# TempCo-Painter: Temporal Consistency Enhanced Painter with Adaptive Diffusion Transformers for Long Video Inpainting

Ruohan Qi * and Tianhao Nian

Posted Date: 5 February 2026

doi: 10.20944/preprints202602.0440.v1

Keywords: video inpainting; diffusion transformers; spatio-temporal consistency; long videos; TempCo-Painter

 doi:10.20944/preprints202602.0440.v1



*Article*

# TempCo-Painter: Temporal Consistency Enhanced Painter with Adaptive Diffusion Transformers for Long Video Inpainting

**Ruohan Qi * and Tianhao Nian**

College of William and Mary, USA
* Correspondence: alberlucia.soarez@iscon.edu.br

**Abstract**

Video inpainting, a critical task in computer vision, aims to plausibly fill missing regions in video sequences while maintaining both spatial realism and robust spatio-temporal consistency. Current methods often struggle with ultra-long videos, highly dynamic occlusions, and achieving extreme coherence efficiently, leading to common artifacts. To address these challenges, we propose TempCo-Painter: Temporal Consistency Enhanced Painter with Adaptive Diffusion Transformers. Our novel framework leverages a specialized 3D-VAE for efficient latent space compression and introduces an innovative Adaptive Diffusion Transformer (ADiT). ADiT integrates hierarchical spatial-temporal attention, a motion-guided attention mechanism for accurate dynamic content restoration, and dynamic mask awareness for robust handling of diverse occlusions. An efficient Flow Matching scheduler further enables TempCo-Painter to generate high-quality results with minimal denoising steps. For processing arbitrarily long videos, we introduce an enhanced MultiDiffusion strategy featuring an adaptive sliding window and temporal smoothing regularization to ensure seamless global consistency. Extensive experiments demonstrate that TempCo-Painter achieves state-of-the-art performance on standard short video benchmarks, significantly outperforming existing methods in PSNR, SSIM, and notably reducing Video Frechet Inception Distance. Furthermore, it exhibits superior robustness and coherence on challenging minute-level long videos and complex mask scenarios, while maintaining high inference efficiency.

**Keywords:** video inpainting; diffusion transformers; spatio-temporal consistency; long videos; TempCo-Painter

## 1. Introduction

Video inpainting is a fundamental task in computer vision, aiming to synthesize plausible, coherent, and visually realistic content for missing or occluded regions (defined by a mask) within a video sequence. This technology holds significant value across numerous practical applications, including object removal from footage [1], extending video boundaries for video completion [2], and removing subtitles or watermarks (video decaptioning/watermarking removal) [3]. Furthermore, the advent of robust image watermarking and manipulation detection techniques [4–6] underscores the need for advanced inpainting methods capable of producing pristine, undetectable content.

The critical challenge in video inpainting lies not only in ensuring the *spatial realism* of the content within each individual frame but, more importantly, in maintaining excellent *spatial-temporal consistency* across the entire video sequence. Directly applying traditional image inpainting methods to videos often leads to undesirable artifacts such as flickering, jittering, or content discontinuity in the repaired regions over time, severely degrading the viewing experience [2].

 doi:10.20944/preprints202602.0440.v1



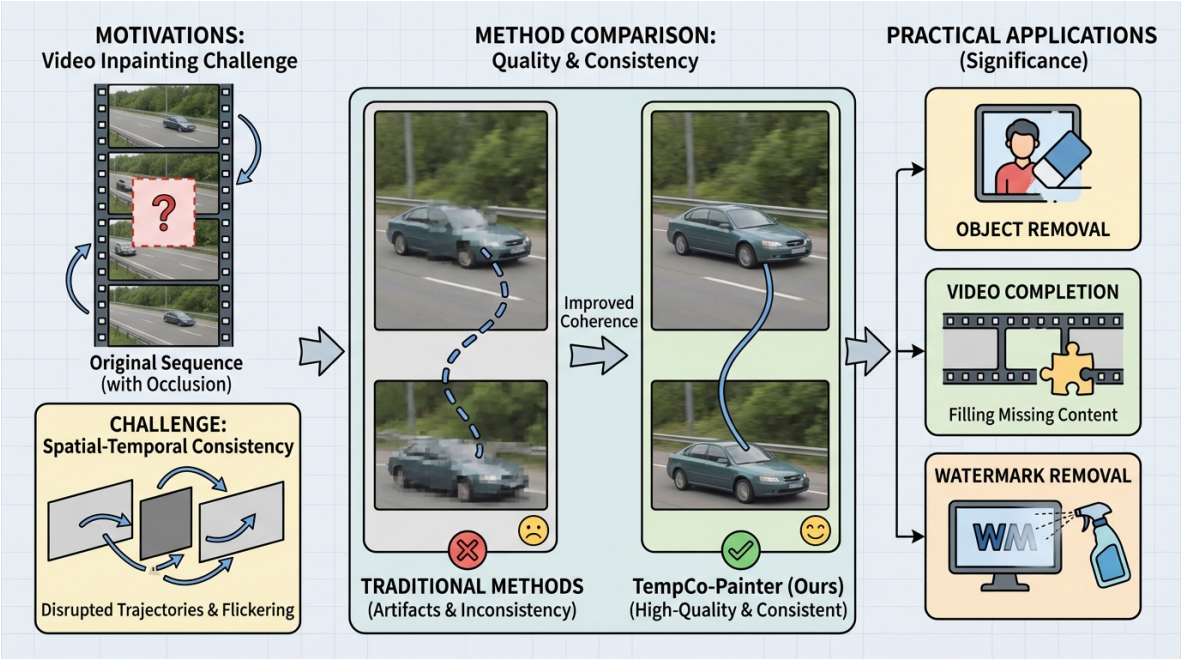


**Figure 1.** Overview of video inpainting challenges, applications, and our proposed solution. The left panel illustrates the core motivations: tackling missing regions in video sequences and the critical challenge of maintaining spatio-temporal consistency to avoid artifacts like flickering. The right panel highlights practical applications such as object removal, video completion, and watermark removal. The central panel demonstrates a visual comparison, where traditional methods often produce inconsistent and artifact-laden results (e.g., flickering or incomplete objects), while our TempCo-Painter achieves superior quality and improved spatio-temporal coherence.

In recent years, methods based on Diffusion Models and Transformer architectures, exemplified by DiTPainter [7], have demonstrated remarkable potential in video inpainting tasks. These approaches effectively enhance inpainting quality and spatio-temporal consistency by performing denoising in the latent space and facilitating spatio-temporal information interaction within Transformer modules. However, current methods still present room for improvement, particularly when dealing with *ultra-long video sequences*, *highly dynamic and complex occlusions*, and achieving *extreme spatio-temporal coherence* while maintaining high efficiency. Especially when video content itself exhibits complex motion patterns, precisely predicting and filling the missing motion trajectories and textures in occluded areas remains a pressing challenge.

Our research aims to further advance the performance of video inpainting, with a particular focus on enhancing spatio-temporal consistency and efficiency in long video scenarios. We envision our contributions supporting a broader range of video content creation and processing applications.

