VidEvo: Evolving Video Editing through Exhaustive Temporal Modeling

Sizhe Dang[1], **Huan Liu**[1 *], **Mengmeng Wang**[2,3 *], **Xin Lai**[1], **Guang Dai**[3], **Jingdong Wang**[4]
[1]Xi'an Jiaotong University
[2]Zhejiang University of Technology
[3]SGIT AI Lab, State Grid Corporation of China
[4]Baidu Inc.
darknight1118@stu.xjtu.edu.cn, huanliu@xjtu.edu.cn, mengmengwang@zju.edu.cn
laixin@xjtu.edu.cn, guang.gdai@gmail.com, wangjingdong@outlook.com

Abstract

Text-guided video editing (TGVE) has become a recent hotspot due to its entertainment value and practical applications. To reduce overhead, existing methods primarily extend from text-to-image diffusion models and typically involve reconstruction and editing phases. However, challenges persist, particularly in enhancing temporal consistency of a video while adhering to textual alignment requirements. A crucial factor leading to the aforementioned issue is the inadequate and implicit tuning of the attention module within existing methods, which is specifically designed to capture temporal information. In light of this, we introduce VidEvo, a novel one-shot video editing method that leverages explicit cues derived from the original video to enhance temporal modeling. By integrating null-video embedding (NVE) and window-frame attention (WFA) components, VidEvo facilitates the smooth and coherent generation of videos from global and local perspectives simultaneously. To be specific, NVE learns a set of multi-scale temporal embeddings within the visual space during the reconstruction phase. These embeddings are subsequently directly injected into the attention module of the editing phase, explicitly augmenting the temporal consistency of the entire video. On the other hand, WFA enhances local temporal modeling by dynamically optimizing attention mechanisms between adjacent frames, which improves temporal coherence with reduced computational costs. Experimental evaluations show that VidEvo enhances frame-to-frame temporal consistency. Ablation studies confirm NVE and WFA's effectiveness and their plug-and-play capability with other methods.

1 Introduction

Recent progress in large-scale diffusion models [Dhariwal and Nichol, 2021] represents a notable shift in AI-Generated Content (AIGC), surpassing the capabilities of GANs [Reed *et al.*, 2016] and auto-regressive models [Ramesh *et al.*, 2021]. Leading-edge models including GLIDE [Nichol *et al.*, 2022], Imagen [Saharia *et al.*, 2022], Stable Diffusion [Rombach *et al.*, 2022], and DALL-E2 [Ramesh *et al.*, 2022] have advanced the frontiers of image generation, offering higher fidelity and complexity. This development has set the stage for text-guided image editing (TGIE) techniques, including ControlNet [Zhang *et al.*, 2023], Plug-and-Play [Tumanyan *et al.*, 2023], and Prompt-to-Prompt (P2P) [Hertz *et al.*, 2022], which further enhance user control over image generation and improve editing precision. Building on these advancements, the editing focus is now shifting towards dynamic video content [Dang *et al.*, 2024c]. This transition highlights the need for text-guided video editing (TGVE) architectures, which bridge the gap between textual prompts and video editing.

(a) Source video (b) vid2vid (c) TAV (d) Ours

Figure 1: Comparison of one-shot video reconstruction. The blue box highlights content preservation errors, and the yellow box marks temporal inconsistencies between frames.

In addressing TGVE challenges, one approach involves training models directly with large text-video datasets, though limitations like dataset scarcity and high computational costs restrict accessibility for many researchers. Alternatively, utilizing pre-trained TGIE models offers a more cost-effective strategy. This approach expands the TGIE model from 2D to 3D, with the core of this approach being the application of temporal modeling. Specifically, the TGIE-expended ap-

*Corresponding authors. This work was completed during the internship at SGIT AI Lab, State Grid Corporation of China.

Figure 2: Video editing utilizing our VidEvo framework. We demonstrate content editing (second row) and style transformation (third row).

proach can be grouped into three categories: a) Train-based methods [Guo *et al.*, 2023], which involve training a temporal module on large datasets without retraining the entire model. b) Zero-shot methods, which adapt existing text to image (T2I) architectures to video without additional training. c) One-shot methods, which fine-tune pre-trained T2I models using just one video instance to enhance temporal representation. Given train-based methods' high training demand and zero-shot methods' challenges in accurately capturing temporal details from the source video, we opt for the one-shot method. This method balances lower training requirements with effective temporal modeling.

