported results. Although we include supervised methods just for reference given their very different setting compared to the zero-shot approach we focus on, their relatively poor performance highlights their limited generalizability when confronted with real-world distribution shifts. For zero-shot competitors, our method consistently outperforms them by large margins in both spatial metrics and, most significantly, temporal consistency. These results validate the effectiveness of our proposed formulation in Eq. (12) and our hierarchical framework for bilevel temporal consistency described in Section 3.2.

A particularly interesting observation is that CG-based methods (Kwon and Ye 2024-10-04,b) perform *uniformly* poorly on motion deblur-related tasks, as is evident in Tables 4 and 6. This observation is consistent with our theoretical analysis in Section 2, which highlighted that *CG requires the degradation operator $\mathcal{A}$ be known, symmetric and positive definite* (Shewchuk 1994). When CG deals with VR tasks that violate these mathematical requirements—such as motion deblurring—numerical issues arise, leading to poor performance. This limitation comprises a critical weakness in the SOTA CG-based methods, whereas our method is versatile and remains effective for different degradation scenarios.

4.2 Ablation studies

Effects of noise prior and progressive warping Table 7 presents the quantitative results of our ablation studies on the DAVIS dataset for video super-resolution. The integration of noise prior significantly improves the baseline model performance in terms of both PSNR and SSIM, while substantially reducing WE, signifying enhanced semantic-level temporal consistency. Progressive warping yields even stronger results with marked improvements in all metrics compared to the baseline, particularly in WE, indicating superior pixel-level temporal consistency. Notably, combining both components produces the optimal configuration, with our complete model outperforming an SOTA method (Kwon and Ye 2024b) in all metrics. These results confirm that the noise prior and progressive warping mechanisms complement each other in promoting bilevel temporal consistency.

Effect of controllable residuals We also study how the learning rate for residuals (LR_r) and the radius of residual balls (ie, controlling the magnitude of residuals) affect the performance. Intuitively, the residuals should not be too large to allow substantial deviation from the image manifold, and not be too small to limit their powers in modeling reasonable framewise deviation from the shared frame. This intuition is confirmed by our data in Table 8: a restrictive radius (0.1) prevents adequate motion learning, whereas a moderate radius (1.0) allows effective discrepancy modeling; similarly, learning rate calibration is critical—a high rate (0.01) causes residual learning to dominate and downplays the influence from the DM prior, while an insufficient rate (0.0001) limits framewise adaptation. The optimal configuration (LR_r=0.001, radius=1.0) achieves the best performance by balancing these competing factors.

Method	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
SOTA (Kwon and Ye 2024b)	26.03	0.717	0.339	1.411
Base	24.70	0.612	0.366	1.398
Base with noise prior	26.09	0.703	0.410	1.057
Base with warping	27.14	0.736	**0.301**	0.943
Base with both	**27.95**	**0.790**	0.321	**0.725**

Table 7: Ablation study on essential components for bilevel temporal consistency, performed on DAVIS dataset for video super-resolution $\times 4$. (**Bold**: best, under: second best)

LR_r	Radius	PSNR↑	SSIM↑	LPIPS↓	WE(10^{-2})↓
0.01	0.1	24.36	0.654	0.436	1.557
	1.0	27.19	0.741	0.402	0.804
	10.0	27.16	0.739	0.405	0.806
0.001	0.1	25.74	0.719	0.374	1.161
	1.0	**27.95**	**0.790**	**0.321**	**0.725**
	10.0	27.91	0.789	0.324	0.737
0.0001	0.1	25.11	0.677	0.389	1.298
	1.0	26.17	0.709	0.366	1.096
	10.0	26.17	0.710	0.369	1.089

Table 8: Ablation study on the trainable residuals, performed on DAVIS dataset for video super-resolution $\times 4$.

5 Discussion

In this paper, we focus on solving video restoration (VR) problems using pre-trained image DMs. We systematically address three key challenges: approximation errors through a novel MAP framework that reparameterizes frames directly in the seed space of LDMs; temporal inconsistency via a hierarchical bilevel consistency strategy; and high computational and memory demands through reformulation and low-rank decomposition. Comprehensive experiments show that our method significantly improves over state-of-the-art methods across various VR tasks without task-specific training, in terms of both frame quality and temporal consistency. As for limitations, our work remains primarily empirical, and we leave a solid theoretical understanding for future research.

Acknowledgment

Part of the work was done when Wang H. was interning at Amazon.com, Inc. during the summer of 2024. Wang H. and Sun J. are also partially supported by a UMN DSI Seed Grant. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported in this article.

References

Cao, C.; Yue, H.; Liu, X.; and Yang, J. 2024. Zero-shot Video Restoration and Enhancement Using Pre-Trained Image Diffusion Model. *arXiv.org*.

