# Zero-Shot Video Restoration and Enhancement Using Pre-Trained Image Diffusion Model

**Cong Cao[1], Huanjing Yue[1*], Xin Liu[2], Jingyu Yang[1]**

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2]Computer Vision and Pattern Recognition Laboratory, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland
caocong_123@tju.edu.cn, huanjing.yue@tju.edu.cn, linuxsino@gmail.com, yjy@tju.edu.cn

## Abstract

Diffusion-based zero-shot image restoration and enhancement models have achieved great success in various tasks of image restoration and enhancement. However, directly applying them to video restoration and enhancement results in severe temporal flickering artifacts. In this paper, we propose the first framework for zero-shot video restoration and enhancement based on the pre-trained image diffusion model. By replacing the spatial self-attention layer with the proposed short-long-range (SLR) temporal attention layer, the pre-trained image diffusion model can take advantage of the temporal correlation between frames. We further propose temporal consistency guidance, spatial-temporal noise sharing, and an early stopping sampling strategy to improve temporally consistent sampling. Our method is a plug-and-play module that can be inserted into any diffusion-based image restoration or enhancement methods to further improve their performance. Experimental results demonstrate the superiority of our proposed method.

**Code** — https://github.com/cao-cong/ZVRD

## Introduction

Recently, Denoising Diffusion Probabilistic Models (DDPMs) (Dhariwal and Nichol 2021) have demonstrated advanced generative capabilities surpassing those of GANs, inspiring further exploration of restoration and enhancement methods based on diffusion models. Different from using supervised learning and diffusion framework to train models for specific restoration and enhancement tasks (Saharia et al. 2022; Yin et al. 2023), the works in (Song and Ermon 2019; Chung et al. 2022; Fei et al. 2023; Shi and Liu 2024) employ a pre-trained image diffusion model for universal zero-shot image restoration and enhancement. These methods constrain the content between generated results in the reverse diffusion process and degraded images. However, due to the absence of temporal modeling in pre-trained image diffusion models, although these methods have shown promising results in image restoration and enhancement, their direct application to video restoration and enhancement can lead to significant temporal flickering.

With the emergence of powerful pre-trained text-to-image diffusion models, such as Stable Diffusion (Rombach et al. 2022), using off-the-shelf text-to-image diffusion model for zero-shot video editing has garnered increasing attention (Wu et al. 2023; Yang et al. 2023). To generate temporally consistent edited video, the motion information from the original video is typically utilized to design various temporal modules (Cong et al. 2023). However, predicting motion becomes more challenging when dealing with video restoration and enhancement tasks since input videos suffering from various degradations. In order to address this issue, we propose Short-Long-Range (SLR) temporal attention which consists cross-neighbour-frame attention and self-corrected trajectory attention. The cross-neighbor-frame attention implicitly models short-range temporal correlation without explicitly estimating motion, while the self-corrected trajectory attention compensates for inaccurate explicit motion estimation to capture long-range temporal correlation. The explicitly estimated motion information is utilized to construct guidance for pixel-level temporal consistency, which is a complementary of semantic-level consistency guidance. We observe that temporal flickering is mainly caused by inherent stochasticity in the diffusion model. Therefore, we introduce spatial-temporal noise sharing to mitigate this stochasticity effect. Additionally, we propose an early stopping sampling strategy since flicking details are usually generated during sampling in the later stage.

In summary, there are mainly three contributions in this work. First, we propose the first framework for Zero-shot Video Restoration and enhancement using a pre-trained image Diffusion model (ZVRD). Second, we propose SLR temporal attention, temporal consistency guidance, spatial-temporal noise sharing, and early stopping sampling strategy to maintain temporal consistency during video restoration and enhancement. Third, extensive experiments demonstrate the effectiveness of our method.

## Related Works

### Diffusion-Based Zero-Shot Image Restoration and Enhancement

The success of diffusion-based generative models has enlightened diffusion-based image restoration and enhancement methods. These methods can be divided into two cate-

*Corresponding author

gories. One category is designed for each specific task and utilizes paired data for supervised training (Saharia et al. 2022; Yin et al. 2023). The other category is a universal zero-shot method for different image restoration tasks based on a pre-trained image diffusion model (Song and Ermon 2019; Wang, Yu, and Zhang 2022; Chung et al. 2022; Fei et al. 2023; Shi and Liu 2024). Zero-shot methods utilize a pre-trained off-the-shelf diffusion model as the generative prior, which requires no additional training. The key to zero-shot methods is to constrain the result in the reverse diffusion process to have consistent content as degraded images. DDNM (Wang, Yu, and Zhang 2022) refines only the null-space contents during the reverse diffusion process to preserve content consistency. DPS (Chung et al. 2022) extends diffusion solvers to efficiently handle general noisy non-linear inverse problems via approximation of the posterior sampling. GDP (Fei et al. 2023) applies different loss functions between result and degraded image, and guides the reverse diffusion process with gradient. But these methods are designed for image restoration problems, there exists severe temporal flickering when applied to degraded videos.