TGVE typically involves reconstruction and editing phases, both using U-Net as the backbone. Previous one-shot methods struggle with robust temporal modeling (e.g., video shaking or jumping) and more severe content preservation issues during reconstruction, which impacts the subsequent editing phase negatively, as shown in Fig. 1. The vid2vid method utilizes null-text inversion but suffers from a lack of global temporal modeling, leading to content preservation problems. The TAV method, which tunes the model, yields approximate reconstruction results but lacks precision due to insufficient local temporal modeling. The core issue lies in the simplistic and implicit tuning of existing attention modules in U-Net over a single video, which exerts limited influence on attention and inadequately captures the temporal consistency. Hence, a critical question emerges: **How can we extract more temporal information from a single video?**

A straightforward idea, prompted by the limitations of implicitly tuning existing attention modules in U-Net, is to explore explicit cues from the original video to enhance both global and local temporal consistency [Dang *et al.*, 2024a]. In light of this, we developed the VidEvo model, which introduces two key innovations: the null-video embedding (NVE) and the window-frame attention (WFA), to achieve improved video editing outcomes, as illustrated in Fig. 2. Specifically, the NVE is designed to capture global temporal features, providing a robust foundation for maintaining coherence across the entire video sequence. Diverging from previous methods that integrate null embedding into the text space [Mokady *et al.*, 2023], primarily designed for image editing tasks, our method explicitly learns a set of multi-scale temporal embeddings during the reconstruction phase in the visual space to capture frame-to-frame consistency. Rather than implicitly tuning attention, we explicitly integrate this embedding into the temporal attention to enhance coherence during the editing phase. Concurrently, the WFA module enhances local temporal cohesion by seamlessly integrating structural details with the context of adjacent frames. This method not only preserves structural integrity within frames but also ensures smooth transitions between them, efficiently reducing computational overhead compared to traditional self-attention mechanisms. It is noteworthy that NVE and WFA offer plug-and-play integration with existing video editing methods, also supporting methods like TokenFlow [Geyer *et al.*, 2024] and SVD [Blattmann *et al.*, 2023] that require optical flow and motion information. Our key contributions include:

- We introduce a novel text-guided one-shot video editing method, dubbed VidEvo. It incorporates exhaustive temporal modeling in both the reconstruction and editing phase, enhancing temporal consistency and content preservation in video editing.
- Our methodology features NVE in the reconstruction process to improve global frame coherence and enrich the source video's representation for the subsequent editing. Simultaneously, WFA is applied across both reconstruction and editing phases. This adaptation enhances temporal modeling, ensuring frame-to-frame consistency and preserving local details and textures in the target video.
- Our experimental analysis showcases our method's efficiency and robust performance in enhancing frame consistency and textual alignment. Moreover, ablation studies and orthogonality tests reveal the effectiveness and plug-and-play capability of our modules.

2 Related Work

The landscape of AIGC has seen a significant shift with the advent of diffusion models in T2I tasks, outperforming GANs and auto-regressive models. This evolution has extended to video synthesis [Ho *et al.*, 2022; Wang *et al.*, 2023a; Wang *et al.*, 2023c], where various control signals drive model performance. Among these, text-guided approaches have received extensive focus, evidenced by models like pose-guided [Ma *et al.*, 2023] and motion-guided [Hu and Xu, 2023], signifying a broad spectrum of research in controlled video generation. Similar to video understanding tasks [Dang *et al.*, 2023a; Dang *et al.*, 2023b; Dang *et al.*, 2024b], the progression from TGIE to TGVE synthesis is underscored by the critical need for effective temporal modeling to produce coherent video content.

Training-based methods focus on enhancing temporal consistency by introducing specialized layers or adapters into T2I models. Notable examples include GEN-1 [Esser *et al.*, 2023], Control-A-Video [Chen *et al.*, 2023], and AnimateDiff, which integrate temporal dynamics into T2I diffusion models. These methods, while adept at controlling video content, typically require extensive datasets (e.g., WebVid [Bain *et al.*, 2021]) and substantial computational resources.

Zero-shot methods utilize pre-trained models in a zero-shot manner, negating the need for additional training in video editing tasks. This category includes FateZero [Qi *et al.*, 2023], T2V-Zero [Khachatryan *et al.*, 2023] and Pix2Video [Ceylan *et al.*, 2023], which maintain motion fidelity through spatial-temporal blocks and self-attention features. Additionally, InFusion [Khandelwal, 2023] presents an approach by integrating residual and attention features specific to the edit prompt, enhancing zero-shot editing capabilities while ensuring uniformity across edited and unedited aspects. However, these methods can encounter spatio-temporal distortions in practical video editing scenarios.