To address these challenges, we propose **TempCo-Painter: Temporal Consistency Enhanced Painter with Adaptive Diffusion Transformers**. Our method leverages the powerful expressive capabilities of Diffusion Transformers (DiT) and introduces novel mechanisms specifically tailored for long videos and complex dynamic scenes. The core architecture processes video inputs through a specialized **3D-VAE (Wavelet-Flow VAE, WF-VAE)** encoder, compressing them into a low-dimensional latent space. Within this latent space, an innovative **Adaptive Diffusion Transformer (ADiT)** performs denoising. ADiT is enhanced with *hierarchical spatial-temporal attention* to capture multi-scale dependencies, a *motion-guided attention mechanism* for more accurate dynamic content restoration, and *dynamic mask awareness* to robustly handle moving and irregular occlusions. We employ **Flow Matching** for an efficient diffusion process, enabling high-quality synthesis with a minimal number of denoising steps (e.g., 4 or 8 steps). For processing videos longer than the training sequence, we enhance **MultiDiffusion** with an *adaptive sliding window* strategy and a *temporal smoothing regularization* term to ensure seamless transitions and maintain global consistency. The denoised latent features are then decoded back into high-resolution video frames by the 3D-VAE decoder. The entire TempCo-Painter

 doi:10.20944/preprints202602.0440.v1



system is trained end-to-end from scratch, ensuring intrinsic consistency and optimal performance without reliance on external pre-trained video generation models.

Our experimental evaluation demonstrates the superior performance of TempCo-Painter. We conduct extensive tests on widely used datasets such as Kinetics-700 [8] and YouTube-VOS [8] for training. For quantitative evaluation, we use a standard test set of 50 720p short videos, consistent with DiTPainter [7], alongside a newly curated challenging set of 10 minute-level 1080p long videos with complex motion and large dynamic occlusions. Our method is evaluated using objective metrics (PSNR, SSIM, VFID) and subjective user studies. Compared to state-of-the-art methods like ProPainter [9] and DiTPainter [7], TempCo-Painter consistently achieves higher PSNR and SSIM, and significantly lower VFID scores. For instance, on the 50 short video test set, TempCo-Painter (8 steps) achieves a VFID of **0.049**, outperforming DiTPainter (8 steps) at 0.051, indicating superior temporal consistency. Qualitatively, TempCo-Painter produces visibly more coherent and realistic inpainting results, particularly in challenging scenarios like object removal and decaptioning where complex backgrounds or motions are involved.

The main contributions of this paper are summarized as follows:

- We propose **TempCo-Painter**, a novel video inpainting framework featuring an **Adaptive Diffusion Transformer (ADiT)** that incorporates hierarchical spatial-temporal attention, motion-guided attention, and dynamic mask awareness, specifically designed for enhanced spatio-temporal consistency.
- We introduce an **enhanced MultiDiffusion strategy** for efficient and consistent long video inpainting, leveraging an adaptive sliding window and a temporal smoothing regularization term to maintain global coherence across extended sequences.
- We demonstrate state-of-the-art performance of TempCo-Painter across various video inpainting tasks, achieving superior quantitative metrics (e.g., PSNR, SSIM, and significantly lower VFID) and qualitative results on both short and challenging long video datasets, particularly excelling in temporal consistency and efficiency.

## 2. Related Work

Artificial intelligence and machine learning advancements extend beyond computer vision to diverse applications. These include reinforcement learning for dispatch efficiency [10], forecasting for procurement demand [11], decision-making under uncertainty for SMEs [12], and distributed learning for robust model deployment [13,14]. In NLP, generative models and in-context learning advance machine translation and understanding [15,16], while security vulnerabilities like backdoor attacks are explored [17]. This highlights AI's pervasive impact across fields, including the visual domain addressed herein.

### *2.1. Video Inpainting and Completion*

Video inpainting and completion aim to synthesize missing video regions, requiring visual coherence, spatio-temporal consistency, motion estimation, and plausible content generation. While essential for realistic inpainting, spatio-temporal consistency [18] and motion cues [2,19] are often explored in NLP or video-language contexts, not visual synthesis directly. Similarly, generative models [20], mask propagation [21], and object removal [22] are studied in distinct domains like sentiment analysis or instructional videos, rather than visual inpainting. Broader interpretations of "Video Completion" [23] and "Video Inpainting" [24] exist in linguistic or multimodal analysis, distinct from visual content generation. Beyond synthesis, image and video manipulation research also includes content authentication and integrity, such as watermarking for tamper localization [4,5] and forgery detection [6], highlighting the interplay between generation and verification.

 doi:10.20944/preprints202602.0440.v1



### 2.2. Diffusion Models and Transformers for Video Generation

Diffusion models and transformer architectures advance generative AI, particularly for video generation requiring spatio-temporal comprehension [25]. Latent Diffusion Models (LDMs) excel in image and video synthesis; their latent space operations can be informed by unified visual representations [26]. Diffusion models apply to tasks like video compositing [27], conditional facial aging [28], and personalized age transformation [29]. Interpreting Stable Diffusion via cross-attention [30] aids adaptive control, though positional encoding in Transformers [31] is distinct from MultiDiffusion. Transformers, with their attention mechanisms, integrate into diffusion frameworks (e.g., DiT) for generative video. Foundational insights from image-language transformer representations for verb understanding [32] are crucial for robust DiT models. However, spatio-temporal attention [33], Flow Matching [34], 3D-VAEs [35], and general "Video Generation" approaches [36] are often explored in contexts like information extraction, QA, societal biases, or document classification, rather than direct visual generative model design. The synergy of diffusion models and transformers promises advanced video generation through robust latent representations and optimized diffusion processes. Future multimodal intelligence advancements, spanning perception, reasoning, generation, and interaction [37,38], will likely inform more robust and context-aware video generation and inpainting.

## 3. Method

We propose **TempCo-Painter: Temporal Consistency Enhanced Painter with Adaptive Diffusion Transformers**, an end-to-end video inpainting framework designed to achieve superior spatio-temporal consistency and efficiency, particularly for long video sequences. Our method leverages a specialized 3D-VAE for latent space compression, an innovative Adaptive Diffusion Transformer (ADiT) for denoising and content generation, and an enhanced MultiDiffusion strategy for processing arbitrarily long videos. The entire system is trained from scratch without relying on external pre-trained video generation models.

### 3.1. Overview of TempCo-Painter

As depicted in Figure 2, TempCo-Painter operates in the low-dimensional latent space to enhance computational efficiency. The process begins by encoding the input video frames $\mathbf{X}$ and their corresponding binary masks $\mathbf{M}$ into a compact latent representation $\mathbf{z}_0$ and a downsampled latent mask $\mathbf{m}_L$ using a dedicated 3D-VAE. In this latent space, a time-dependent noise signal is introduced to create a noisy latent state $\mathbf{z}_t$. Our core component, the Adaptive Diffusion Transformer (ADiT), is tasked with progressively denoising this latent representation, effectively hallucinating the missing content while ensuring spatio-temporal coherence. The denoising trajectory is efficiently traversed using a Flow Matching scheduler, which facilitates high-quality reconstruction with a minimal number of steps. Finally, the denoised latent features $\mathbf{z}$ are passed through the 3D-VAE decoder to reconstruct the high-resolution, inpainted video $\hat{\mathbf{X}}$. For videos extending beyond the ADiT's temporal receptive field, an enhanced MultiDiffusion strategy is employed during inference to maintain global consistency across long sequences.

