EXTENDING THE TEMPORAL HORIZON OF VIDEO EDITING VIA SELF-EDIT ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent video editing models are trained on datasets of mostly short videos, which limits their temporal horizon, *i.e.*, the temporal range over which their editing capability is reliably maintained. Although a long video can be processed at once, we find that edits succeeding on short inputs break down into visual artifacts or inaccurate modifications. Editing it in shorter segments instead avoids this breakdown but yields a different edit in each segment, which poses a trade-off between editing quality and temporal consistency. To resolve this, we propose a simple yet effective method, coined **Horizon Extension via Self-edit Adaptation (HESA)**, which adapts the model to its own edit that stays within the horizon yet spans the full length of the video. Specifically, we compose consecutive frame groups sampled across the video into a short adaptation clip, covering diverse temporal states despite inter-group discontinuities. Since this discontinuity makes the clip harder to edit even within the horizon, we edit it with source-aligned generation, which keeps the editing trajectory aligned with the source clip. We then train a LoRA on the adaptation clip alone, with its edited latent as the flow-matching target. Finally, the adapted model edits the full video with the same source-aligned generation, avoiding a mismatch between the adaptation target and the inference trajectory. Consequently, without any additional training data, HESA extends the temporal horizon of pretrained video editing models, improving both editing quality and temporal consistency on long and dynamic videos.

1 INTRODUCTION

Recent video editing datasets (Zi et al., 2025; Wu et al., 2025; Bai et al., 2026) have enabled models trained directly for video editing (Yang et al., 2026; Cong et al., 2026), moving beyond approaches that extend image editing models. However, these datasets mostly consist of short videos, restricting the temporal range (*i.e.*, temporal horizon) observed during training. Although these models can process a long video at once, we observe that an edit succeeding on a short clip can break down into visual artifacts or inaccurate modifications. Editing the video in shorter segments avoids this breakdown, but independent editing can produce a different edit in each segment, as shown in Fig. 1.

Achieving both editing quality and temporal consistency thus requires a single edit for the whole video, with quality comparable to segment-wise editing. An insight comes from prior approaches that extend image editing models to video (Geyer et al., 2024; Liu et al., 2025), which obtain such an edit by editing only a few keyframes sampled across the video. These keyframes compactly capture the diverse temporal states of the video without requiring the full video to be processed at once.

In this paper, we propose a simple yet effective framework, coined **Horizon Extension via Self-edit Adaptation (HESA)**, which extends the temporal horizon of a pretrained video editing model by adapting it to its own edit spanning the whole video, without any additional training data. As illustrated in Fig. 2, this edit can be obtained using an adaptation clip, *i.e.*, a short clip constructed by uniformly sampling frames across the video while remaining within the model's temporal horizon. Unlike prior approaches that use edited keyframes as temporal references at inference, HESA incorporates the desired edit into the model itself through two components, source-aligned clip editing and self-edit adaptation. The first edits the adaptation clip with source-aligned generation based on FlowAlign (Kim et al., 2026), which keeps the editing trajectory aligned with the source clip to compensate for the temporal discontinuity of the clip. The second trains a LoRA (Hu et al., 2021)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