One-shot-tuned methods, like ControlVideo [Zhao *et al.*, 2023] and Tune-A-Video (TAV) [Wu *et al.*, 2023], strike a balance by fine-tuning pre-trained T2I models on specific video instances. Video-P2P [Liu *et al.*, 2023] and Vid2vid-Zero [Wang *et al.*, 2023b] further adapt this approach, employing null-text inversion and specialized attention mechanisms for bidirectional temporal modeling. These methods, while efficient, still confront challenges in computational intensity and internal frame structure focus. Due to the different editing structures and high resource overhead of One-shot-tuned methods like TokenFlow and SVD, we primarily conducted experiments on P2P-based editing methods.

3 Method

In this section, we introduce VidEvo, a framework designed for real-world text-guided video editing. Through exhaustive global and local temporal modeling, our proposed method preserves the reconstruction quality of the original video while achieving vivid and realistic video editing effects. Given a source video sequence $\mathcal{V} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$ of n frames, a source prompt $\mathcal{P}$ and a target editing text prompt $\mathcal{P}^E$, our framework needs to generate a modified target sequence $\mathcal{V}^E = \{\mathcal{I}_1^E, \mathcal{I}_2^E, \ldots, \mathcal{I}_n^E\}$ that fulfills the textual editing requirements. For instance, to edit a video with $\mathcal{P}$=*"jeep car"* to $\mathcal{P}^E$=*"red toy car"*, the user simply alters the prompt accordingly while VidEvo will be responsible for maintaining the video's structure and temporal consistency, as shown in Fig. 2.

Algorithm 1 VidEvo video editing

Input: A source video $\mathcal{V}$, a target prompt $\mathcal{P}^E$
Output: An edited video $\mathcal{V}^E$

$z_T \sim N(0, I)$ a unit Gaussian random variable
Latent features from DDIM inversion: $\{z_T^*, \ldots, z_0^*\}$
Initialize $z_T^I \leftarrow z_T^*$, $\varnothing_T \leftarrow \mathbf{0}$, $\mathcal{P} \leftarrow \mathcal{F}(\mathcal{P}^E, \mathcal{V})$;
for $t = T$ **to** 1 **do**
  for $j = 0$ **to** $N-1$ **do**
    $\varnothing_t \leftarrow \varnothing_t - \eta \nabla_{\varnothing} \left\| z_{t-1}^* - z_{t-1}(z_t^I, \varnothing_t, \mathcal{P}) \right\|_2^2$
  end for
  $z_{t-1}^I \leftarrow z_{t-1}(z_t^I, \varnothing_t, \mathcal{P})$, $\varnothing_{t-1} \leftarrow \varnothing_t$
end for
Initialize $z_T^E \leftarrow z_T^*$
for $t = T$ **to** 1 **do**
  if $t < \tau_{null}$ **then**
    $\varnothing_t \leftarrow \varnothing$
  end if
  $z_{t-1}^*, M_t \leftarrow DM(z_t^*, \mathcal{P}, t, \varnothing_t)$
  $M_t^E \leftarrow DM(z_t^E, \mathcal{P}^E, t, \varnothing_t)$
  $\hat{M}_t \leftarrow Edit(M_t, M_t^E, t)$
  $z_{t-1}^E \leftarrow DM(z_t^E, \mathcal{P}^E, t, \varnothing_t)\{M_t \leftarrow \hat{M}_t\}$
end for
$\mathcal{V}^E \leftarrow \mathcal{D}\{z_0^E\}$
return $\mathcal{V}^E$

The whole editing process of our method consists of two phases: reconstruction and editing, as shown in Fig. 3.

Reconstruction. This phase aims to convert $\mathcal{V}$ into latent representations, which should have the ability to reconstruct the source video itself. Related existing methods [Liu *et al.*, 2023; Wang *et al.*, 2023b] employing null-text inversion in this phase, which struggles with temporal inconsistency and content distortion as shown in Fig. 7. We address this with our novel null-video inversion technique (detailed in Section 3.1), which learns global temporal embeddings to represent the video structures and keep the temporal continuity.

Editing. This phase utilizes the attention control method P2P to achieve zero-shot video editing. However, directly applying these strategies or existing improved methods [Liu *et al.*, 2023; Wang *et al.*, 2023b] still encounters issues such as temporally inconsistent editing results or high computational costs. Instead, we propose WFA (discussed in section 3.2) to specifically enhance the focus on local temporal features, aiming to simultaneously maintain temporal continuity and reduce computational overhead. Furthermore, section 3.3 outlines how we explore such editing transformations effectively executed within VidEvo.

