Appendix

A LongV-EVAL Details

We collected 75 videos, each approximately one minute long with a frame rate of 20-30 fps, from *https://mixkit.co/, https://www.pexels.com, and https://pixabay.com.* These videos cover diverse categories including humans, animals, natural landscapes, and human landscapes, with detailed category distributions shown in Fig.A1. To annotate these data with high-quality editing prompts, we first input the video V and prompt P_1 into Video-Llava, where P_1 is *"Please add a caption to the video in great detail."* This generates a detailed textual description C of the video.

Next, we input prompt P_2 into GPT-4, where P_2 has three different forms to generate three distinct editing prompts for the same video. The forms of P_2 are as follows:

- *"I have a video caption: C. Imagine that you have modified the **main object** of the video content (such as color change, similar object replacement, etc.). After editing, add a concise one-sentence caption of the edited video (with emphasis on the edited part, no more than 15 words), not the original video content. The answer should contain only the caption, without any additional content."*
- *"I have a video caption: C. Imagine that you have modified the **background** of the video content (such as background tone replacement, similar background replacement, etc.). After editing, add a concise one-sentence caption of the edited video (with emphasis on the edited part, no more than 15 words), not the original video content. The answer should contain only the caption, without any additional content."*
- *"I have a video caption: C. Imagine that you have applied Van Gogh, Picasso, Da Vinci, Mondrian, watercolors, comics, or **drawings style transfer** to the video. After editing, add a concise one-sentence caption of the edited video (with emphasis on the style, no more than 15 words), not the original video content. The answer should contain only the caption, without any additional content."*

This process eventually generates three editing prompts for each video, as shown in Fig.A3. And a word cloud visualization of all prompts (Fig.A2) demonstrates their diversity and comprehensiveness, presenting substantial challenges for long video editing models.

Figure A1: Category distribution of the LongV-EVAL benchmark dataset. The 75 videos span four primary semantic classes: human activities, animal behaviors, natural landscapes, and human landscapes, ensuring comprehensive coverage of common long video editing scenarios.

Figure A2: Word cloud visualization of editing prompts generated through our multimodal annotation pipeline. The prompts (average length: 13.12 words) reflect the benchmark's challenging diversity for text-driven video editing models. Font sizes correlate with term frequency across all 225 prompts.

B Additional Qualitative Results

As shown in Fig.A7 and Fig.A4, our method can edit over a thousand video frames (even 10k frames) on a single NVIDIA A800 (80GB), while maintaining temporal consistency and achieving high editing quality.

In addition, we also compare with TokenFlow on the official examples used by TokenFlow, as shown in Fig.A5, and find that our method can better preserve details (fingers under the basketball), as well as more realistically preserve the background content (background behind the sculpture man).

C User Study Details

We randomly selected 20 video-text pairs from our dataset for a user study, comparing them with the five baselines mentioned in the main text. For each pair, 50 participants were asked to evaluate and select the best video from the six options based on the following criteria:

- **Video Quality**: The edited video should appear realistic and not easily identifiable as AI-generated. Only the parts specified by the prompt should be edited, while the content

Figure A3: Examples of results for dataset annotation. Each source video is accompanied by three different prompts that focus on three aspects: foreground, background, and style.

Figure A4: Additional Qualitative Results. Our method can support processing videos up to 10k frames in a single inference while maintaining high editing quality and temporal consistency.

not mentioned in the prompt should remain consistent with the source video.

- **Temporal Consistency**: The same object should remain consistent at any point in the long video, and the transitions between frames should be as smooth as in the source video.

D Visualization of Adaptive Attention Slimming

As shown in Fig.A8, the eighth frame serves as the *query* in this attention operation. By employing our proposed method, a portion of the tokens can be automatically discarded to save computational

Figure A5: Additional Qualitative Comparison. We compare with TokenFlow on the official examples used by TokenFlow and find that our method can better preserve details (fingers under the basketball) and more realistically preserve the background content (background behind the sculpture man).

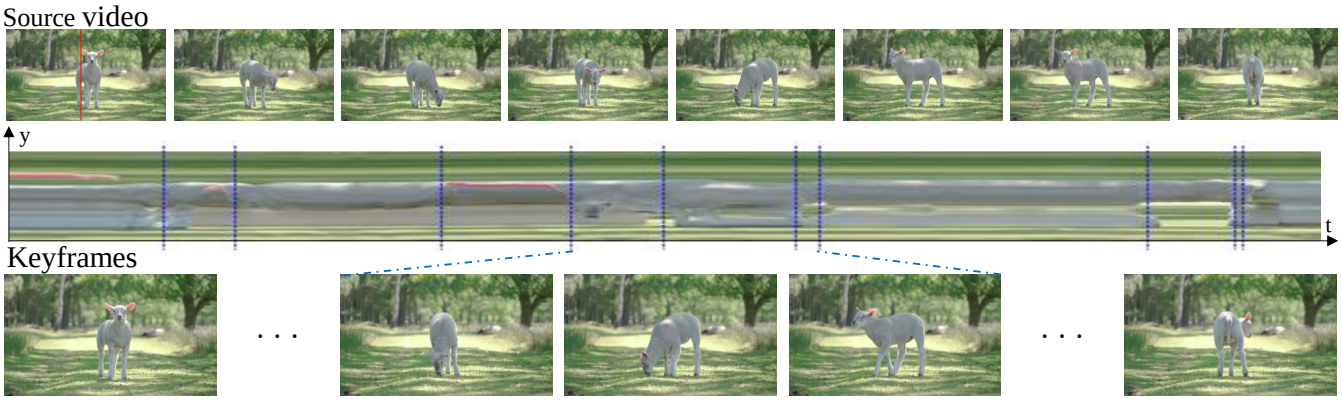

Figure A6: y-t plot. We extracted a vertical column of pixels from the center of each video frame and then sequentially stitched these columns together from left to right to get the y-t plot. The blue lines in the figure indicate the points where the video is segmented.