### Diffusion-Based Zero-Shot Video Editing

Along with the development of powerful pre-trained text-to-image diffusion models, such as Stable Diffusion (Rombach et al. 2022), diffusion-based zero-shot video editing has gained increasing attention, which utilizes the off-the-shelf text-to-image diffusion model and mainly solves the temporal consistency problem. FateZero (Qi et al. 2023) follows Prompt-to-Prompt (Hertz et al. 2022) and fuses the attention maps to preserve the motion and structure consistency. Text2Video-Zero (Khachatryan et al. 2023) proposes cross-frame attention for better temporal consistency. (Cong et al. 2023; Yang et al. 2024a) propose optical flow-guided attention and spatial-temporal correspondence-guided attention, respectively. Inspired by these works, we propose to use the pre-trained image diffusion model for zero-shot video restoration and enhancement. Different from these zero-shot video editing methods that use Stable Diffusion, we use an unconditional image diffusion model (Dhariwal and Nichol 2021) pre-trained on ImageNet, which is commonly used in zero-shot image restoration.

### Video Restoration and Enhancement

The existing video restoration methods need to be trained for every single task. Temporal mutual self-attention is proposed to exploit temporal information in video super-resolution and video deblurring (Liang et al. 2024). The work in (Yang et al. 2024b) explores colors of exemplars and utilizes them to help video colorization by temporal feature fusion with the guidance of semantic image prior. The work in (Zheng and Gupta 2022) explores zero-shot image (video) enhancement by utilizing non-reference loss functions, but still needs training on unpaired data with diverse illumination conditions. Different from the above methods, our method is a training-free zero-shot method, which is universal to different restoration and enhancement tasks. Recently, the work in (Yeh et al. 2024) adapts image restoration model for video restoration without training. But it is based on image latent diffusion model which has been specifically trained for restoration in a supervised manner. However, most zero-shot image restoration diffusion methods are based on (Dhariwal and Nichol 2021). For the U-Net of (Dhariwal and Nichol 2021), the attention module only exists on the features with $32 \times 32$ and lower resolution, the higher resolution features can also cause the temporal inconsistency of output. Therefore, besides the SLR temporal attention, we further propose temporal consistency guidance and spatial-temporal noise sharing to solve this problem. The token merging of the attention module in (Yeh et al. 2024) is not enough to maintain the temporal consistency of (Dhariwal and Nichol 2021). Our method can be applied to both zero-shot and supervised diffusion-based image restoration (enhancement) models for video restoration (enhancement).

## Background

Diffusion models transform target data distribution into simple noise distribution and recover data from noise. We follow the diffusion model defined in denoising diffusion probabilistic models (DDPM) (Ho, Jain, and Abbeel 2020). DDPM defines a T-step forward process and a T-step reverse process. The forward process adds random noise to data step by step, while the reverse process constructs target data samples step by step.

### The Reverse Diffusion Process

The Reverse diffusion Process is a Markov chain that denoises a sampled Gaussian noise to a clean image step by step. Starting from noise $x_T \sim \mathcal{N}(0, \boldsymbol{I})$, the reverse process from latent $\boldsymbol{x}_T$ to clean data $\boldsymbol{x}_0$ is defined as:

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right), \Sigma_\theta \boldsymbol{I}\right) \quad (1)$$

The mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)$ is the target we want to estimate by a neural network $\boldsymbol{\theta}$. The variance $\Sigma_\theta$ can be either time-dependent constants (Ho, Jain, and Abbeel 2020) or learnable parameters (Nichol and Dhariwal 2021). The reverse process yields the previous state $\boldsymbol{x}_{t-1}$ from the current state $\boldsymbol{x}_t$:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right) + \sqrt{\Sigma_\theta}\boldsymbol{z} \quad (2)$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$. In practice, $\tilde{\boldsymbol{x}}_0$ is usually predicted from $\boldsymbol{x}_t$, then $\boldsymbol{x}_{t-1}$ is sampled using both $\tilde{\boldsymbol{x}}_0$ and $\boldsymbol{x}_t$, where $\tilde{\boldsymbol{x}}_0$ is computed as:

$$\tilde{\boldsymbol{x}}_0 = \frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(\boldsymbol{x}_t, t\right)}{\sqrt{\bar{\alpha}_t}} \quad (3)$$

and $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

## Method

### Overall Framework

Given a degraded video with $N$ frames $\{I_i\}_{i=0}^N$, our goal is to restore or enhance it to a normal-light clean video $\{\bar{I}_i\}_{i=0}^N$. Our method leverages a pre-trained image diffusion model (Dhariwal and Nichol 2021) for zero-shot video restoration and enhancement. The work in (Dhariwal and Nichol 2021)