3.1 Global Null-Video Inversion

To effectively edit real-world videos, video reconstruction is an essential first step. The goal here is to develop a robust representation of the temporal relationships within the source

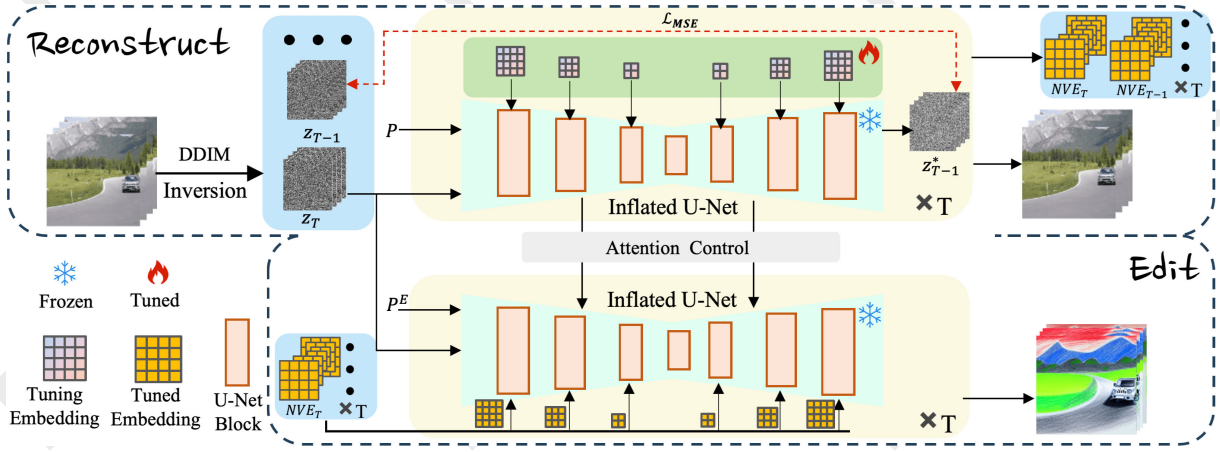

Figure 3: **VidEvo Pipeline Overview**. The reconstruction phase (top) utilizes DDIM inversion to derive latents from a video, optimizing NVE sequence. In the editing phase, these embedding facilitate global temporal modeling for enhanced reconstruction, while the attention map manipulation enables precise video editing. Source Prompt P and Target Prompt P^E, see Fig. 2 for reference. More details about the U-Net block can be seen in Fig. 5.

video for subsequent editing. Although an effective DDIM inversion scheme has been suggested for unconditional diffusion models, it falls short for text-guided diffusion models, especially when classifier-free guidance is necessary for meaningful editing. Given the success of null-text inversion [Mokady *et al.*, 2023] in image reconstruction, vid2vid directly applies it to video reconstruction. Unfortunately, this approach struggles to capture temporal information between frames (as depicted in Fig. 1). This is because the optimized embedding is the same for all video frames, unable to capture the temporal information between frames. To address this limitation, we introduce NVE, an innovative method that enhances video reconstruction and strengthens global temporal modeling. By incorporating our embedding within the image channel and optimizing on the unconditioned branch, we leverage the richer information capacity of images.

Next, we formalize this process. The diffusion model [Ho *et al.*, 2020; Sohl-Dickstein *et al.*, 2015] serves as the foundation of our approach. The network ϵ_θ is trained to predict artificial noise, following the objective:

$$\min_\theta \mathbb{E}_{z_0\sim\mathcal{N}(0,I),t\sim\text{Uniform}(1,T)} \left\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{C})\right\|_2^2, \tag{1}$$

where $\mathcal{C}$ represents control conditions such as the text condition, and z_t is a noised sample where noise is added to the sampled data z_0 according to time step t. During inference, starting from the noisy state z_T, noise is gradually removed using the noise predicted by ϵ_θ to obtain z_0. We employ the deterministic DDIM sampling:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(z_t, t, \mathcal{C}), \tag{2}$$

where α_t are the noise scales. z_0 is the latent space representation of the real image frame x_0, obtained by mapping through an image encoder $z_0 = \mathcal{D}(x_0)$, and an image decoder is employed at the end of the diffusion backward process to reconstruct x_0 from z_0.