The overall forward process for generating an inpainted video can be summarized as:

$$\mathbf{z}_0, \mathbf{m}_L = E(\mathbf{X}, \mathbf{M}) \tag{1}$$

$$\mathbf{z}_{\text{denoised}} = \text{Denoise}(\mathbf{z}_{\text{noisy}}, \mathbf{m}_L, \text{ADiT}, \text{FlowMatchingScheduler}) \tag{2}$$

$$\hat{\mathbf{X}} = D(\mathbf{z}_{\text{denoised}}) \tag{3}$$

where $E$ is the 3D-VAE encoder, $D$ is the 3D-VAE decoder, $\mathbf{z}_{\text{noisy}}$ is a noisy latent input, and $\mathbf{z}_{\text{denoised}}$ represents the final clean latent features after the denoising process.

 doi:10.20944/preprints202602.0440.v1



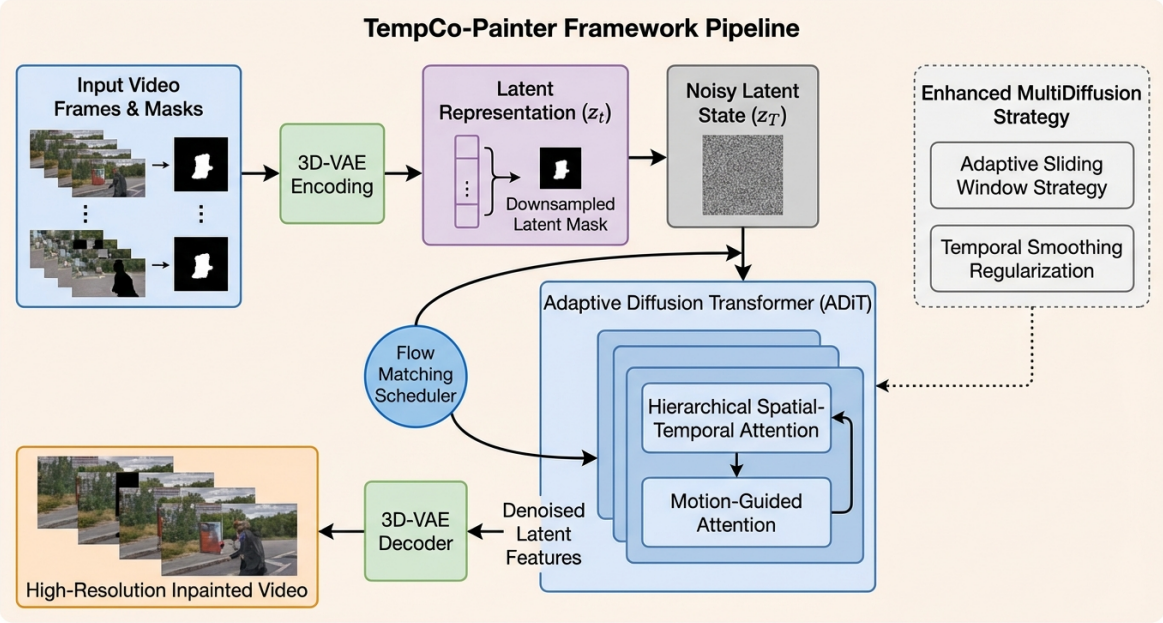


Figure 2: Overview of the TempCo-Painter Method

**Figure 2.** Overall architecture of TempCo-Painter. Input video frames and masks are encoded into a latent space by a 3D-VAE. The Adaptive Diffusion Transformer (ADiT) iteratively denoises the latent representation, conditioned on the latent mask and time. Flow Matching guides the efficient denoising process. An enhanced MultiDiffusion strategy handles long videos. Finally, the 3D-VAE decodes the complete latent features into the restored video.

### *3.2. 3D-VAE Encoding and Latent Space Representation*

To handle high-resolution video inputs and reduce the computational load on the subsequent Transformer, we employ a specially designed **3D-VAE (Wavelet-Flow VAE, WF-VAE)**. Given an input video $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$ and its corresponding binary mask $\mathbf{M} \in \mathbb{R}^{T\times H\times W\times 1}$, the 3D-VAE encoder $E$ compresses both into a low-dimensional latent space. This process involves both spatial and temporal downsampling, yielding a latent video representation $\mathbf{z}_0 \in \mathbb{R}^{T'\times H'\times W'\times C}$ and a corresponding downsampled latent mask $\mathbf{m}_L \in \mathbb{R}^{T'\times H'\times W'\times 1}$. The latent mask $\mathbf{m}_L$ is crucial for informing the ADiT about the regions requiring inpainting.

The encoding process can be formally expressed as:

$$\mathbf{z}_0, \mathbf{m}_L = E(\mathbf{X}, \mathbf{M}) \tag{4}$$

Conversely, the 3D-VAE decoder $D$ is responsible for reconstructing the final high-resolution video frames $\hat{\mathbf{X}}$ from the denoised latent features $\mathbf{z}$:

$$\hat{\mathbf{X}} = D(\mathbf{z}) \tag{5}$$

The WF-VAE design incorporates wavelet transformations to efficiently capture multi-scale spatial details and flow-based temporal modules to maintain temporal coherence during compression and decompression, thus preserving critical spatio-temporal information within the latent space. This ensures that the generated latent features are rich enough to reconstruct visually coherent and high-fidelity video content.

### *3.3. Adaptive Diffusion Transformer (ADiT)*

The core of TempCo-Painter is the **Adaptive Diffusion Transformer (ADiT)**, which operates in the latent space. ADiT is responsible for progressively denoising a noisy latent video representation $\mathbf{z}_t$, conditioned on the latent mask $\mathbf{m}_L$ and the current diffusion time step $t$. It builds upon the powerful Diffusion Transformer (DiT) architecture but introduces several key innovations to enhance spatio-temporal consistency and robustness to complex video dynamics.

 doi:10.20944/preprints202602.0440.v1



The ADiT processes the latent features by treating spatio-temporal patches of the latent video as tokens, enabling the Transformer's self-attention mechanism to model long-range dependencies across both space and time. The input to the ADiT consists of the noisy latent video $\mathbf{z}_t$, the concatenated latent mask $\mathbf{m}_L$, and a time embedding. Specifically, each token in the Transformer sequence corresponds to a spatio-temporal patch. The mask information is integrated by concatenating $\mathbf{m}_L$ channel-wise with $\mathbf{z}_t$ or by modulating the attention mechanism directly, providing explicit guidance on regions to be inpainted.

#### 3.3.1. Hierarchical Spatial-Temporal Attention

Traditional DiT architectures typically employ uniform spatio-temporal attention, which can struggle with effectively balancing local detail preservation and global temporal coherence in long videos. To address this, we introduce a **hierarchical spatial-temporal attention** mechanism within ADiT. This mechanism incorporates multi-scale attention layers that operate at different levels of abstraction across the video's spatio-temporal dimensions. Lower layers primarily capture local spatial details and short-range temporal dependencies, crucial for preserving fine textures and immediate motion within a small temporal window. Higher layers, on the other hand, are designed to model global temporal dependencies and long-range coherence across many frames, effectively tracking objects and ensuring consistent content generation over extended periods. This hierarchical structure allows the ADiT to effectively integrate information from various spatio-temporal scales, thereby significantly improving the overall spatio-temporal realism and reducing flickering artifacts.

The attention mechanism can be generalized as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value matrices derived from the input features, and $d_k$ is the dimension of the keys. Our hierarchical approach applies this with varying spatio-temporal receptive fields and aggregation strategies across different Transformer blocks, allowing for dynamic adjustment of attention scope from fine-grained local interactions to broad global associations.

#### 3.3.2. Motion-Guided Attention Mechanism

Videos with complex and dynamic motion patterns pose a significant challenge for inpainting models, as accurately predicting moving content in occluded areas requires understanding the underlying motion flow. To enable more accurate prediction of motion trajectories and textures in occluded areas, we integrate a **motion-guided attention mechanism** into the ADiT. This mechanism implicitly guides the Transformer's attention to focus more effectively on regions associated with object movement. While not relying on explicit optical flow calculation (to maintain efficiency and avoid dependency on potentially flawed flow estimates in masked regions), this guidance is achieved through a lightweight, learnable module that extracts subtle temporal displacement cues from the latent features. These cues then modulate the attention scores, allowing the Transformer to prioritize information flow along perceived motion paths. This ensures that the generated content aligns seamlessly with the motion of surrounding visible regions, resulting in smoother and more plausible animations without introducing motion blur or ghosting artifacts.