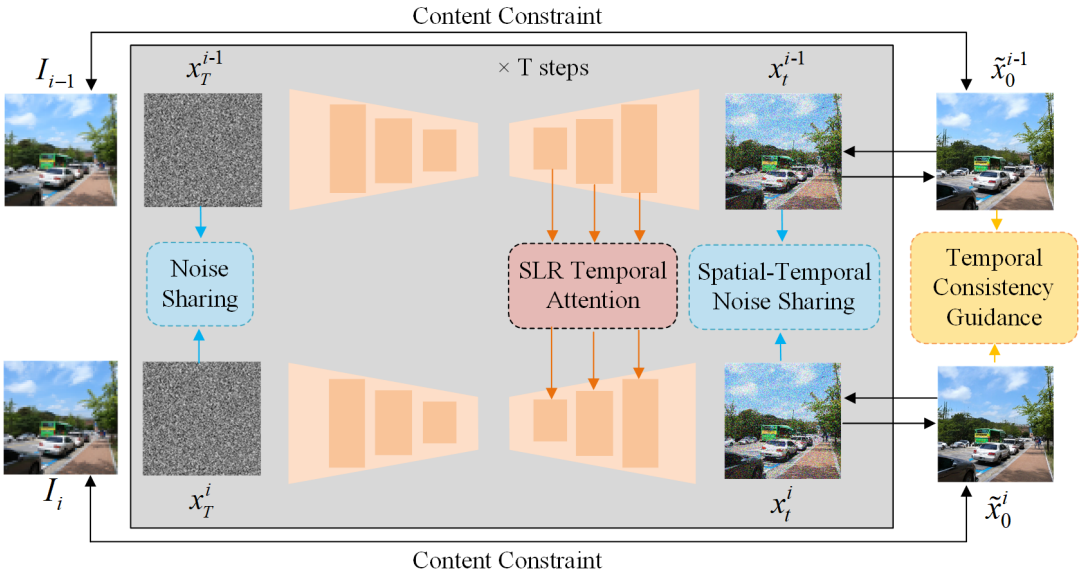


Figure 1: Framework of the proposed zero-shot video restoration and enhancement.

utilizes a U-Net constructed by layers of 2D convolutional residual blocks and spatial self-attention blocks. We replace all $3 \times 3$ 2D convolutions with inflated $1 \times 3 \times 3$ 3D convolutions to process video. For better temporal consistency, we propose SLR temporal attention, temporal consistency guidance, spatial-temporal noise sharing, and early stopping sampling strategy, the framework is illustrated in Fig. 1. It's worth noting that our method is a plug-and-play module, meaning it can be easily incorporated into any diffusion-based image restoration or enhancement method.

## SLR Temporal Attention

We propose SLR temporal attention to strengthen the temporal consistency of video restoration and enhancement results. Since the decoder layers are less noisy than the encoder layer in the sampling, we replace spatial self-attention layers in the U-Net decoder with our SLR temporal attention layer, which consists of two modules: cross-neighbor-frame attention and self-corrected trajectory attention, as shown in Fig. 2.

**Cross-Neighbor-Frame Attention** For the spatial self-attention layer, the query, key, and value $Q$, $K$, $V$ are obtained by linear projection of the feature $v_i$ from $I_i$, the corresponding self-attention output is produced by $\textit{Self\_Attn}(Q,K,V) = \textit{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$ with

$$Q = W^Q v_i, K = W^K v_i, V = W^V v_i, \tag{4}$$

where $W^Q$, $W^K$, $W^V$ are pre-trained matrices that project the inputs to $Q$, $K$, $V$ respectively. Cross-frame attention uses the key $K'$ and value $V'$ from other frames, which has been widely used for zero-shot video editing. For video editing, besides the previous frame, the first frame is also used to maintain global coherence in terms of generated content. However, for video restoration, we find that the bidirectional neighbor frames are more suitable for maintaining the temporal consistency. The cross-neighbor-frame attention output is produced by $\textit{CrossNeighFrame\_Attn}(Q,K',V') = \textit{Softmax}(\frac{QK'^T}{\sqrt{d}}) \cdot V'$ with

$$Q = W^Q v_i, K' = W^K(v_{i-1}||v_{i+1}), V' = W^V(v_{i-1}||v_{i+1}), \tag{5}$$

where $||$ represents concatenation. To reduce the computation cost, we only leverage the previous and next neighbor frame to capture the short-range temporal correlation between frames, and utilize the following self-corrected trajectory attention to capture the long-range temporal correlation. As shown in Fig. 2, the colored $\times$ denotes the position of query in attention, and the colored square denotes the position of key and value. The output of the cross-neighbor-frame serves as the query $\hat{Q}$ for self-corrected trajectory attention.

**Self-Corrected Trajectory Attention** To further improve the temporal consistency, optical flow-guided attention (Cong et al. 2023) was proposed, which samples patch trajectories according to optical flow and performs the attention on the patch embeddings in the same trajectory. However, inaccurate flow-based trajectories from inaccurate optical flow limit the performance. Especially for video restoration and enhancement, the input degraded frames damage the optical flow calculation. Performing restoration twice and calculating the optical flow after the first restoration without flow-guided attention results in a doubling of the inference time, and the gap between two restoration processes also influences the suitability of the optical flow. In view of this, we propose a self-corrected strategy to progressively correct the trajectories in the sampling process. For the U-Net in (Dhariwal and Nichol 2021), the original spatial self-attention is applied to the features with resolution $32 \times 32$, $16 \times 16$, and $8 \times 8$. We inject our self-corrected trajectory attention to the largest resolution $32 \times 32$. In the $t$ step of sampling, for each pixel