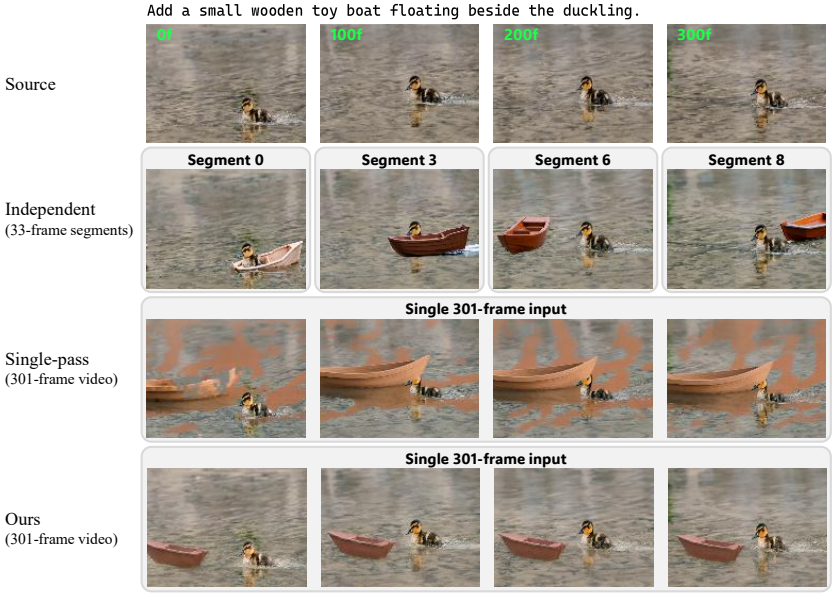

Figure 1: **Limited temporal horizon in video editing.** Independent short-segment editing maintains editing quality but yields inconsistent results across segments, whereas single-pass long-video editing suffers from degraded editing quality. Our method extends the temporal horizon of the pretrained model while maintaining consistent, high-quality edits over the full video.

with the edited adaptation clip latent as the flow-matching target while keeping the pretrained model frozen. As the adaptation clip covers the content appearing throughout the video (*i.e.*, its diverse temporal states), HESA learns how this content should be edited from only a few frames. Consequently, when editing the full video with the same source-aligned generation, the adapted model maintains the editing quality learned from the adaptation clip over a broader temporal range.

Through extensive experiments across varying temporal horizons and scene dynamics, we demonstrate the effectiveness of HESA and its compatibility with two pretrained video editing models (Yang et al., 2026; DecartAI-Team, 2025). Despite its simplicity, HESA improves the overall V2V-Bench (Liu et al., 2026) score by 8.3% over single-pass editing and, at more than $9\times$ the training length, even surpasses the model's short-range baseline score of 0.72, achieving 0.74. We further observe larger gains on dynamic scenes than on non-dynamic ones (9.7% compared with 7.0%), suggesting that self-edit adaptation is particularly effective under substantial scene changes. Our analysis of the adaptation clip shows that effective adaptation requires both broad temporal coverage and local temporal continuity. It should capture diverse states throughout the video while preserving locally consecutive frames, so that its local temporal structure matches that of continuous video inputs at inference. Finally, we show that source-aligned generation and self-edit adaptation play complementary roles. Source-aligned generation helps the edit follow the source video's temporal evolution and structure, but can weaken the intended edit; self-edit adaptation compensates for this loss, improving edit faithfulness beyond the original pretrained model while largely preserving the benefits of source alignment.

Our contributions are summarized as follows:

- We show that a pretrained video editing model can extend its temporal horizon by adapting to its own edit of a short adaptation clip that sparsely covers the full video. Although the model never sees the full video during adaptation, the learned edit transfers to all its frames.
- We demonstrate that HESA consistently improves editing performance across two pretrained video editing models while extending their temporal horizons to more than $9\times$ the training range. It improves the overall V2V-Bench score by 8.3% over single-pass editing and yields larger gains on dynamic videos.
- Our analysis shows that effective adaptation requires both broad temporal coverage and local continuity. We further show that source-aligned generation and self-edit adaptation play complementary roles in preserving source alignment and edit faithfulness.