In text-guided generative tasks, it is crucial to enhance the influence of the textual prompt on $\epsilon_\theta(z_t, t, \mathcal{C})$. Classifier-free guidance [Ho and Salimans, 2021] achieves this by interleaving unconditional and text-conditioned predictions. Given an empty text prompt embedding $\psi(\text{""})$ and a guidance scale w, the classifier-free guidance output is formulated as:

$$\tilde{\epsilon}_\theta(z_t, t, \mathcal{C}, \psi(\text{""})) = w\cdot\epsilon_\theta(z_t, t, \mathcal{C}) + (1-w)\cdot\epsilon_\theta(z_t, t, \psi(\text{""})), \tag{3}$$

Through the above formulation, we can generate video frames that align with textual descriptions from noise. However, this alone does not achieve our goal of reconstructing real video by extracting information from the original video. We first use DDIM inversion [Song *et al.*, 2020] to obtain the results of each step from z_0 to z_T:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_\theta(z_t, t, \mathcal{C}), \tag{4}$$

We can then begin our optimization.

In practice, we optimize frames at a guidance scale of $w = 1$, yielding a series of pivotal latent codes $\{z^*_{t,i}\}^{T,n}_{t=0,i=1}$, where i indexes each of the n frames within the time step t. For simplicity, we denote these as $\{z^*_t\}^T_{t=0}$. NVE is defined for each temporal step t as $\varnothing_t$, with the aim of optimizing the set $\{\varnothing_t\}^T_{t=1}$.

The optimization begins with z^I_t initialized as z^*_T, the starting point of our process. At each subsequent time step t, descending from T to 1, we perform N iterations of optimization. Employing DDIM inversion with a default guidance scale of $w = 7.5$, our objective at each time step t is to minimize the following:

$$\min_{\varnothing_t} \left\|z^*_{t-1} - z_{t-1}(z^I_t, \varnothing_t, \mathcal{C})\right\|^2. \tag{5}$$

After each iteration, z^I_{t-1} is updated according to:

$$z^I_{t-1} = z_{t-1}(z^I_t, \varnothing_t, \mathcal{C}). \tag{6}$$

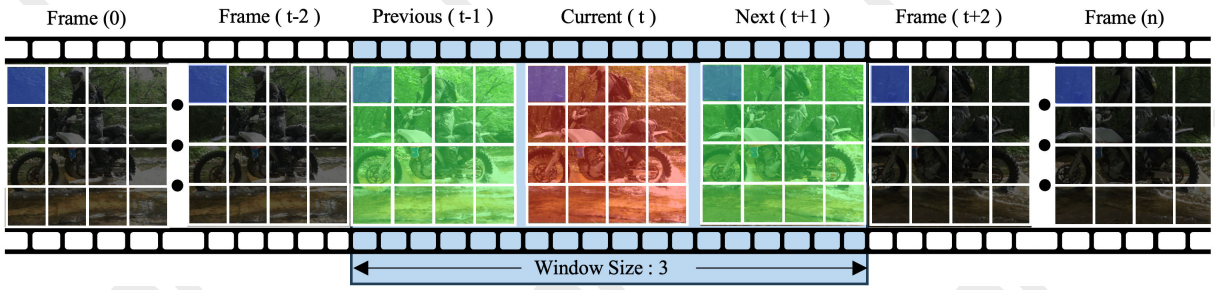

Figure 4: **Illustration of the WFA mechanism.** When the first token of the current frame serves as the query, traditional self-attention consider only the red tokens as key, while TA takes into account all blue tokens as key, and STA utilizes all the tokens as key. In contrast, our WFA selectively focuses on the red and green tokens within a specified window size as key, thereby bolstering local temporal modeling by integrating immediate sequential context with the targeted frame's position.

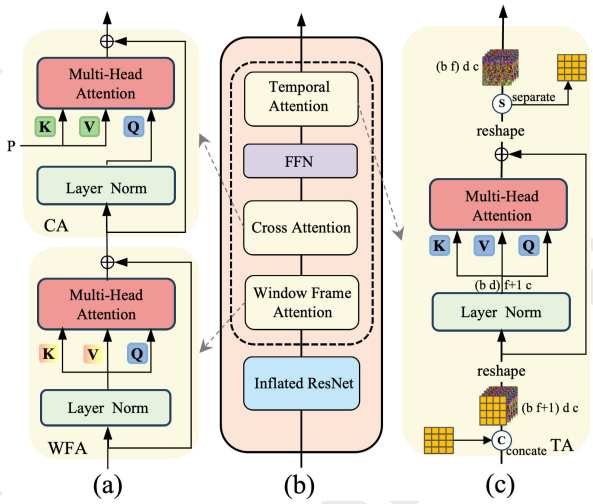

Figure 5: More details about the U-Net block. (a) WFA and cross attention. (b) U-Net block. (c) TA where we add NVE.