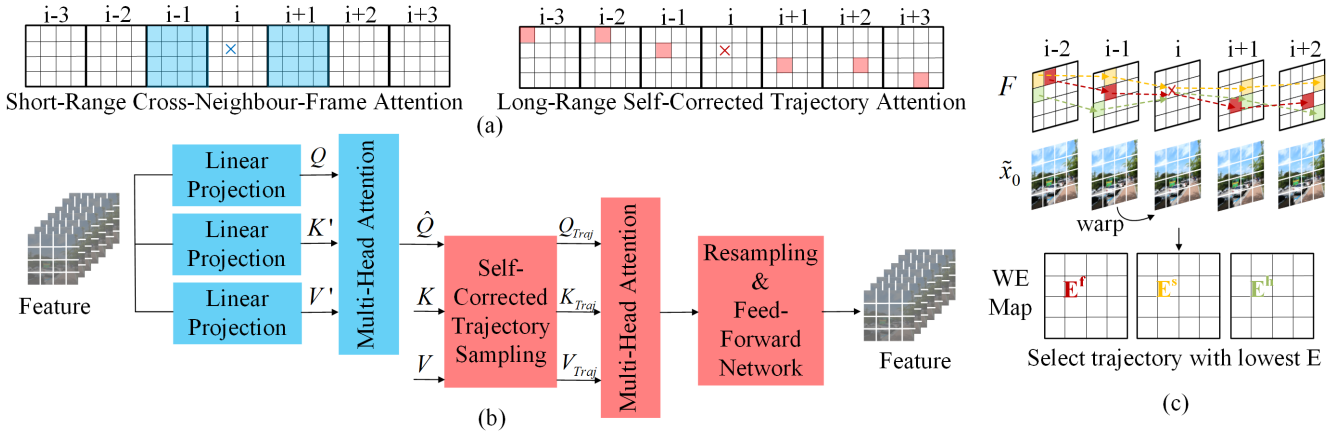


Figure 2: Architecture of the proposed SLR Temporal Attention. (a) The two modules in SLR temporal attention: cross-neighbor-frame attention and self-corrected trajectory attention, focus on short-range and long-range temporal correlation between frames, respectively. (b) The cross-neighbor-frame attention is applied first, and its output serves as the query for the self-corrected trajectory attention. (c) The procedure of self-corrected trajectory sampling. The red, yellow, and green trajectories denote the flow-based, similarity-based, and historically-best trajectories, respectively.

in the feature map, its flow-based trajectory can be calculated from the downsampled optical flow. The optical flow is calculated on the $\tilde{\boldsymbol{x}}_0$ of $t+1$ step. The clean image $\tilde{\boldsymbol{x}}_0$ can be directly inferred when given $\boldsymbol{x}_t$ by the Eq. 3 in every timestep $t$. As $t$ decreases, $\tilde{\boldsymbol{x}}_0$ will have better quality, resulting in more accurate optical flow. In addition, we propose the similarity-based trajectory and historically-best trajectory to correct the flow-based trajectory.

Given the diffusion feature $F_i$ and $F_{i-1}$ of frame $I_i$ and $I_{i-1}$, the cosine similarity between pixel pairs $(p, q)$ in the feature can be formulated as

$$S(p,q) = \frac{F_i(p) \cdot F_{i-1}(q)}{||F_i(p)||\,||F_{i-1}(q)||}. \tag{6}$$

The similarity-based trajectory between frame $I_i$ and $I_{i-1}$ can be obtained from the pixel pairs with the highest similarity. We define the best trajectory for a pixel in the feature as the trajectory that can achieve the best temporal consistency on the corresponding patch of $\tilde{\boldsymbol{x}}_0$. For every timestep $t$, the historically-best trajectory is defined as the best trajectory at step $t+1$. The inaccurate flow-based trajectory can be compensated for by similarity-based trajectory and historically-best trajectory. Specifically, for each pixel in feature $F_i$ at step $t$, we compute $\tilde{\boldsymbol{x}}_0$ of three different trajectories, respectively. Each pixel corresponds to a $8\times8$ patch in $\tilde{\boldsymbol{x}}_0$, we warp the previous frame and compute the average warp error of this patch area. The trajectory with the lowest warp error serves as the final trajectory of this pixel at step $t$, i.e., the best trajectory at step $t$ and the historically-best trajectory for the step $t-1$. This procedure is shown in Fig. 2 (c). The trajectory attention output is produced by $\text{Traj\_Attn}(Q_{Traj}, K_{Traj}, V_{Traj}) = \text{Softmax}(\frac{Q_{Traj}K_{Traj}^T}{\sqrt{d}}) \cdot V_{Traj}$ with

$$Q_{Traj} = \hat{Q}[p], K_{Traj} = K[Traj{-}p], V_{Traj} = V[Traj{-}p]. \tag{7}$$

where [p] denotes the sampling value of pixel $p$, and [Traj-p] denotes the sampling values of pixels in the trajectory of $p$ except for $p$ itself. We find that the $\tilde{\boldsymbol{x}}_0$ is not always clean in the whole sampling process. At the beginning of sampling, $\tilde{\boldsymbol{x}}_0$ has a lower signal-to-noise ratio, where the image contents are unrecognizable and have a lot of noise. In the middle part of sampling, $\tilde{\boldsymbol{x}}_0$ has smooth content which cannot be used to compute precise optical flow. Only in the second half of sampling does the diffusion model slowly generate rich content and details, which are suitable for computing precise optical flow. In practice, we only apply self-corrected trajectory attention after the current diffusion step $t{<}T_{TA}$, where $T_{TA}$ is set to 100 for the GDP backbone. We utilize RAFT (Teed and Deng 2020; Jeong and Ye 2023; Cong et al. 2023) to calculate optical flow, and utilize forward-backward consistency check to generate occlusion mask for warp error (Lai et al. 2018).