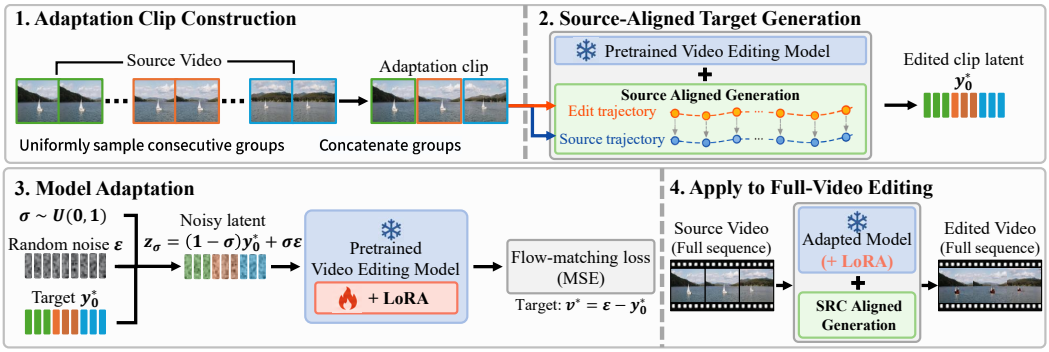

Figure 2: **Overview of Horizon Extension via Self-edit Adaptation.** We adapt the model to a source-aligned edit of an adaptation clip sampled across the video, then apply it to the full video.

2 METHOD

Here, we describe HESA, which extends the temporal horizon of a pretrained video editing model by adapting it to its own edit within the horizon, without any additional training data. As illustrated in Fig. 2, we first construct a short adaptation clip from consecutive frame groups uniformly sampled across the source video. The pretrained model edits this clip with source-aligned generation, and we train a LoRA (Hu et al., 2021) with the edited latent as the flow-matching (Lipman et al., 2022) target while keeping the pretrained model frozen. The adapted model then edits the full video with the same source-aligned generation, without further optimization. This separates obtaining the edited target within the model's temporal horizon from applying the learned edit over the full video.

2.1 SELF-EDIT ADAPTATION

We adapt the pretrained video editing model using an adaptation clip that covers temporally diverse states of the source video. The adaptation procedure consists of adaptation clip construction, source-aligned target generation, and model adaptation.

Adaptation Clip Construction.

We construct the adaptation clip by uniformly sampling consecutive frame groups from the source video and concatenating them into a sequence. This design preserves local temporal continuity while providing compact coverage of temporally diverse parts of the video.

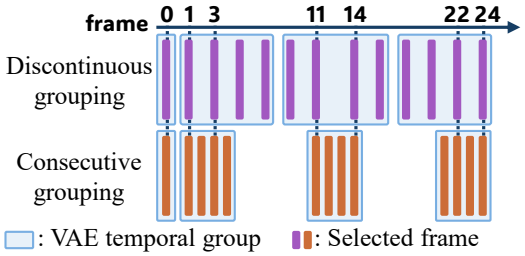

Figure 3: **Adaptation clip design.** Sampling individual frames can introduce temporal discontinuities within VAE groups, whereas consecutive group sampling preserves temporal continuity.

Local temporal continuity is required to preserve the temporal input structure expected by the video VAE. For the video VAE used in our backbone, Wan-VAE (Wan et al., 2025), consecutive input frames are temporally compressed into a shorter sequence of latent representations. As shown in Fig. 3, uniformly sampling individual frames can introduce temporal discontinuities in the video VAE input. Such frame-wise sampling produces a temporal structure different from that of continuous video sequences and the inputs encountered during full-video inference. This mismatch can hinder model adaptation and limit transfer of the learned edit to full-video editing. We further validate the effect of this temporal mismatch through an ablation study in Sec. 3.5.

Source-Aligned Target Generation. We generate the edited latent used for adaptation by editing the adaptation clip obtained in the previous stage. Specifically, we use the final latent representation of the edited clip, denoted as y_0^*, for model adaptation.

Although our clip construction preserves local temporal continuity, the adaptation clip still compresses temporally distant parts of the full video into a short sequence, which can degrade editing

quality. To mitigate this issue, we employ source-aligned generation inspired by FlowAlign (Kim et al., 2026), which keeps the editing trajectory aligned with the source clip and stabilizes editing of the adaptation clip.

