SCIENCE CHINA
Information Sciences

• LETTER •

October 2025, Vol. 68, Iss. 10, 200112:1–200112:2
https://doi.org/10.1007/s11432-024-4596-1

Special Topic: Large Multimodal Models

LOVECon: text-driven training-free long video editing with ControlNet

Zhenyi LIAO[1], Qingsong XIE[2] & Zhijie DENG[1*]

[1]*School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*OPPO Guangdong Mobile Telecommunications Co., Ltd., Dongguan 523902, China*

Received 14 May 2024/Revised 29 April 2025/Accepted 21 August 2025/Published online 28 September 2025

Citation Liao Z Y, Xie Q S, Deng Z J. LOVECon: text-driven training-free long video editing with ControlNet. Sci China Inf Sci, 2025, 68(10): 200112, https://doi.org/10.1007/s11432-024-4596-1

Diffusion models have gained tremendous popularity in image generation. Recently, leveraging pre-trained conditional diffusion models for training-free video editing has gained increasing attention due to its promise in film production, advertising, etc. Yet, seminal studies in this line fall short in generation length, temporal coherence, and fidelity to the source video.

In this work, we aim to bridge the gap, establishing a simple and effective baseline for text-driven training-free LOng Video Editing with ControlNet, dubbed as LOVECon. Technically, LOVECon follows the basic video editing pipeline based on Stable Diffusion [1] and ControlNet [2], with an additional step of splitting long videos into consecutive windows to accommodate limited computational memory. On top of these, we also introduce a novel cross-window attention mechanism to maintain coherence in style and subtleties across windows. To ensure the structural fidelity to the original source video, we enrich the latent states of edited frames with information extracted from the source video through DDIM inversion [3]. Additionally, LOVECon incorporates a video interpolation model, which polishes the latent states of the edited frames in the late stages of generation, to alleviate frame flickering. These techniques contribute to smoother transitions in long videos and significantly mitigate visual artifacts.

Methodology. As illustrated in Figure 1, given a video of M frames, we evenly split it into multiple consecutive windows of size K, and use $\tilde{x}^{i,j}$ ($x^{i,j}$) to refer to the j-th source (edited) frame in the i-window, $i \in [1, M/K]$, $j \in [1, K]$. The editing is governed by a source prompt p_{src} and a target prompt p_{tgt}. p_{obj} denotes the specific tokens in p_{src} that indicate the editing objects. $\tilde{z}_t^{i,j}$ and $z_t^{i,j}$ denote the latent states of $\tilde{x}_t^{i,j}$ and $x_t^{i,j}$ in the latent diffusion models, with subscript t representing t-th timestep. We denote the features corresponding to $z_t^{i,j}$ in the model as $h_t^{i,j} \in \mathbb{R}^{L\times D}$ with L referring to the sequence length.

Cross-window attention. Original Stable Diffusion models leverage self-attention and apply computations to the L tokens in $h_t^{i,j}$ to account for intra-image attention. Existing studies [4] refurbish it as a spatial-temporal one to include inter-image information for video editing, using $[h_t^{i,j}, h_t^{i,K/2}] \in \mathbb{R}^{2L\times D}$ to construct the key and value matrices for attention computation, where $h_t^{i,K/2}$ refers to the feature of the middle frame in the window of concern. However, such a strategy still cannot tackle the inter-window inconsistency in long video editing.

We advocate further improving it by including more contextual information. Specifically, we extend $h^{i,j}$ to $[h_t^{1,1}, h_t^{i-1,K}, h_t^{i,j}, h_t^{i,K}] \in \mathbb{R}^{4L\times D}$ for computing the key and value matrices in the attention, where $h_t^{1,1}, h_t^{i-1,K}$, and $h_t^{i,K}$ denote the features of the first frame of the first window, the last frame of the previous window, and the last frame of the current window, respectively. The first one provides guidance on global style for the frames in the current window, and the others govern the variation dynamics. Compared to video editing methods relying on costly fully cross-frame attention, which performs attention over all frames in the window, our method can significantly reduce memory usage while keeping global consistency.

Latent fusion. To maintain the structural information of the source video, we propose to fuse the latent states of the source frames with those of the edited ones. Specifically, at t-th denoising step, there is

$$z_t^{i,j} = m^{i,j} \odot z_t^{i,j} + (1 - m^{i,j}) \odot \tilde{z}_t^{i,j}, \tag{1}$$

where $\odot$ is the element-wise multiplication, $m^{i,j}$ denotes a time-independent mask to identify the regions that require editing, and $\tilde{z}_t^{i,j}$ denotes the outcomes of DDIM inversion [3] for the source frames. Besides, $m^{i,j}$ should not be specified manually as that can be laborious and time-consuming for long videos. Instead, we estimate $m^{i,j}$ by (1) collecting the cross-attention maps between textual features of p_{obj} and visual features of $\tilde{z}_t^{i,j}$ during the DDIM inversion procedure, (2) binarizing the time-dependent cross-attention maps via thresholding, and (3) aggregating them into a global binary one via pooling.