## Temporal Consistency Guidance

Since the attention module only exists on the features with $32 \times 32$ and lower resolution, the higher resolution ($64 \times 64$, $128 \times 128$, $256 \times 256$) features can also cause the temporal inconsistency of output. Therefore we propose the temporal consistency guidance to directly constrain the final output. The temporal consistency guidance is categorized into pixel-level consistency and semantic-level consistency. For pixel-level consistency, we compute the optical flow and occlusion mask between the $\tilde{\boldsymbol{x}}_0$ of frame $I_i$ and $I_{i-1}$ (denoted by $\tilde{\boldsymbol{x}}_0^i$ and $\tilde{\boldsymbol{x}}_0^{i-1}$) in the step $t+1$, then constrain $\tilde{\boldsymbol{x}}_0^i$ and $\tilde{\boldsymbol{x}}_0^{i-1}$ in step $t$ with pixel-level consistency

$$\mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{PC} = \sum_{i=0}^{N} M_i \left\| \tilde{\boldsymbol{x}}_0^i - warp(\tilde{\boldsymbol{x}}_0^{i-1}, f_i) \right\|_1 \tag{8}$$

where $M_i$ is the predicted occlusion mask, $f_i$ is the predicted optical flow. For semantic-level consistency, the neighbour

frames should have similar sematic information. We utilize the image encoder of CLIP to extract the embedding $E^i$ and $E^{i-1}$ of $\tilde{\boldsymbol{x}}_0^i$ and $\tilde{\boldsymbol{x}}_0^{i-1}$, the semantic-level consistency can be formulated as

$$\mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{SC} = 1 - \frac{E_i \cdot E_{i-1}}{||E_i||\,||E_{i-1}||}. \tag{9}$$

the totally temporal consistency can be formulated as

$$\mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{TC} = \mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{PC} + \gamma \mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{SC} \tag{10}$$

Then we apply gradient guidance (Fei et al. 2023) to guide the sampling process. Specifically, we sample $\boldsymbol{x}_{t-1}$ by $\mathcal{N}\left(\mu + s\nabla_{\tilde{\boldsymbol{x}}_0}\mathcal{L}_{\tilde{\boldsymbol{x}}_0}^{TC}, \sigma^2\right)$, $s$ is gradient scale. Since only in the second half of sampling, $\tilde{\boldsymbol{x}}_0$ is suitable to compute optical flow, we apply pixel-level consistency guidance after the current diffusion step $t{<}T_{TC}$, which is set to 300 for the GDP backbone. We apply semantic-level consistency guidance throughout the entire sampling process, compensating for the absence of pixel-level consistency guidance in the early steps.

## Spatial-Temporal Noise Sharing

Recently, (Chen et al. 2024) demonstrates that the denoising process plays an important role in the denoising diffusion model. Actually, the noise in the sampling process controls the final generated color and details. For the same degraded frame, different noise $\boldsymbol{x}_T$ and $\boldsymbol{z}$ in the reverse diffusion process will lead to different colors and details in the result. For better temporal consistency, we propose to share the same $\boldsymbol{x}_T$ and $\boldsymbol{z}$ in Eq. 2 between all frames, which encourages the diffusion model to generate the same details in the static areas. We used the predicted optical flow and occlusion mask to blend the $\boldsymbol{z}$ of degraded frame $I_i$ and $I_{i-1}$, which are denoted by the $\boldsymbol{z}^i$ and $\boldsymbol{z}^{i-1}$. We propose to blend $\boldsymbol{z}$ rather than to blend $\tilde{\boldsymbol{x}}_0$, $\tilde{\boldsymbol{x}}_t$ or U-Net feature since the latter usually leads to motion ghost and unpleasant artifacts. The blending process can be formulated as

$$\boldsymbol{z}^i = M_i(\lambda \boldsymbol{z}^i + (1-\lambda)\boldsymbol{z}^{i-1}) + (1-M_i)\boldsymbol{z}^i \tag{11}$$

The blending process shares noise between the corresponding pixels in different frames, which encourages the diffusion model to generate the same details in these dynamic areas.