At each denoising step, let p_t and q_t denote the edited and source trajectory states, respectively, with corresponding velocity estimates v_p and v_q. We estimate their clean latents as

$$\hat{p}_0 = p_t - \sigma_t v_p, \qquad \hat{q}_0 = q_t - \sigma_t v_q, \tag{1}$$

where σ_t denotes the flow noise level at step t. We then apply the source-directed correction as

$$y_t \leftarrow y_t + \Delta\sigma_t \left(v_p - v_q\right) + \gamma \left(\hat{q}_0 - \hat{p}_0\right), \tag{2}$$

where $\Delta\sigma_t = \sigma_{t+1} - \sigma_t$, y_t denotes the editing latent being updated, and γ controls the source-alignment strength.

This correction encourages the edited trajectory to remain aligned with the source during generation. The resulting final latent y_0^* is then used for model adaptation. The effect of source alignment is further evaluated in Sec. 3.6.

Model Adaptation. We transfer the edit encoded in y_0^* into the pretrained model by optimizing lightweight LoRA parameters with a standard flow-matching objective. At each adaptation step, we sample Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and a noise level σ, and construct the noisy latent as

$$z_\sigma = (1 - \sigma) y_0^* + \sigma\epsilon. \tag{3}$$

The corresponding target velocity is given by

$$v^* = \epsilon - y_0^*. \tag{4}$$

Let $v_{\theta,\phi}$ denote the velocity predicted by the pretrained model with LoRA parameters ϕ, where the pretrained parameters θ remain frozen. We optimize only the LoRA parameters using

$$\mathcal{L}_{\text{adapt}} = \left\| v_{\theta,\phi}(z_\sigma, \sigma, c) - v^* \right\|_2^2, \tag{5}$$

where c denotes the conditioning information used during adaptation.

This objective adapts the LoRA parameters to the edit represented by y_0^*. Because the adaptation clip comprises consecutive frame groups sampled across the video, y_0^* captures editing outcomes associated with diverse temporal states. Adaptation uses only this short clip and its source-aligned edit; the remaining video frames are not processed until full-video inference.

2.2 Full-Video Editing with the Adapted Model

After adaptation, we apply the adapted model directly to the full video without further optimization. Importantly, we use the same source-aligned generation procedure as in adaptation, maintaining a consistent generation trajectory between adaptation and full-video inference. The effect of this consistent source alignment is further evaluated in Appendix B.

Our framework therefore requires optimization only on the short adaptation clip, while full-video editing is performed in a single pass with the adapted model. This makes the overall procedure simple and computationally efficient.

3 Experiments

3.1 Experimental Settings

Implementation details. We use VideoCoF (Yang et al., 2026) as the video editing backbone, built on Wan2.1-T2V-14B (Wan et al., 2025). LoRA (Hu et al., 2021) modules are inserted into the value and output projections of the self-attention layers in blocks 30–39. We use rank 4 with $\alpha = 4$ and a scaling factor of 1.0. For VideoCoF's [source|ground|edit] sequence, LoRA updates are applied only to the edit-token rows, leaving the source and ground-conditioning tokens unchanged. We train the LoRA for 500 steps using AdamW (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-3} and zero weight decay. The LoRA parameters are maintained in FP32, while the backbone is executed