Chan, K. C. K.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Investigating Tradeoffs in Real-World Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5962–5971.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022-09-29. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*.

Chung, H.; Lee, S.; and Ye, J. C. 2023-10-13. Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems. In *The Twelfth International Conference on Learning Representations*.

Coleman, R. 2012. *Calculus on normed vector spaces*. Springer Science & Business Media.

Daras, G.; Nie, W.; Kreis, K.; Dimakis, A.; Mardani, M.; Kovachki, N. B.; and Vahdat, A. 2024. Warped Diffusion: Solving Video Inverse Problems with Image Diffusion Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9935–9946.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hyvärinen, A. 2005. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24): 695–709.

Kupyn, O.; Martyniuk, T.; Wu, J.; and Wang, Z. 2019. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8878–8887.

Kwon, T.; and Ye, J. C. 2024-10-04. Solving Video Inverse Problems Using Image Diffusion Models. In *The Thirteenth International Conference on Learning Representations*.

Kwon, T.; and Ye, J. C. 2024b. VISION-XL: High Definition Video Inverse Problem Solver using Latent Image Diffusion Models.

Li, T.; Wang, H.; Zhuang, Z.; and Sun, J. 2023a. Deep Random Projector: Accelerated Deep Image Prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 18176–18185. IEEE.

Li, T.; Zhuang, Z.; Wang, H.; and Sun, J. 2023b. Random Projector: Efficient Deep Image Prior. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, 1–5. IEEE.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2024. VRT: A Video Restoration Transformer. *IEEE Transactions on Image Processing*, 33: 2171–2182.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Meister, S.; Hur, J.; and Roth, S. 2018-04-27. UnFlow: Unsupervised Learning of Optical Flow With a Bidirectional Census Loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Number: 1.

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study.

Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C. C.; and Luo, P. 2022-11. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7474–7489. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023-10-13. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shewchuk, J. 1994. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain.

Song, J.; Meng, C.; and Ermon, S. 2020-10-02. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8934–8943.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Computer Vision – ECCV 2020*, 402–419. Springer, Cham. ISBN 978-3-030-58536-5. ISSN: 1611-3349.

Tsai, F.-J.; Peng, Y.-T.; Lin, Y.-Y.; Tsai, C.-C.; and Lin, C.-W. 2022. Stripformer: Strip Transformer for Fast Image Deblurring. In *Computer Vision – ECCV 2022*, 146–162. Springer, Cham. ISBN 978-3-031-19800-7. ISSN: 1611-3349.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep Image Prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Wang, H.; Liu, Y.; Liu, H.; Wang, C.-C.; Guo, Y.; Li, H.; Wang, B.; and Sun, J. 2025. Temporal-Consistent Video Restoration with Pre-trained Diffusion Models.

Wang, H.; Zhang, X.; Li, T.; Wan, Y.; Chen, T.; and Sun, J. 2024. DMPlug: A Plug-in Method for Solving Inverse Problems with Diffusion Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C. K.; and Loy, C. C. 2024-12-01. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *International Journal of Computer Vision*, 132(12): 5929–5949. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 12 Publisher: Springer US.

Wu, J.-H.; Tsai, F.-J.; Peng, Y.-T.; Tsai, C.-C.; Lin, C.-W.; and Lin, Y.-Y. 2024. ID-Blau: Image Deblurring by Implicit Diffusion-based reBLurring AUgmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25847–25856.

Xu, R.; Li, X.; Zhou, B.; and Loy, C. C. 2019. Deep Flow-Guided Video Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 3723–3732. Computer Vision Foundation / IEEE.

Yeh, C.; Lin, C.-Y.; Wang, Z.; Hsiao, C.-W.; Chen, T.-H.; Shiu, H.-S.; and Liu, Y.-L. 2024-10-04. DiffIR2VR-Zero: Zero-Shot Video Restoration with Diffusion-based Image Restoration Models.

Zhong, Z.; Gao, Y.; Zheng, Y.; and Zheng, B. 2020. Efficient Spatio-Temporal Recurrent Neural Network for Video Deblurring. In *Computer Vision – ECCV 2020*, 191–207. Springer, Cham. ISBN 978-3-030-58539-6. ISSN: 1611-3349.

Zhou, S.; Li, C.; Chan, K. C. K.; and Loy, C. C. 2023. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10477–10486.

Zhou, S.; Yang, P.; Wang, J.; Luo, Y.; and Loy, C. C. 2024. Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2535–2545.

Zhuang, Z.; Li, T.; Wang, H.; and Sun, J. 2024-02-01. Blind Image Deblurring with Unknown Kernel Size and Substantial Noise. *International Journal of Computer Vision*, 132(2): 319–348. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 2 Publisher: Springer US.