## Early Stopping Sampling Strategy

In the above sections, we find that $\boldsymbol{x}_0$ firstly reconstructs the low-frequency component of the frame, then reconstructs the high-frequency component in the sampling, the temporal flicker easily increases at the end of the reverse diffusion process. Besides, the real-world degraded images often suffer from noise. When enhancing low-light videos, the diffusion model reconstructs the high-frequency noise at the end of sampling, thereby reducing temporal consistency. We propose an early stopping sampling strategy, which stops sampling after $T_{ES}$, preventing $\boldsymbol{x}_0$ from reconstructing noise or unconsistent high-frequency details. We take the early stopping $\boldsymbol{x}_0$ as the final result.

| Methods | PSNR↑ | SSIM↑ | FID↓ | WE↓ | FS↑ | OFME↓ |
|---|---|---|---|---|---|---|
| VRT | **23.68** | **0.7434** | 157.87 | 0.4797 | 0.9858 | **0.1563** |
| DDNM | 23.46 | 0.6876 | 110.13 | 1.3103 | 0.9513 | 0.3212 |
| DDNM+ZVRD | 23.53 | 0.6925 | **106.84** | 0.5339 | 0.9754 | 0.2596 |
| GDP | 20.44 | 0.5252 | 171.59 | 4.0327 | 0.8950 | 4.3595 |
| GDP+ZVRD | 21.39 | 0.5843 | 167.44 | **0.4234** | **0.9885** | 0.9948 |

Table 1: Quantitative comparison with state-of-the-art methods for 4× video super-resolution. The best results are highlighted in bold and the second best results are underlined. WE is expressed as a percentage (%). VRT is a supervised method, the others are zero-shot methods.

| Methods | PSNR↑ | SSIM↑ | FID↓ | WE↓ | FS↑ | OFME↓ |
|---|---|---|---|---|---|---|
| BiSTNet | 23.67 | 0.9920 | 131.04 | 1.1752 | 0.9787 | **0.1213** |
| DDNM | 24.60 | 0.9932 | 123.29 | 2.4315 | 0.9371 | 0.7094 |
| DDNM+ZVRD | **24.86** | **0.9945** | **121.87** | 1.2992 | **0.9866** | 0.1386 |
| GDP | 24.58 | 0.9333 | 134.56 | 1.3125 | 0.9176 | 0.3658 |
| GDP+ZVRD | 24.64 | 0.9416 | 133.39 | **0.9208** | 0.9850 | 0.2875 |

Table 2: Quantitative comparison with state-of-the-art methods for video colorization. The best results are highlighted in bold and the second best results are underlined. WE is expressed as a percentage (%). BiSTNet is a supervised method, the others are zero-shot methods.

# Experiments

## Test Datasets

For video super-resolution, we collected 18 gt videos from commonly used test datasets REDS4 (Nah et al. 2019), Vid4 (Liu and Sun 2013), and UDM10 (Yi et al. 2019). For video deblurring, we collected 10 ground truth (GT) videos from the dataset REDS (Nah et al. 2019). For video denoising, we collected 15 GT videos from the commonly used test dataset Set8 (Tassano, Delon, and Veit 2020) and DAVIS (Pont-Tuset et al. 2017). For video inpainting, we collected 20 GT videos from the commonly used DAVIS (Pont-Tuset et al. 2017) dataset. For video colorization, we use the GT videos from the Videvo20 (Lai et al. 2018) dataset, which is one of the mainly used datasets for video colorization. We follow (Chung et al. 2022; Wang, Yu, and Zhang 2022; Fei et al. 2023) to apply linear degradation to GT videos to construct corresponding degraded videos for video super-resolution, deblurring, denoising, inpainting, and colorization respectively. For low-light video enhancement, we collected 10 paired low-normal videos from the DID dataset (Fu et al. 2023) which was captured in the real world. Due to the slow sampling speed of DDPM and a test video containing a lot of frames, we first center crop the frames along the shorter edge and then resize them to 256×256, which matches the image size of the diffusion model. Our method can be combined with a patch-based strategy in (Fei et al. 2023) to process any-size videos.


Figure 3: Visual quality comparison for video super-resolution. Zoom in for better observation.

| Methods | PSNR↑ | SSIM↑ | FID↓ | WE↓ | FS↑ | OFME↓ |
|---|---|---|---|---|---|---|
| FastLLVE | 12.76 | 0.6572 | 261.69 | 0.7236 | 0.9791 | 0.5306 |
| SGZ | 17.22 | 0.6576 | **49.49** | 0.4548 | 0.9904 | 0.3844 |
| GDP | 17.35 | 0.8072 | 62.05 | 0.6029 | 0.9827 | 0.3533 |
| GDP+ZVRD | **17.56** | **0.8237** | 60.54 | **0.3352** | **0.9910** | **0.3181** |

Table 3: Quantitative comparison with state-of-the-art methods for low-light video enhancement. The best results are highlighted in bold and the second best results are underlined. WE is expressed as a percentage (%). FastLLVE is a supervised method, the others are zero-shot methods.

| Methods | PSNR↑ | SSIM↑ | FID↓ | WE↓ | FS↑ | OFME↓ |
|---|---|---|---|---|---|---|
| DiffBIR | 24.47 | 0.6727 | 32.03 | 0.6907 | 0.9825 | 0.1328 |
| DiffIR2VR | 24.49 | 0.6718 | 30.41 | 0.6818 | 0.9820 | 0.1312 |
| DiffBIR+ZVRD | **24.55** | **0.6799** | **30.12** | **0.4923** | **0.9898** | **0.1145** |

Table 4: Quantitative comparison with state-of-the-art methods for 4× blind video super-resolution on the DAVIS dataset. The best results are highlighted in bold and the second best results are underlined. WE is expressed as a percentage (%).