resources. The content closer to the *query* frame is retained more, while the content further away from the *query* frame is discarded more. This is because, with a larger period, a significant amount of content dissimilar to the *query* appears in the frames, and attending to this content does not contribute to the continuity and consistency of the video. Conversely, the content closer to the query is crucial for maintaining the smoothness of the video. Therefore, using our proposed method not only saves memory but also minimally impacts the quality of video generation.

E Visualization of Keyframe Selection

To visualize the *Adaptive Keyframe Selection*, we extracted a vertical column of pixels from the center of each video frame. We then sequentially stitched these columns together from left to right to create a y-t diagram, as shown in Fig.A6. The blue dashed lines in the figure indicate the points where we segmented the video. It can be observed that each segmentation point corresponds to a significant change in the video content. Moreover, the keyframes obtained from each segment always contain different content. This demonstrates the effectiveness of our method.

F Limitations

Our method adopts the motion information from the source video as a reference to generate non-key frames. Therefore, our approach performs exceptionally well when the image structure remains unchanged. However, it often produces unsatisfactory results when changes in object shapes are required. Additionally, since our method is training-free and directly employs image editing techniques, it primarily addresses the issue of temporal consistency. Consequently, the editing capability of our method may be influenced by the performance of the image editing techniques used.

Figure A7: Additional Qualitative Results. Our method supports a wide variety of text-driven video edits and maintains high editing quality and temporal consistency even for videos exceeding a thousand frames.

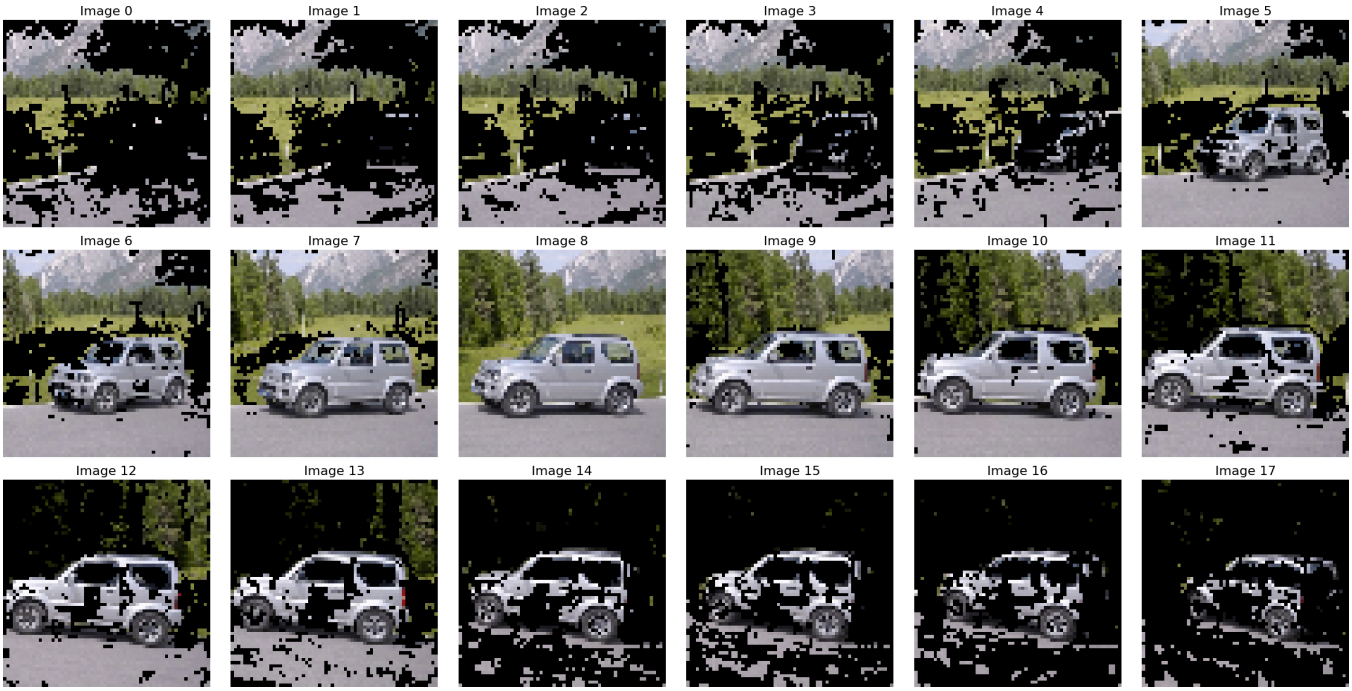

Figure A8: We retain only the tokens corresponding to the regions shown in the figure for K and V during the self-attention computation. In the scenario illustrated here, the eighth frame serves as the query. It can be observed that the content closer to the query frame is automatically retained more, while the content further away from the query frame is discarded more. This automatic selection can save substantial computational resources while maintaining the continuity and consistency of video generation.