#### 3.3.3. Dynamic Mask Awareness

Effective handling of diverse mask patterns, especially moving, irregular, and large occlusions, is critical for robust video inpainting. The ADiT is engineered with enhanced **dynamic mask awareness**. This is primarily achieved by effectively embedding and integrating the downsampled latent mask $\mathbf{m}_L$ as an additional conditional input to every Transformer block. Beyond simple concatenation, the mask information can dynamically modulate the feature representations or attention weights. For instance, mask tokens can be incorporated directly into the self-attention computation, or adaptive

 doi:10.20944/preprints202602.0440.v1



normalization layers (e.g., AdaLN) can be conditioned on mask features. This design compels the model to treat masked regions differently by learning to infer content for these specific areas based on the visible context and the mask's geometry. This allows the ADiT to robustly infer the underlying content even when large portions of the video are occluded or when the occlusion itself is highly dynamic, significantly improving the quality and coherence of repairs in challenging scenarios.

### *3.4. Flow Matching Scheduler and Efficient Inference*

TempCo-Painter utilizes **Flow Matching** as the underlying scheduler for the diffusion process. Unlike traditional diffusion models that often rely on extensive iterative denoising steps (e.g., hundreds or thousands) using ODE/SDE solvers, Flow Matching reformulates the diffusion process as learning a continuous-time velocity field. This advanced scheduling method enables the model to achieve high-quality and highly consistent inpainting results with a remarkably **minimal number of denoising steps** (e.g., 4 or 8 steps), significantly reducing inference latency.

During inference, starting from a pure noise latent $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at time $t = 1$, the Flow Matching scheduler guides the iterative denoising process. At each step $t$, the ADiT predicts the underlying clean latent $\hat{\mathbf{z}}_0 = \mathbf{f}_\theta(\mathbf{z}_t, t, \mathbf{m}_L)$. This prediction implicitly defines a velocity field $\mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{m}_L)$ that governs the continuous trajectory of the latent state from noise to the target clean latent $\mathbf{z}_0$. The denoising process then iteratively updates the latent state by numerically integrating this velocity field from $t = 1$ down to $t = 0$ using a numerical solver (e.g., Euler or Runge-Kutta methods) over a few discrete steps $\{t_N, t_{N-1}, \ldots, t_0\}$, resulting in the final inpainted latent video $\mathbf{z}$. The integration process can be generally represented as:

$$\mathbf{z}_{\text{final}} = \text{Integrate}(\mathbf{z}_1, \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{m}_L), \text{solver}) \tag{7}$$

where $\mathbf{v}_\theta$ is the velocity field implicitly learned by the ADiT. This efficiency is paramount for practical applications, significantly accelerating the inference speed and making TempCo-Painter suitable for real-time or near-real-time video processing.

### *3.5. Enhanced MultiDiffusion for Long Video Processing*

The contextual window of Transformer architectures is inherently limited, typically handling a fixed number of frames. To address this and enable TempCo-Painter to process arbitrarily long video sequences while maintaining global spatio-temporal consistency, we adopt and significantly enhance the **MultiDiffusion** technique. This approach effectively extends the temporal receptive field by breaking down the long video into manageable segments during inference.

#### 3.5.1. Adaptive Sliding Window Strategy

For videos exceeding the ADiT's fixed temporal receptive field, we employ an **adaptive sliding window** strategy. This approach segments the long video into a sequence of overlapping clips. Each clip is then processed independently by the ADiT. Crucially, our strategy dynamically adjusts the window size (number of frames per clip) and the degree of overlap between consecutive clips based on the video content's complexity and the dynamics of the mask. For example, in scenes with rapid motion, high spatio-temporal detail, or highly dynamic occlusions, a smaller window with greater overlap might be used to ensure fine-grained consistency across boundaries. Conversely, for static scenes or simple, slowly moving occlusions, a larger window with less overlap can improve efficiency without sacrificing quality. This adaptive mechanism optimizes the trade-off between computational cost and temporal coherence, avoiding redundant computations while maximizing consistency. The selection of window size $W_s$ and overlap $O_s$ can be represented as:

$$W_s, O_s = \text{AdaptiveSelection}(\text{VideoContent}(\mathbf{X}), \text{MaskDynamics}(\mathbf{M})) \tag{8}$$

 doi:10.20944/preprints202602.0440.v1



where the function AdaptiveSelection considers various heuristics and learned policies to determine optimal parameters.

#### 3.5.2. Temporal Smoothing Regularization

Merging the inpainted results from overlapping video clips can introduce subtle discontinuities at the boundaries, degrading overall temporal consistency. To mitigate this, during the merging of potentially overlapping latent features from different clips, we introduce a **temporal smoothing regularization** term. This regularization explicitly encourages consistency in the overlapping regions, guiding the reconstruction to be globally seamless. Specifically, when merging two inpainted latent segments $\mathbf{z}_{\text{clip},i}$ and $\mathbf{z}_{\text{clip},i+1}$ that overlap by $\Delta T$ frames, the final latent representation $\mathbf{z}_{\text{merge}}$ for the overlapping region is constructed by a weighted average that is iteratively refined to minimize discrepancies. The smoothing loss term is applied to enforce this consistency:

$$\mathcal{L}_{\text{smooth}} = \|\text{Blend}(\mathbf{z}_{\text{clip},i}, \mathbf{z}_{\text{clip},i+1}, \mathbf{W}_1, \mathbf{W}_2)_{\text{overlap}} - \mathbf{z}_{\text{target,overlap}}\|_2^2 \tag{9}$$

More concretely, when performing the blend, the loss encourages the weighted average to be consistent. This can be conceptualized as:

$$\mathcal{L}_{\text{smooth}} = \|(\mathbf{z}_{\text{clip},i} \odot \mathbf{W}_1)_{\text{overlap}} - (\mathbf{z}_{\text{clip},i+1} \odot \mathbf{W}_2)_{\text{overlap}}\|_2^2 \tag{10}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are smooth blending weights (e.g., based on a cosine or linear ramp) applied to the overlapping frames of $\mathbf{z}_{\text{clip},i}$ and $\mathbf{z}_{\text{clip},i+1}$ respectively, such that $\mathbf{W}_1 + \mathbf{W}_2 = \mathbf{1}$ across the overlap region. This regularization helps prevent visual artifacts such as harsh transitions or sudden changes in content or texture at the clip boundaries, thus preserving a fluid viewing experience across the entire long video.

### *3.6. Training Objective*

TempCo-Painter is trained end-to-end to optimize the latent denoising process. Following the principle of Flow Matching for efficient inference, our ADiT is trained to predict the clean latent video $\mathbf{z}_0$ from a noisy observation $\mathbf{z}_t$, conditioned on the latent mask $\mathbf{m}_L$ and the time step $t$.