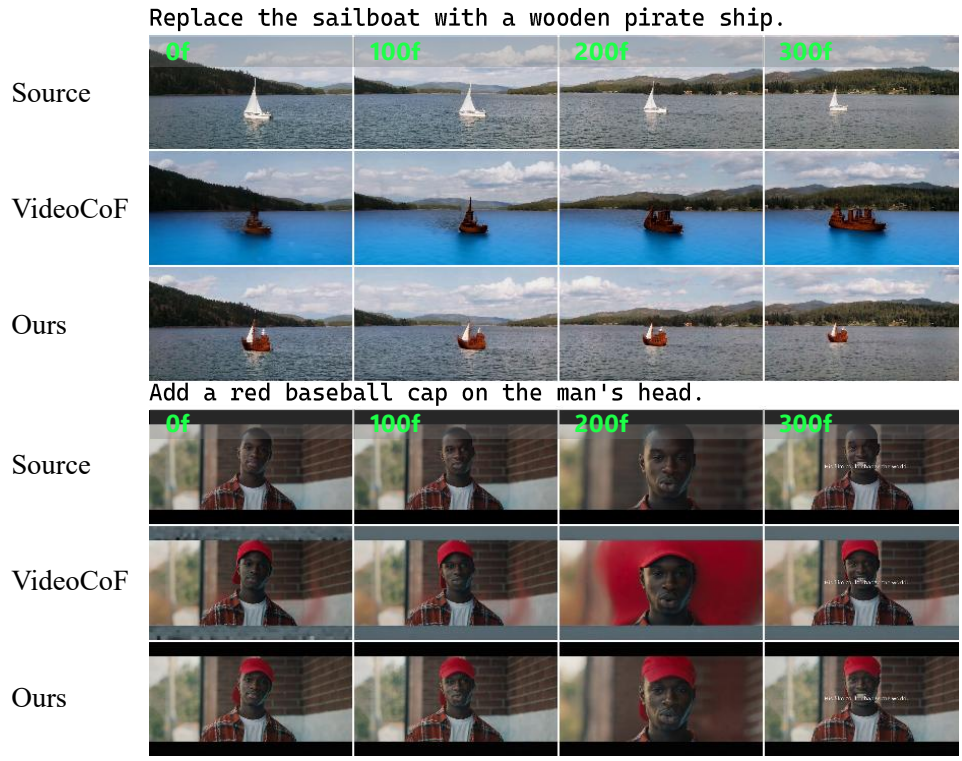

Figure 4: **Qualitative comparison over extended temporal horizons.** We show single-pass editing results on 301-frame videos at frames 0, 100, 200, and 300. VideoCoF introduces unintended changes to surrounding regions as the temporal horizon increases, whereas our method maintains the target edit while better preserving the source content across distant frames.

in BF16. For source-aligned generation, we set the alignment strength to $\gamma = 0.002$. We use a 17-frame adaptation clip consisting of four consecutive 4-frame groups and one additional frame required by the VAE's temporal compression scheme. Adaptation takes approximately 15 minutes per sample on an NVIDIA B200 GPU.

Datasets. VideoCoF is trained on 33-frame clips from Señorita-2M (Zi et al., 2025), and existing Señorita-based evaluations are likewise limited to 33-frame clips. This makes it difficult to assess performance beyond the model's training temporal range. We therefore construct an evaluation set of 24 videos, consisting of 12 videos collected by us and 6 videos randomly selected from each of LAMBDA (Si et al., 2025) and AdsQA (Long et al., 2025). For each video, we use Claude Fable 5.1 (Anthropic, 2026) to generate three editing prompts tailored to the video content, resulting in a total of 72 video-prompt pairs. As advertisement-video datasets, LAMBDA and AdsQA contain frequent shot transitions and substantial scene changes, allowing us to evaluate editing stability under diverse temporal dynamics. All videos are trimmed to 301 frames and resized to a height of 480 pixels while preserving their original aspect ratios. This setting enables evaluation substantially beyond the training temporal range of VideoCoF.

Metrics. We use V2V-Bench (Liu et al., 2026) to evaluate video editing quality across Temporal Alignment, Structural Fidelity, Semantic Alignment, Video Quality, and Transformation Quality. These metrics primarily assess how well the edited video remains aligned with the source while preserving visual quality and realizing the target edit. We additionally use VBench (Huang et al., 2024) to measure the internal consistency of the edited video through Subject Consistency and Background Consistency. Together, V2V-Bench and VBench provide complementary evaluation of source alignment, editing quality, and temporal consistency within the edited video.