Having optimized the NVE, we can now proceed with zero-shot editing without altering the network structure. The embedding is simply injected into the U-Net. Unlike textual elements that predominantly affect the cross-attention mechanism, our image channel impacts all attention and ResNet blocks. Given our primary goal of interacting temporal information across all frames with NVE, we strategically inject it into the temporal attention within the attention block, as demonstrated in Fig. 5 (c). To better illustrate the temporal information captured by the NVE, we visualize the outer layer embeddings from the final time step, as shown in Fig. 6. The fifth column visualizes the impact of NVE in editing. It demonstrates that this embedding provides global temporal information for the initial consecutive frames. For a more discussion on the timing and placement of NVE injection, please refer to the section 4.4.

3.2 Local Window-Frame Attention

Local temporal modeling in video editing presents unique challenges, particularly due to the dynamic nature of video content. While traditional 2D diffusion U-Nets are adept at capturing intra-frame spatial relationships, their extension to 3D self-attention often falls short in addressing the critical temporal dimension required for video continuity (see Fig. 4).

Key-frame attention [Qi *et al.*, 2023] and sparse-causal attention [Wu *et al.*, 2023] (SCA) have shown promise in video generation by focusing on key frames and their immediate predecessors. However, their application in video editing is limited by an underutilization of temporal progression in the source material. Vid2vid's spatio-temporal attention (STA) aims to rectify this by introducing bi-directional temporal modeling, where each frame element considers the entirety of elements in all other frames. This approach, while thorough, can become computationally intensive with an increasing number of frames and may introduce irrelevant information in cases of significant content variation.

To address these issues, we propose the WFA mechanism as an alternative to traditional self-attention. As shown in Fig. 4, this mechanism uses a window size of λ (e.g., 3) to allow each token (e.g., the blue token in frame x_t) to be influenced by tokens within this window span (e.g., the red and green tokens). Formally, with a window λ, we establish boundary points at $begin = t - \frac{\lambda-1}{2}$ and $end = t + \frac{\lambda-1}{2}$, allowing frame x_t to focus on frames $x_{begin:end}$. The query (Q), key (K), and value (V) computations in this context are redefined as:

$$Q = W^Q x_i, \quad K = W^K x_{\text{begin:end}}, \quad V = W^V x_{\text{begin:end}}.$$

Here, W^Q, W^K, and W^V are the pre-trained projection weights within the self-attention layers, shared across spatial and temporal tokens.

This WFA (using a window size of 3) enables our VidEvo to effectively enhance local temporal details between adjacent frame information. Further discussion on its impact is provided in Section 4.4.

3.3 Attention Control

Once the optimized video embeddings are obtained and the 3D U-Net architecture is adapted, VidEvo is primed to manipulate attention for achieving precise video editing outcomes. The attention control mechanism between reconstructed and target edited videos is a common strategy to enable desired modifications. At a single-frame level, this is achieved through a P2P control scheme, typical in image editing. For example, a word swap operation is represented as:

Method	Frame consistency			Textual alignment		Runtime [min]		Memory [GB]
	CLIP Score↑	FID↓	User Vote↑	CLIP Score↑	User Vote↑	Training	Inference	Max Memeory
T2V-Zero	0.921	30.17	17.3%	0.901	19.1%	**0**	1.5	28.9
TAV	0.941	26.98	18.8%	0.802	19.6%	10.3	0.6	**10.4**
Vid2vid	0.942	26.74	18.0%	0.786	16.8%	11.1	2.2	18.2
Video-P2P	0.947	25.03	19.9%	0.889	20.9%	15.8	1.5	29.4
VidEvo	**0.969**	**23.32**	**26.2%**	**0.904**	**23.7%**	7.2	**1.5**	23.8

Table 1: Comparison of P2P-based Video Editing Methods. We reported only Max Memeory for simplicity.

Figure 6: Visualization of the NVE. The fifth column illustrates the global information captured by NVE.

$$Edit\left(M_t, M_t^E, t\right) := \begin{cases} M_t^E & \text{if } t < \tau \\ M_t & \text{otherwise} \end{cases}, \tag{7}$$

where M_t and M_t^E are the attention maps of the original and edited videos at time step t, and τ is the threshold time step.

The full VidEvo algorithm is presented in Algorithm 1. Here, $\mathcal{F}$ denotes the video caption model or GPT-4 used for generating prompts, τ_{null} is a parameter determining the timing for injecting null-video embeddings, and DM refers to the use of the Stable Diffusion Model. The function *Edit* denotes the attention editing operations of the P2P method. For further details on the *Edit* function, we suggest reading the article on the P2P method for a complete understanding.