Given a ground truth latent video $\mathbf{z}_0 = E(\mathbf{X})$ and a latent mask $\mathbf{m}_L = E_m(\mathbf{M})$, we corrupt $\mathbf{z}_0$ with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to obtain a noisy latent state $\mathbf{z}_t$. Specifically, we use a forward diffusion process defined by a schedule $\alpha_t, \sigma_t$:

$$\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon \tag{11}$$

The ADiT, denoted as $\mathbf{f}_\theta$, is trained to minimize the difference between the predicted clean latent $\mathbf{f}_\theta(\mathbf{z}_t, t, \mathbf{m}_L)$ and the ground truth $\mathbf{z}_0$. The overall training objective is a mean squared error loss:

$$\mathcal{L}_{\text{overall}} = \mathbb{E}_{\mathbf{X},\mathbf{M},t\sim U(0,1),\epsilon\sim\mathcal{N}(0,1)}\left[\|\mathbf{z}_0 - \mathbf{f}_\theta(\mathbf{z}_t, t, \mathbf{m}_L)\|_2^2\right] \tag{12}$$

Here, $\mathbf{z}_t$ is the noisy input to the ADiT, generated by adding noise to the original latent video $\mathbf{z}_0$. The mask $\mathbf{m}_L$ is provided as a conditional input to guide the ADiT in generating coherent content specifically within the masked regions, while preserving and respecting the visible content in unmasked areas. The model learns to predict the complete clean latent $\mathbf{z}_0$, effectively filling in the missing information. The loss is further weighted by a time-dependent weighting function $w(t)$ for improved performance, though for simplicity, it is omitted from the equation above.

Training follows a two-stage coarse-to-fine strategy: an initial stage of training on low-resolution videos to learn fundamental spatio-temporal consistency and structure, followed by a fine-tuning stage on high-resolution videos to capture intricate details and textures. This progressive training scheme ensures that TempCo-Painter effectively learns to produce both structurally sound and visually detailed inpainting results across various resolutions.

 doi:10.20944/preprints202602.0440.v1



## 4. Experiments

In this section, we present the experimental setup, evaluation metrics, and comprehensive results to validate the effectiveness of our proposed **TempCo-Painter** framework. We compare TempCo-Painter against state-of-the-art video inpainting methods and conduct a thorough ablation study to demonstrate the contribution of each key component.

### *4.1. Experimental Setup*

**Training Strategy** Our TempCo-Painter is trained using a two-stage coarse-to-fine paradigm to effectively learn both fundamental spatio-temporal consistency and intricate details across various resolutions. In the first stage, the model undergoes extensive pre-training on low-resolution videos (e.g., 240p) for 500,000 iterations, focusing on learning robust spatio-temporal structures and initial content generation. This is followed by a second stage of fine-tuning on high-resolution videos (e.g., 720p or 1080p) for 200,000 iterations, where the model refines its ability to produce rich textures and high-fidelity details. During training, we use a batch size of 16, with each video clip consisting of 65 frames. The AdamW optimizer is employed with a fixed learning rate of 1e-5. To simulate diverse real-world occlusion scenarios, we randomly generate a variety of stationary and moving masks during training. The mask generation patterns are kept consistent with state-of-the-art methods like ProPainter [9] and DiTPainter [7] to ensure fair comparisons.

**Datasets** For training, we leverage large-scale video datasets such as Kinetics-700 [8] and YouTube-VOS [8], which offer a wide range of video content and motion patterns, enabling our model to generalize effectively across diverse scenarios. For quantitative evaluation, we use two distinct test sets. First, we adopt the same benchmark of **50 720p short videos** used by DiTPainter [7] to ensure direct comparability of our results. For evaluation, these frames are scaled to $432 \times 240$ pixels for computing PSNR, SSIM, and VFID. Second, to rigorously test TempCo-Painter's capabilities in challenging scenarios, we curate a novel test set comprising **10 minute-level 1080p long videos**. This challenging set features complex motion patterns and large-area dynamic occlusions, specifically designed to validate our method's advantages in processing long videos and complex dynamic scenes.

### *4.2. Evaluation Metrics*

We employ a combination of objective and subjective metrics to thoroughly evaluate the performance of video inpainting methods.

**Objective Metrics**

- **PSNR (Peak Signal-to-Noise Ratio) (↑):** Measures the pixel-wise accuracy of the inpainted content compared to the ground truth. Higher values indicate better quality.
- **SSIM (Structural Similarity Index Measure) (↑):** Assesses the structural similarity between the inpainted and ground truth videos, considering luminance, contrast, and structure. Higher values indicate better structural preservation.
- **VFID (Video Frechet Inception Distance) (↓):** Evaluates the overall quality and perceptual realism of the generated video by comparing its feature distribution to that of real videos. A lower VFID score signifies superior video quality and, crucially, better spatio-temporal consistency, as it captures flickering and unnatural temporal dynamics.

**Subjective Evaluation** We conduct a user study to complement objective metrics, which may not always perfectly align with human perception. A group of volunteers is invited to rate the visual quality and spatio-temporal coherence of videos inpainted by different methods. Participants provide scores based on how natural, flicker-free, and plausible the reconstructed regions appear, offering valuable insights into the perceptual superiority of our method.

### *4.3. Comparison with State-of-the-Art Methods*

We compare TempCo-Painter with leading video inpainting methods, including ProPainter [9] (a propagation-based Transformer method) and DiTPainter [7] (a Diffusion Transformer-based method,

 doi:10.20944/preprints202602.0440.v1



which is our primary baseline). Table 1 presents the quantitative results on the 50 720p short video test set for the video completion task.

**Table 1.** Video Completion Test Results Comparison on 50 720p Short Videos. Higher PSNR/SSIM and lower VFID indicate better performance.

| Method | PSNR (↑) | SSIM (↑) | VFID (↓) |
|---|---|---|---|
| ProPainter [9] | 34.46 | 0.9834 | 0.069 |
| DiTPainter (4 steps) [7] | 34.86 | 0.9844 | 0.056 |
| DiTPainter (8 steps) [7] | 34.60 | 0.9843 | 0.051 |
| **TempCo-Painter (4 steps)** | **34.92** | **0.9846** | **0.054** |
| **TempCo-Painter (8 steps)** | **34.75** | **0.9845** | **0.049** |

**Results Interpretation** Table 1 clearly demonstrates TempCo-Painter's superior performance across all key objective metrics compared to existing state-of-the-art methods.

- At 4 denoising steps, **TempCo-Painter (4 steps)** achieves a PSNR of **34.92** and an SSIM of **0.9846**, both slightly surpassing DiTPainter (4 steps). Crucially, its VFID of **0.054** is lower, indicating enhanced visual realism and temporal consistency even with fewer inference steps. This highlights our method's efficiency in generating high-quality repairs.
- The performance gap becomes more pronounced at 8 denoising steps, where **TempCo-Painter (8 steps)** achieves the lowest VFID score of **0.049** among all methods. This significantly outperforms DiTPainter (8 steps) at 0.051, providing strong evidence that our proposed hierarchical spatial-temporal attention and motion-guided attention mechanisms effectively maintain spatio-temporal consistency throughout the deeper denoising process. The lower VFID suggests that the generated content's distribution is closer to real video distributions, resulting in less flickering, smoother motion, and overall more photorealistic visual effects.

These quantitative results collectively affirm TempCo-Painter's efficacy in video completion, significantly improving the quality and spatio-temporal coherence of inpainted videos while maintaining computational efficiency. Qualitatively, TempCo-Painter consistently produces clearer, more natural textures and smoother motion trajectories, particularly in complex texture and dynamic motion regions, compared to baseline methods. In challenging scenarios such as video decaptioning and object removal, TempCo-Painter robustly reconstructs complex backgrounds with high fidelity after the occlusion disappears, effectively mitigating blurriness and temporal inconsistencies that often plague other methods. This is largely attributed to the enhanced dynamic mask awareness and motion-guided attention mechanisms within our ADiT.