Nonetheless, Eq. (1) totally discards the structural information of the source frames in the masked regions, but

* Corresponding author (email: zhijied@sjtu.edu.cn)

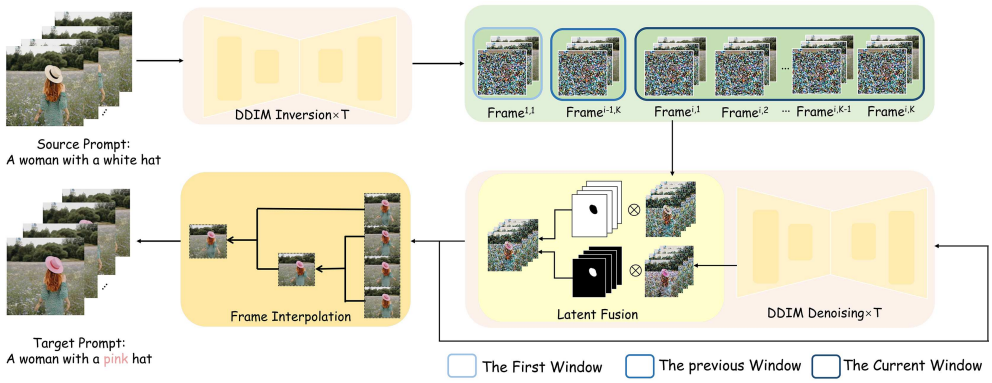

Figure 1 (Color online) Method overview. LOVECon is built upon Stable Diffusion and ControlNet (omitted in the plot for simplicity) for long video editing. LOVECon splits the source video into consecutive windows and edits sequentially, where cross-window attention is employed to improve inter-window consistency. LOVECon fuses the latent states of the edited frames with those of the source frames from DDIM Inversion [3] to maintain the structure of the source video. LOVECon further incorporates a video interpolation model to address the frame-level flickering issue (omitted it during the latent fusion stage for simplicity).

such information could be beneficial in the early stages of the reverse diffusion process. To mitigate this, we refine the latent fusion mechanism as

$$z_t^{i,j} = m^{i,j} \odot [\gamma z_t^{i,j} + (1-\gamma)\tilde{z}_t^{i,j}] + (1-m^{i,j}) \odot \tilde{z}_t^{i,j}, \quad (2)$$

where $\gamma \in (0,1)$ represents the trade-off coefficient and will be set to 1 after some timestep T_0.

Frame interpolation. While the cross-window attention and latent fusion mechanisms effectively preserve the global style and finer details, we still observe that the generation may suffer from frame-level flickering issues.

To address this, we introduce a video frame interpolation model to polish the latent states $z_t^{i,j}$. Given that interpolation models usually operate in the pixel space, we first project $z_t^{i,j}$ back to images $x_{t\to 0}^{i,j}$ with the one-step DDIM sampling [5] and the decoder of the latent diffusion model. Given these, the video interpolation model ψ produces new frames via

$$\dot{x}_{t\to 0}^{i,j} = \psi(\dot{x}_{t\to 0}^{i,j-1}, x_{t\to 0}^{i,j+1}), \ j = 2, \ldots, K-1, \quad (3)$$

where $\dot{x}_{t\to 0}^{i,1} := x_{t\to 0}^{i,1}$ and $\dot{x}_{t\to 0}^{i,K} := x_{t\to 0}^{i,K}$. $\dot{x}_{t\to 0}^{i,j}$ contains almost the same contents as $x_{t\to 0}^{i,j}$ while being more smoothing in the temporal axis. We map them back to the latent space via the encoder $\mathcal{E}$ of the latent diffusion model for the following sampling process.

Nonetheless, the repetitive use of the encoder and the decoder can lead to degradation in image quality and an accumulation of information loss. We empirically address this issue by performing interpolation only two times, once at the last timestep of the reverse diffusion process within the window and once after the process within the whole video. We select these two positions because the generated images contain less noise at the end of the denoising process, which enables more reasonable frame interpolation. Compared to editing without interpolation, albeit with a slightly increased processing time, the issue of flickering is significantly reduced.

Experiments. To verify the effectiveness of our framework, we compare it with recent baselines such as Text2Video-Zero [4]. We conduct extensive evaluations regarding video-text alignment, temporal consistency, and fidelity to the source video.

The experimental results demonstrate the following. (1) Our framework received greater preference across all metrics in the user study and achieved the best results in 3 out of 4 metrics for object evaluation, despite a lower score in video-text alignment. This indicates the overall effectiveness of our approach. (2) Improved results in temporal coherence, as reflected in both the object metrics and the user study, highlight the effectiveness of the cross-window attention and frame interpolation modules. Additionally, enhanced fidelity results underscore the impact of the latent fusion technique. More details and results of the experiments are in the supplementary files.

Conclusion. In this work, we propose a video editing framework called LOVECon, which consists of three modules: cross-window attention for maintaining temporal coherence in long video editing, latent fusion to ensure fidelity to the source video, and frame interpolation to address flickering issues.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62306176, 92470118), Natural Science Foundation of Shanghai (Grant No. 23ZR1428700), CCF-ALIMAMA TECH Kangaroo Fund (Grant No. CCF-ALIMAMA OF 2025010), and CCF-Zhipu Large Model Innovation Fund (Grant No. CCF-Zhipu202412).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695
2 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3836–3847
3 Mokady R, Hertz A, Aberman K, et al. Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 6038–6047
4 Khachatryan L, Movsisyan A, Tadevosyan V, et al. Text2video-zero: text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 15954–15964
5 Song J, Meng C, Ermon S. Denoising diffusion implicit models. In: Proceedings of International Conference on Learning Representations, 2020