4 Results

In this section, we present quantitative and qualitative analyses, ablation studies, and orthogonality analyses. Our method is primarily evaluated on the DAVIS [Pont-Tuset *et al.*, 2017] dataset for comparison with existing works.

4.1 Quantitative analysis

Since there is no unified quantitative benchmark for video editing, we performed our evaluation using the pre-trained CLIP model [Radford *et al.*, 2021], FID [Heusel *et al.*, 2017] and user studies. Similar to GROUND-A-VIDEO [Jeong and Ye, 2023], for textual alignment, we calculate average cosine similarity between the target prompt and the edited frames. For frame consistency, we compute CLIP image features for all frames in output video and then calculate the average cosine similarity between all pairs of video frames.

Automatic evaluation metrics such as CLIP score only roughly reflect human judgment. Therefore, we also conducted user studies on frame consistency and textual alignment. For this, we enlisted 100 participants to rank the editing outcomes based on frame-to-frame consistency and the degree to which the edited results match the target prompt. As indicated in Table 1, our proposed VidEvo method outperforms all competing methods in terms of frame consistency and textual alignment. Particularly for frame consistency, our method achieves a 2% increase in CLIP Score and a 6.8% reduction in FID compared to the state-of-the-art one-shot methods. Moreover, VidEvo accomplishes this without incurring excessive memory costs and even accelerates inference speeds through the implementation of WFA.

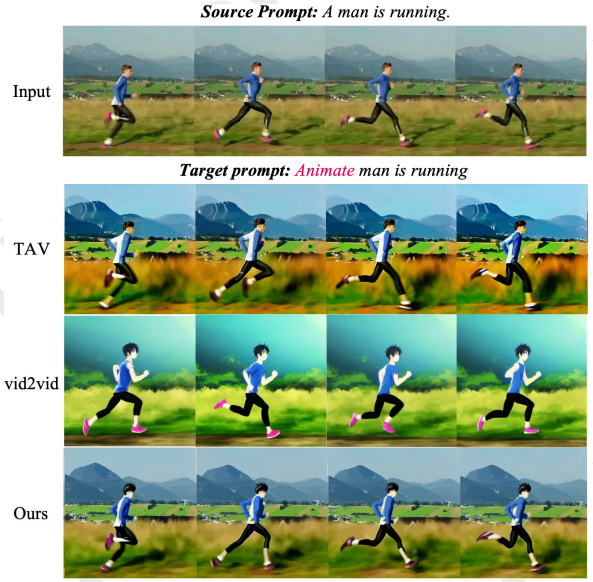

Figure 7: Comparison with other methods. Our VidEvo achieves both temporal consistency and fidelity to the source video.

4.2 Qualitative analysis

The results of various methods applied to subject editing are shown in Fig. 7. To effectively assess temporal consistency, we selected videos with substantial motion. The TAV method, having undergone temporal modeling and fine-tuning, generally exhibits better frame-to-frame consistency, though with some discrepancies noticeable in the second column compared to the input video. However, it falls short in the degree of character cartoonization. The vid2vid approach aligns well with the editing objective, yet shows weaker temporal coherence. Our VidEvo method strikes a balance, achieving cartoonization of characters while maintaining temporal con-

Method	CLIP-F↑	FID↓	CLIP-T↑	RT	MEM
P2P-direct	0.853	34.71	0.807	14.5	18.4
w/o NVE	0.902	27.47	0.835	1.5	20.5
w/o WFA	0.864	27.18	0.822	11.2	23.8
Ours	0.969	23.23	0.904	8.7	23.8

Table 2: Ablation Study

Method	CLIP-F↑	FID↓	CLIP-T↑
TAV+VidEvo	+0.029	-3.81	+0.002
vid2vid+VidEvo	+0.008	-1.41	+0.111
TokenFlow+VidEvo	+0.021	-0.91	+0.094
SVD+VidEvo	+0.006	-0.71	+0.014

Table 3: Module Orthogonality Validation

sistency and fidelity to the input video.Our VidEvo framework displays commendable results in the localized editing of structure and color, as well as in the global editing of style.

4.3 Ablation

After showcasing the VidEvo's capabilities, this section focuses on the ablation study of various components within our framework as shown in Fig. 8 and Table 2. Table 2 lists the metrics, with CLIP-F and CLIP-T denoting Frame consistency and Textual alignment under CLIP Score. Runtime is calculated as the total of training and inference times.