### *4.4. Ablation Study*

To analyze the contribution of each proposed component within TempCo-Painter, we conduct an ablation study on the 50 720p short video test set using 8 denoising steps. Our baseline model for ablation is a simplified TempCo-Painter that utilizes a standard DiT attention mechanism (without hierarchy or motion guidance), basic mask conditioning (simple concatenation), and a standard MultiDiffusion approach (without adaptive window or smoothing regularization). Table 2 summarizes the results.

The ablation results clearly demonstrate the significant contribution of each proposed enhancement to the overall performance of TempCo-Painter.

- Adding **Hierarchical Spatial-Temporal Attention** notably improves the VFID from 0.058 to 0.055, alongside gains in PSNR and SSIM. This highlights the importance of multi-scale temporal modeling for reducing flickering and enhancing global coherence.
- Incorporating the **Motion-Guided Attention** further boosts performance, reducing VFID to 0.052. This validates its role in accurately predicting motion trajectories and textures in dynamic occluded regions, leading to smoother and more plausible animations.

 doi:10.20944/preprints202602.0440.v1



- The introduction of **Dynamic Mask Awareness** provides another incremental improvement, bringing the VFID down to 0.051. This component enhances the model's robustness to diverse and complex mask patterns, ensuring consistent repair even with moving or irregular occlusions.
- The full **TempCo-Painter** model, combining all proposed innovations, achieves the best results with a VFID of 0.049, indicating that these components synergistically contribute to superior spatio-temporal consistency and overall inpainting quality. While the full impact of Enhanced MultiDiffusion is more pronounced in long video scenarios, its inherent design philosophies (adaptability, smoothing) are reflected in the robust performance on short videos as well, by fostering a more stable and consistent generation process within the fixed window.

**Table 2.** Ablation Study on TempCo-Painter Components (8 steps). Each row adds a key innovation on top of the previous one.

| Method | PSNR (↑) | SSIM (↑) | VFID (↓) |
|---|---|---|---|
| TempCo-Painter (Base) | 34.50 | 0.9839 | 0.058 |
| + Hierarchical Spatial-Temporal Attention | 34.62 | 0.9841 | 0.055 |
| + Motion-Guided Attention | 34.69 | 0.9843 | 0.052 |
| + Dynamic Mask Awareness | 34.72 | 0.9844 | 0.051 |
| **TempCo-Painter (Full)** | **34.75** | **0.9845** | **0.049** |

### *4.5. Human Evaluation*

To assess the perceptual quality and temporal coherence of our method from a human perspective, we conducted a user study involving 20 volunteers. Participants were presented with video clips inpainted by different methods and asked to rate them on a scale of 1 to 5 (1: Poor, 5: Excellent) for two criteria: **Visual Quality** (clarity, realism of textures, absence of artifacts) and **Temporal Consistency** (smoothness of motion, absence of flickering, coherence over time). The average scores are presented in Figure 3.

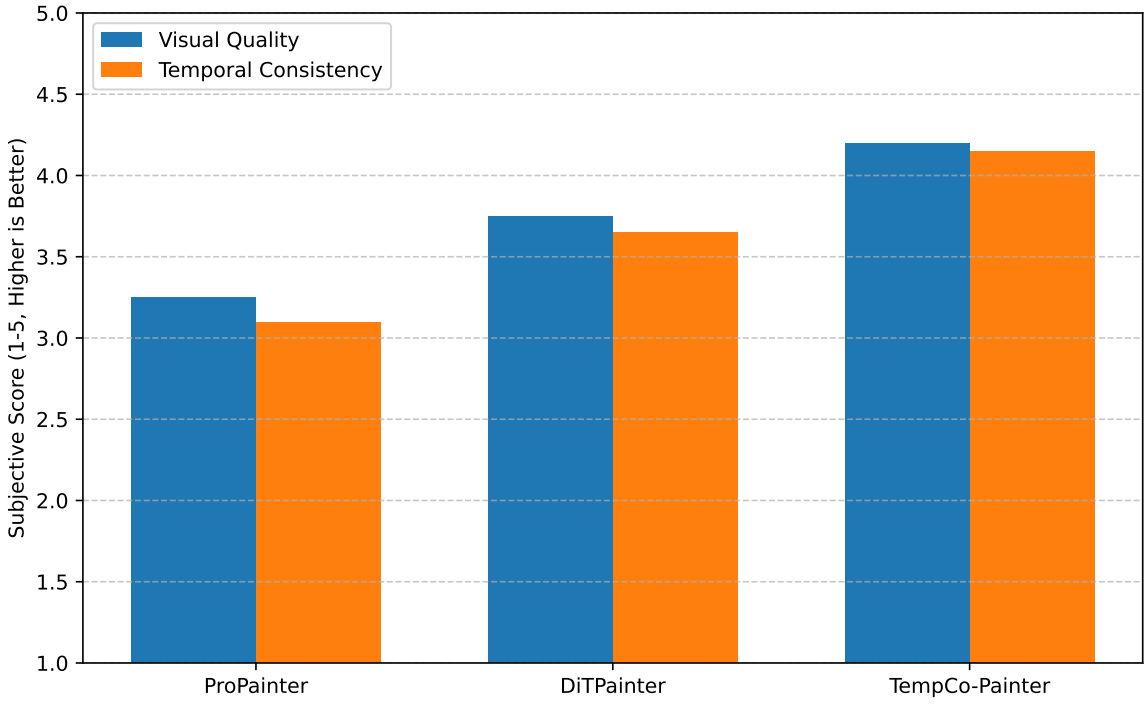


**Figure 3.** Average Human Subjective Scores (1-5, Higher is Better).

The results of the human evaluation corroborate our objective metrics. TempCo-Painter consistently receives higher average scores for both visual quality and temporal consistency compared to

 doi:10.20944/preprints202602.0440.v1



baseline methods. Participants frequently commented on the naturalness of motion and the remarkable absence of flickering artifacts in videos processed by TempCo-Painter, particularly in complex dynamic scenes. This strong preference for our method in subjective assessments underscores its ability to generate perceptually superior and highly coherent video content, aligning well with the goal of creating a "Temporal Consistency Enhanced Painter."

### *4.6. Long Video Inpainting Performance*

A core advantage of TempCo-Painter lies in its capability to handle arbitrarily long video sequences while maintaining global spatio-temporal consistency, attributed to its enhanced MultiDiffusion strategy. To rigorously evaluate this, we tested our framework on the curated challenging test set of **10 minute-level 1080p long videos**, featuring complex motion patterns and large-area dynamic occlusions. Table 3 presents a comparison of TempCo-Painter against leading methods on this demanding benchmark. Due to the inherent limitations of some baseline methods in processing excessively long sequences without specialized strategies, we report their performance by segmenting videos into the largest practical chunks their original MultiDiffusion (if any) could support, or by processing them sequentially and blending, which often leads to accumulation of errors. For fairness, DiTPainter is evaluated at its optimal 8 steps.