## Comparison with State-of-the-art Methods

We utilize six metrics to evaluate the restoration and enhancement quality. Besides the commonly used metrics PSNR, SSIM, and FID, we utilize Warping Error (WE) (Lai et al. 2018), Frame Similarity (FS) (Wu et al. 2023; Chen et al. 2023; Qi et al. 2023), and optical flow map error (OFME) (Wang et al. 2024; Chen et al. 2023) to evaluate temporal consistency. In our supplementary file, we also provide the user study for temporal consistency evaluation. FS was introduced to assess semantic consistency between generated frames by calculating the similarities of CLIP embeddings of output video frames. OFME was introduced to measure the movement consistency in video synthesis and editing. We extend it to evaluate video restoration and enhancement by calculating the optical flow map error between restored/enhanced frames and ground truth frames. Since our method is a plug-and-play method, we choose three state-of-the-art zero-shot image restoration methods, namely DPS (Chung et al. 2022), DDNM (Wang, Yu, and Zhang 2022) and GDP (Fei et al. 2023) as our compared methods and backbones. We utilize their content constraints in our method and extend them for zero-shot video restoration, respectively. Besides the three backbones, we also compare with VRT (supervised training) (Liang et al. 2024) for video super-resolution and deblurring, FastDVDNet (supervised training) (Tassano, Delon, and Veit 2020) and UDVD (unsupervised training) (Sheth et al. 2021) for video denoising. For video inpainting, we compare with zero-shot image inpainting method RePaint (Lugmayr et al. 2022). For video colorization, we compare with the supervised method BiSTNet (Yang et al. 2024b). For low-light video enhancement, we compared with the supervised method FastLLVE (Li et al. 2023) and zero-shot video enhancement method SGZ (Zheng and Gupta 2022).

Tables 1-3 list the quantitative results for the video super-resolution, video colorization, and low-light video enhancement, respectively. It can be observed that by inserting our method in existing zero-shot image restoration methods (DDNM+ZVRD, GDP+ZVRD, DPS+ZVRD), the temporal consistency can be obviously improved. For 4× video super-resolution, on the basis of DDNM, the WE is decreased to nearly 1/3 of the original, and the FID is increased and outperforms the supervised method VRT. On the basis of GDP, the WE is decreased to about 1/10 of the original, which is better than VRT. Our method achieves nearly 1 dB gain for PSNR. For all tasks, our method can improve the performance in

most of the six metrics. For video colorization, our method can boost DDNM to outperform the supervised method BiST-Net in four metrics. For low-light video enhancement, our method can enhance GDP in five metrics, surpassing other methods. It demonstrates the effectiveness of our method. Due to the page limit, we give the quantitative results for the video deblurring, video denoising, and video inpainting in the supplementary material.

Figs. 3, 4 present the visual comparison results on the evaluation data for video super-resolution and low-light video enhancement, respectively. Due to the page limit, visual comparison results on more tasks are shown in the supplementary file. Fig. 3 presents the results of the four methods on the first and second frames of the video. For GDP, the details of the tree and bus are not consistent on the two frames, and the shape of the car is also obviously different. For DDNM, there are different contents on the window of the bus. Our method (DDNM+ZVRD, GDP+ZVRD) can restore temporal consistent results on both tree and bus. As shown in Fig. 4, GDP has different global light and different details on the table. Our method has better temporal consistency on global light and local details.

Besides the above linear restoration tasks and non-linear, blind enhancement task, our method can also be applied to blind restoration tasks with complex real-world degradation. Following the settings of DiffIR2VR (Yeh et al. 2024), we use DiffBIR as the backbone for blind video super-resolution and evaluate on DAVIS testing sets. Low-quality videos are generated using the degradation pipeline of RealBasicVSR. Our method achieves the best performance for all six metrics as shown in Table 4. Since (Yeh et al. 2024) relies on optical flow which is inaccurate and directly merges similar tokens of attention blocks between frames, they tend to generate blurry results. Our SLR temporal attention provides a softer way to solve the issue of temporal consistency. The self-corrected trajectory attention in SLR temporal attention can adaptly compensate for inaccurate optical flow through similarity-based trajectory and historically-best trajectory. Thus our method generates sharper results.

## Ablation Study

In this section, we perform an ablation study to demonstrate the effectiveness of the proposed SLR Temporal Attention, Temporal Consistency Guidance, Spatial-Temporal Noise Sharing, and Early Stopping Sampling Strategy. Take video super-resolution as an example, Table 5 lists the quantitative comparison results on evaluation data by adding these modules one by one. It can be observed that SLR Temporal Attention can bring 0.68 dB gain for PSNR, 2.7 gain for FID, nearly 1.75 gain for WE, and nearly 2.5 gain for OFME. When adding Temporal Consistency Guidance, WE is decreased by nearly 0.35, and FS is increased by 0.0433. Spatial-Temporal Noise Sharing can bring 0.27 dB gain for PSNR and reduce WE by nearly 1.5. It is ranked the second in terms of gain for the metric WE. Early Stop Sampling Strategy can further reduce the WE, and OFME and improve FS while keeping other metrics basically unchanged. Due to the page limit, we give a more detailed ablation study in the supplementary material.