The first row of Fig. 8 displays the result of directly using P2P, the fourth row highlights the editing effects achieved by the full VidEvo framework, and rows two to three present the results of component-specific ablations. The P2P method performs well within individual frames but lacks consistency across frames, as seen in the third frame where the background shifts from mountains to a city, and in the fourth frame where part of the car turns into a road. The row titled *w/o NVE* represents an ablation of the NVE. This variant exhibits a clear lack of global temporal consistency, not only blurring the entire jeep car but also failing to effectively edit the background, contrary to our intention of transforming the *road* into *snow*. The *w/o WFA* row illustrates the ablation of WFA. The outcome aligns the overall background with the intended snow editing. However, it introduces significant inconsistencies between adjacent frames, such as the complete disappearance of the car in the third frame. This underscores the critical role of WFA, which accounts for temporal information between neighboring frames, achieving effective local temporal modeling.

4.4 Orthogonality and Parameter Analyses

We performed an orthogonality analysis to demonstrate the plug-and-play flexibility of NVE and WFA with other methods. As shown in Table 3, when NVE and WFA were applied to TAV and vid2vid, there was a measurable enhancement in temporal consistency with a 3% increase in CLIP Score and a 7% reduction in FID. These two modules also improve the performance of methods like TokenFlow and SVD, which do not rely on P2P-based editing.

Figure 8: Ablations on direct using P2P (1st row) and the effectiveness of each component in VidEvo (2nd $\sim$ 3rd row). Prompt and source video see Fig. 2.

Different from max memory, real-time memory consumption involves multiple components, including pipeline memory (influenced by base model size, batch inference frame number, and resolution) and memory for the features being tuned. For instance, when fixing the frame count at 8, Video-P2P utilizes 10.4GB for pipeline memory and 19GB for tuning; an SVD-based method uses 22.6GB for pipeline memory and 48GB for motion tuning; and VidEvo employs 17.2GB for pipeline memory alongside 6.6GB for NVE tuning.

What's more, we further analyzed the positioning and hyperparameters of NVE and WFA. For the precise timing and placement of NVE injection, we found that symmetrically adding NVE at both shallow and deep layers of the U-Net is most effective. It is beneficial to apply injection at only the coarsest layer within each U-Net block for efficiency.

Our ablation studies reveal that for videos with minimal motion, a window size of 3 for our WFA effectively maintains temporal consistency and achieves robust results without significant computational overhead. Additionally, placing temporal attention after the Feedforward Network proves most conducive to video editing, synergizing with textual prompts to ensure stable and coherent editing transitions.

5 Conclusion

This paper introduces VidEvo, a one-shot framework designed specifically to address the complex challenges of text-guided real-world video editing, with significant progress made in enhancing temporal coherence. We design two exhaustive temporal modeling modules for both the reconstruction and editing phases, including global NVE and local WFA. They ensure high temporal consistency in both the global video structure and local temporal details of the edited target videos. Moreover, VidEvo's plug-and-play characteristic allows it to integrate seamlessly with various video editing methods, yielding superior results.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (NSFC) [grant numbers 72274152, 62403429 and 62202367] and the Zhejiang Provincial Natural Science Foundation of China [grant number LQN25F030008].

References

[Bain *et al.*, 2021] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[Ceylan *et al.*, 2023] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[Chen *et al.*, 2023] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023.

[Dang *et al.*, 2023a] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32:3924–3938, 2023.

[Dang *et al.*, 2023b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4512–4526, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Esser *et al.*, 2023] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[Geyer *et al.*, 2024] Michal Geyer, Omer Bar-Tal, Shai Bagon, Tali Dekel, et al. Tokenflow: Consistent diffusion features for consistent video editing. *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, 2024. Accepted.

[Guo *et al.*, 2023] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Ho and Salimans, 2021] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Ho *et al.*, 2022] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*. NeurIPS, 2022.

[Hu and Xu, 2023] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023.

[Jeong and Ye, 2023] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[Khachatryan *et al.*, 2023] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023.

[Khandelwal, 2023] Anant Khandelwal. Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2023.

[Liu *et al.*, 2023] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023.

[Ma *et al.*, 2023] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Nichol *et al.*, 2022] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[Pont-Tuset *et al.*, 2017] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[Qi *et al.*, 2023] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2020.

[Tumanyan *et al.*, 2023] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[Wang *et al.*, 2023a] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[Wang *et al.*, 2023b] Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.

[Wang *et al.*, 2023c] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023.

[Wu *et al.*, 2023] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 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*, pages 7623–7633, 2023.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhao *et al.*, 2023] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023.