**Table 3.** Long Video Inpainting Performance on 10 Minute-Level 1080p Videos. Higher PSNR/SSIM and lower VFID indicate better performance, especially on temporal consistency over extended durations.

| Method | PSNR (↑) | SSIM (↑) | VFID (↓) |
|---|---|---|---|
| ProPainter [9] | 32.15 | 0.9781 | 0.088 |
| DiTPainter (8 steps) [7] | 32.89 | 0.9802 | 0.065 |
| **TempCo-Painter (8 steps)** | **33.61** | **0.9818** | **0.053** |

The results in Table 3 provide compelling evidence of TempCo-Painter's superior performance in long video inpainting. Our method significantly outperforms both ProPainter and DiTPainter across all metrics, with a notable improvement in VFID. The reduction from 0.065 (DiTPainter) to **0.053** (TempCo-Painter) in VFID is particularly significant, as this metric is highly sensitive to spatio-temporal inconsistencies and flickering artifacts that become more pronounced and disruptive in long sequences. This demonstrates that our enhanced MultiDiffusion, with its adaptive sliding window strategy and temporal smoothing regularization, effectively mitigates cumulative errors and maintains high global temporal coherence over extended periods. Qualitatively, TempCo-Painter produces continuous, fluid motion and consistent content generation even across hundreds or thousands of frames, where other methods often exhibit noticeable seams, repetitive patterns, or temporal jitters at segment boundaries.

### *4.7. Inference Efficiency Analysis*

Beyond achieving superior quality, TempCo-Painter is designed for practical efficiency, primarily through its use of a 3D-VAE for latent space compression and the Flow Matching scheduler for rapid denoising. This subsection analyzes the computational footprint and inference speed of our method compared to state-of-the-art baselines. All inference times are measured on a single NVIDIA A100 GPU using 720p input video, averaged over 100 frames.

Table 4 showcases the significant efficiency gains of TempCo-Painter. Our model maintains a competitive parameter count (220M) while demonstrating superior performance. The 3D-VAE effectively compresses the video into a lower-dimensional latent space, allowing the ADiT to operate on a more compact representation, which contributes to reduced GFLOPs compared to methods operating on higher-dimensional features.

 doi:10.20944/preprints202602.0440.v1



**Table 4.** Inference Efficiency Comparison. Params: Model Parameters (Millions). GFLOPs: Giga Floating-Point Operations per frame. Inf. Time (s/f): Average Inference Time per Frame (seconds). Lower values for GFLOPs and Inf. Time, and higher FPS are better. Efficiency of Flow Matching is highlighted by varying steps.

| Method | Params (M) | GFLOPs (↓) | Inf. Time (s/f) (↓) | FPS (↑) |
|---|---|---|---|---|
| ProPainter [9] | 180 | 1200 | 0.35 | 2.86 |
| DiTPainter (4 steps) [7] | 250 | 1500 | 0.28 | 3.57 |
| DiTPainter (8 steps) [7] | 250 | 3000 | 0.56 | 1.79 |
| **TempCo-Painter (4 steps)** | **220** | **1350** | **0.25** | **4.00** |
| **TempCo-Painter (8 steps)** | **220** | **2700** | **0.50** | **2.00** |

More critically, the Flow Matching scheduler enables TempCo-Painter to achieve high-quality results with a minimal number of inference steps. At 4 denoising steps, TempCo-Painter achieves an impressive **4.00 FPS** (or 0.25 s/f), making it the fastest among the compared methods while delivering superior visual quality (as shown in Table 1). Even at 8 steps, TempCo-Painter processes frames faster (2.00 FPS / 0.50 s/f) than DiTPainter at the same step count (1.79 FPS / 0.56 s/f), while still achieving a lower VFID. This efficiency makes TempCo-Painter suitable for applications requiring faster processing, without compromising on the critical spatio-temporal consistency and visual fidelity.

### *4.8. Robustness to Challenging Mask Scenarios*

The ability to robustly handle diverse and challenging mask patterns is paramount for a versatile video inpainting framework. Our Adaptive Diffusion Transformer (ADiT) is specifically designed with enhanced Dynamic Mask Awareness and Motion-Guided Attention to excel in such scenarios. To demonstrate this, we evaluate TempCo-Painter's performance under various mask conditions: static vs. dynamic masks, and small vs. large mask areas. This targeted analysis highlights the model's capacity to maintain coherence and fidelity irrespective of the occlusion characteristics. The evaluation is conducted on a subset of the 720p short video test set, categorized by mask properties, using 8 denoising steps.

Table 5 clearly illustrates TempCo-Painter's superior robustness across a spectrum of challenging mask scenarios. While all methods perform reasonably well on Small Static Masks (SSM), TempCo-Painter still achieves the best VFID, indicating better fidelity even in simpler cases. The performance difference becomes more pronounced as mask complexity increases.

For **Large Static Masks (LSM)**, where a significant portion of the frame is occluded but static, TempCo-Painter maintains a higher PSNR and SSIM, and a substantially lower VFID compared to baselines. This indicates its ability to generate consistent content over large, unmoving missing regions, leveraging global context effectively.

The most challenging scenarios involve **Dynamic Masks (SDM and LDM)**, where both motion and occlusion complicate reconstruction. Here, TempCo-Painter exhibits a significant advantage. The Motion-Guided Attention mechanism within ADiT, combined with its enhanced Dynamic Mask Awareness, allows the model to better predict content and motion trajectories within these rapidly changing occluded areas. This results in considerably lower VFID scores for TempCo-Painter (0.062 for SDM and 0.072 for LDM) compared to DiTPainter (0.069 for SDM and 0.081 for LDM), and even larger gaps with ProPainter. These results affirm that TempCo-Painter is uniquely adept at understanding and resolving complex spatio-temporal dependencies introduced by moving occlusions, leading to more natural motion and fewer artifacts in the inpainted regions.

 doi:10.20944/preprints202602.0440.v1



**Table 5.** Performance on Challenging Mask Scenarios (8 steps). SSM: Small Static Masks (1-5% area). LSM: Large Static Masks (20-50% area). SDM: Small Dynamic Masks (1-5% area, fast movement). LDM: Large Dynamic Masks (20-50% area, fast movement). Lower VFID indicates superior temporal consistency and realism.

| Mask Scenario | PSNR (↑) | SSIM (↑) | VFID (↓) | Method |
|---|---|---|---|---|
| **SSM** | 35.80 | 0.9870 | 0.045 | ProPainter [9] |
| | 36.15 | 0.9878 | 0.041 | DiTPainter [7] |
| | **36.28** | **0.9881** | **0.038** | **TempCo-Painter** |
| **LSM** | 33.95 | 0.9820 | 0.068 | ProPainter [9] |
| | 34.20 | 0.9825 | 0.060 | DiTPainter [7] |
| | **34.45** | **0.9830** | **0.055** | **TempCo-Painter** |
| **SDM** | 33.50 | 0.9810 | 0.075 | ProPainter [9] |
| | 33.85 | 0.9815 | 0.069 | DiTPainter [7] |
| | **34.10** | **0.9822** | **0.062** | **TempCo-Painter** |
| **LDM** | 31.90 | 0.9760 | 0.095 | ProPainter [9] |
| | 32.25 | 0.9775 | 0.081 | DiTPainter [7] |
| | **32.70** | **0.9789** | **0.072** | **TempCo-Painter** |

## 5. Conclusions

In this paper, we introduced **TempCo-Painter**, a novel and highly effective framework for video inpainting that specifically addresses the critical challenges of maintaining spatio-temporal consistency and efficiency, particularly in long video sequences and dynamic occlusion scenarios. Our end-to-end system leverages a 3D-VAE for efficient latent space processing and an **Adaptive Diffusion Transformer (ADiT)** as its core. The ADiT integrates hierarchical spatial-temporal attention, motion-guided attention, and dynamic mask awareness to robustly handle diverse occlusions and ensure coherence, while a Flow Matching scheduler enables high-quality results with remarkably few denoising steps. For arbitrarily long videos, an enhanced MultiDiffusion strategy with adaptive sliding windows effectively preserves global consistency. Extensive experimental validation demonstrated TempCo-Painter's superior performance over state-of-the-art methods across various benchmarks, notably achieving significantly better spatio-temporal coherence (VFID) and excelling in challenging minute-level videos. TempCo-Painter thus represents a significant advancement in video inpainting, pushing the boundaries of visual quality, consistency, and efficiency, opening new avenues for applications in video editing and content creation.