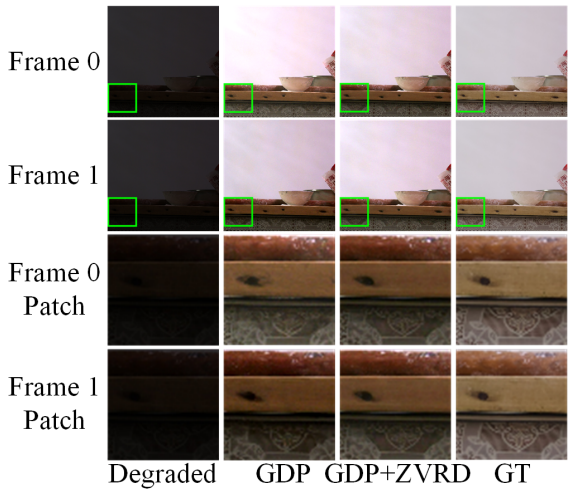


Figure 4: Visual quality comparison for low-light video enhancement. Zoom in for better observation.

| SLTA | × | ✓ | ✓ | ✓ | ✓ |
|---|---|---|---|---|---|
| TCG | × | × | ✓ | ✓ | ✓ |
| STNS | × | × | × | ✓ | ✓ |
| ESSS | × | × | × | × | ✓ |
| PSNR↑ | 20.44 | 21.12 | 21.15 | 21.42 | 21.39 |
| SSIM↑ | 0.5252 | 0.5611 | 0.5623 | 0.5847 | 0.5843 |
| FID↓ | 171.59 | 168.89 | 168.62 | 167.35 | 167.44 |
| WE↓ | 4.0327 | 2.2806 | 1.9281 | 0.4586 | 0.4234 |
| FS↑ | 0.8950 | 0.9220 | 0.9653 | 0.9867 | 0.9885 |
| OFME↓ | 4.3595 | 1.8654 | 1.3540 | 0.9972 | 0.9948 |

Table 5: Ablation study for SLR Temporal Attention (SLTA), temporal consistency guidance (TCG), spatial-temporal noise sharing (STNS) and early stopping sampling strategy (ESSS) on 4× video super-resolution task. WE is expressed as a percentage (%).

## Conclusion

In this paper, we propose the first framework for zero-shot video restoration and enhancement which uses a pretrained image diffusion model and is training-free. By replacing the spatial self-attention layer with the proposed SLR temporal attention layer, the pre-trained image diffusion model can utilize the temporal correlation between frames. To further strengthen the temporal consistency of results, we propose temporal consistency guidance, spatial-temporal noise sharing, and an early stopping sampling strategy. Experimental results demonstrate the superiority of the proposed method.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant 62472308, 62231018 and 62171309.

## References


Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*.

Chen, X.; Liu, Z.; Xie, S.; and He, K. 2024. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

Cong, Y.; Xu, M.; Simon, C.; Chen, S.; Ren, J.; Xie, Y.; Perez-Rua, J.-M.; Rosenhahn, B.; Xiang, T.; and He, S. 2023. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

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.

Fu, H.; Zheng, W.; Wang, X.; Wang, J.; Zhang, H.; and Ma, H. 2023. Dancing in the dark: A benchmark towards general low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12877–12886.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jeong, H.; and Ye, J. C. 2023. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 170–185.

Li, W.; Wu, G.; Wang, W.; Ren, P.; and Liu, X. 2023. Fastllve: Real-time low-light video enhancement with intensity-aware look-up table. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8134–8144.

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*.

Liu, C.; and Sun, D. 2013. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 346–360.

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 (CVPR)*.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

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

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*.

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.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Sheth, D. Y.; Mohan, S.; Vincent, J. L.; Manzorro, R.; Crozier, P. A.; Khapra, M. M.; Simoncelli, E. P.; and Fernandez-Granda, C. 2021. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1759–1768.

Shi, Z.; and Liu, R. 2024. Conditional Velocity Score Estimation for Image Restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 179–188.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Tassano, M.; Delon, J.; and Veit, T. 2020. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1354–1363.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2024. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7623–7633.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2024a. FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8703–8712.

Yang, Y.; Pan, J.; Peng, Z.; Du, X.; Tao, Z.; and Tang, J. 2024b. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yeh, C.-H.; Lin, C.-Y.; Wang, Z.; Hsiao, C.-W.; Chen, T.-H.; and Liu, Y.-L. 2024. DiffIR2VR-Zero: Zero-Shot Video Restoration with Diffusion-based Image Restoration Models. *arXiv preprint arXiv:2407.01519*.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3106–3115.

Yin, Y.; Xu, D.; Tan, C.; Liu, P.; Zhao, Y.; and Wei, Y. 2023. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8145–8156.

Zheng, S.; and Gupta, G. 2022. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, 581–590.