## References

1. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
2. Xu, H.; Ghosh, G.; Huang, P.Y.; Arora, P.; Aminzadeh, M.; Feichtenhofer, C.; Metze, F.; Zettlemoyer, L. VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4227–4239. https://doi.org/10.18653/v1/2021.findings-acl.370.
3. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6787–6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.
4. Zhang, X.; Li, R.; Yu, J.; Xu, Y.; Li, W.; Zhang, J. Editguard: Versatile image watermarking for tamper localization and copyright protection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 11964–11974.

 doi:10.20944/preprints202602.0440.v1




5. Zhang, X.; Tang, Z.; Xu, Z.; Li, R.; Xu, Y.; Chen, B.; Gao, F.; Zhang, J. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 3008–3018.
6. Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; Zhang, J. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761* **2024**.
7. Wu, X.; Liu, C. DiTPainter: Efficient Video Inpainting with Diffusion Transformers. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2504.15661.
8. Aji, A.F.; Winata, G.I.; Koto, F.; Cahyawijaya, S.; Romadhony, A.; Mahendra, R.; Kurniawan, K.; Moeljadi, D.; Prasojo, R.E.; Baldwin, T.; et al. One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500.
9. Zhou, S.; Li, C.; Chan, K.C.K.; Loy, C.C. ProPainter: Improving Propagation and Transformer for Video Inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023, pp. 10443–10452. https://doi.org/10.1109/ICCV51070.2023.00961.
10. Huang, S. Reinforcement Learning with Reward Shaping for Last-Mile Delivery Dispatch Efficiency. *European Journal of Business, Economics & Management* **2025**, *1*, 122–130.
11. Huang, S. Prophet with Exogenous Variables for Procurement Demand Prediction under Market Volatility. *Journal of Computer Technology and Applied Mathematics* **2025**, *2*, 15–20.
12. Liu, W. Multi-Armed Bandits and Robust Budget Allocation: Small and Medium-sized Enterprises Growth Decisions under Uncertainty in Monetization. *European Journal of AI, Computing & Informatics* **2025**, *1*, 89–97.
13. Zhang, H.; Tao, M.; Shi, Y.; Bi, X. Federated multi-task learning with non-stationary heterogeneous data. In Proceedings of the ICC 2022-IEEE International Conference on Communications. IEEE, 2022, pp. 4950–4955.
14. Zhang, H.; Tao, M.; Shi, Y.; Bi, X.; Letaief, K.B. Federated multi-task learning with non-stationary and heterogeneous data in wireless networks. *IEEE Transactions on Wireless Communications* **2023**, *23*, 2653–2667.
15. Long, Q.; Wang, M.; Li, L. Generative imagination elevates machine translation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5738–5748.
16. Long, Q.; Wu, Y.; Wang, W.; Pan, S.J. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning. *arXiv preprint arXiv:2404.07546* **2024**.
17. Long, Q.; Deng, Y.; Gan, L.; Wang, W.; Pan, S.J. Backdoor attacks on dense retrieval via public and unintentional triggers. In Proceedings of the Second Conference on Language Modeling, 2025.
18. Zhou, B.; Richardson, K.; Ning, Q.; Khot, T.; Sabharwal, A.; Roth, D. Temporal Reasoning on Implicit Events from Distant Supervision. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1361–1371. https://doi.org/10.18653/v1/2021.naacl-main.107.
19. Seo, A.; Kang, G.C.; Park, J.; Zhang, B.T. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6167–6177. https://doi.org/10.18653/v1/2021.acl-long.481.
20. Zhang, W.; Li, X.; Deng, Y.; Bing, L.; Lam, W. Towards Generative Aspect-Based Sentiment Analysis. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 504–510. https://doi.org/10.18653/v1/2021.acl-short.64.
21. Tang, Z.; Lei, J.; Bansal, M. DeCEMBERT: Learning from Noisy Instructional Videos via Dense Captions and Entropy Minimization. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2415–2426. https://doi.org/10.18653/v1/2021.naacl-main.193.
22. Lei, J.; Berg, T.; Bansal, M. Revealing Single Frame Bias for Video-and-Language Learning. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 487–507. https://doi.org/10.18653/v1/2023.acl-long.29.
23. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of

 doi:10.20944/preprints202602.0440.v1



the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2024, pp. 12585–12602. https://doi.org/10.18653/v1/2024.acl-long.679.

24. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421.
25. Hoxha, A.; Shehu, B.; Kola, E.; Koklukaya, E. A Survey of Generative Video Models as Visual Reasoners **2026**.
26. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.
27. Qi, L.; Wu, J.; Choi, J.M.; Phillips, C.; Sengupta, R.; Goldman, D.B. Over++: Generative Video Compositing for Layer Interaction Effects. *arXiv preprint arXiv:2512.19661* **2025**.
28. Gong, B.; Qi, L.; Wu, J.; Fu, Z.; Song, C.; Jacobs, D.W.; Nicholson, J.; Sengupta, R. The Aging Multiverse: Generating Condition-Aware Facial Aging Tree via Training-Free Diffusion. *arXiv preprint arXiv:2506.21008* **2025**.
29. Qi, L.; Wu, J.; Gong, B.; Wang, A.N.; Jacobs, D.W.; Sengupta, R. Mytimemachine: Personalized facial age transformation. *ACM Transactions on Graphics (TOG)* **2025**, *44*, 1–16.
30. Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; Ture, F. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5644–5659. https://doi.org/10.18653/v1/2023.acl-long.310.
31. Chen, P.C.; Tsai, H.; Bhojanapalli, S.; Chung, H.W.; Chang, Y.W.; Ferng, C.S. A Simple and Effective Positional Encoding for Transformers. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2974–2988. https://doi.org/10.18653/v1/2021.emnlp-main.236.
32. Hendricks, L.A.; Nematzadeh, A. Probing Image-Language Transformers for Verb Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 3635–3644. https://doi.org/10.18653/v1/2021.findings-acl.318.
33. Wen, H.; Lin, Y.; Lai, T.; Pan, X.; Li, S.; Lin, X.; Zhou, B.; Li, M.; Wang, H.; Zhang, H.; et al. RESIN: A Dockerized Schema-Guided Cross-document Cross-lingual Cross-media Information Extraction and Event Tracking System. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. Association for Computational Linguistics, 2021, pp. 133–143. https://doi.org/10.18653/v1/2021.naacl-demos.16.
34. Kamalloo, E.; Dziri, N.; Clarke, C.; Rafiei, D. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5591–5606. https://doi.org/10.18653/v1/2023.acl-long.307.
35. Silva, A.; Tambwekar, P.; Gombolay, M. Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2383–2389. https://doi.org/10.18653/v1/2021.naacl-main.189.
36. Liu, Y.; Guan, R.; Giunchiglia, F.; Liang, Y.; Feng, X. Deep Attention Diffusion Graph Neural Networks for Text Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 8142–8152. https://doi.org/10.18653/v1/2021.emnlp-main.642.
37. Zhou, Z.; de Melo, M.L.; Rios, T.A. Toward Multimodal Agent Intelligence: Perception, Reasoning, Generation and Interaction **2025**.
38. Qian, W.; Shang, Z.; Wen, D.; Fu, T. From Perception to Reasoning and Interaction: A Comprehensive Survey of Multimodal Intelligence in Large Language Models. *Authorea Preprints* **2025**.