3.2 Performance over Extended Temporal Range

We first examine single-pass editing on 301-frame videos, far beyond VideoCoF's 33-frame training range, as shown in Fig. 4. When VideoCoF is directly applied to this extended temporal range, its

Table 1: **Quantitative comparison on long-horizon video editing.** Our method improves temporal, structural, and semantic alignment with the source while maintaining video quality, transformation quality, and self-consistency comparable to VideoCoF.

	Group	Metric	VideoCoF	HESA
V2V-Bench	Temporal Alignment	frame correspondence	0.69	**0.76** (+10.0%)
		temporal consistency	0.55	**0.62** (+13.4%)
	Structural Fidelity	structural preservation	0.76	**0.82** (+ 8.7%)
		layout adherence	0.72	**0.81** (+11.8%)
	Semantic Alignment	content preservation	0.55	**0.79** (+44.5%)
	Video Quality	aesthetic quality	0.48	0.48
		imaging quality	0.59	**0.62** (+ 5.0%)
		motion smoothness	0.97	0.97
		temporal flickering	0.89	0.89
	Transformation Quality	edit faithfulness	0.60	0.60
	Total		0.68	**0.74** (+ 8.3%)
VBench	Self-Consistency	subject consistency	0.85	0.85
		background consistency	0.88	0.88
	Total		0.87	0.87

editing becomes less stable, whereas our method better preserves both the intended edit and surrounding source content. In the first example, VideoCoF substantially alters the surrounding water while replacing the sailboat. In the second example, it introduces a pronounced background change at 200f. In contrast, our method maintains more consistent edits and better preserves the source content across distant frames. These results indicate that HESA effectively extends the temporal range over which the pretrained model maintains stable editing behavior.

We next quantify this improvement on the 301-frame evaluation set using V2V-Bench and VBench, as reported in Tab. 1. V2V-Bench measures alignment with the source through Temporal Alignment, Structural Fidelity, and Semantic Alignment, whereas VBench Self-Consistency evaluates consistency within the edited video itself. Compared with VideoCoF, our method improves Temporal Alignment by 10.0-13.4%, Structural Fidelity by 8.7-11.8%, and Semantic Alignment by 44.5%, resulting in an overall V2V-Bench improvement of 8.3%. Meanwhile, Subject and Background Consistency remain unchanged, while Video Quality and Transformation Quality remain comparable. These results show that our method extends the temporal editing horizon while substantially improving source alignment without sacrificing the internal consistency of the edited video.

3.3 Temporal Horizon Analysis

We analyze how editing performance changes when the temporal input is extended far beyond the range used by the pretrained model. As shown in Tab. 2, extending VideoCoF from 33 to 301 frames reduces Temporal Alignment by 13.9% and Semantic Alignment by 9.8%, indicating that its short-horizon editing behavior does not transfer reliably to substantially longer inputs. In contrast, our method improves Temporal Alignment by 11.5% over 301-frame VideoCoF, recovering 70% of the degradation caused by extending the temporal input from 33 to 301 frames. It also improves Structural Fidelity and Semantic Alignment by 10.2% and 44.5%, respectively, both exceeding the 33-frame VideoCoF performance. Video Quality and Transformation Quality remain comparable to those of 301-frame VideoCoF. At more than 9× the training length, HESA achieves an overall V2V-Bench score of 0.74, surpassing the 33-frame VideoCoF score of 0.72.

Table 2: **Performance comparison across temporal horizons.** VideoCoF shows degraded alignment from 33 to 301 frames, whereas our method largely recovers it while maintaining comparable video and transformation quality.

Metric Group	VideoCoF		HESA
	33 frames	301 frames	301 frames
Temporal Align.	0.72	0.62	0.69 (+11.5%)
Structural Fid.	0.74	0.74	0.81 (+10.2%)
Semantic Align.	0.61	0.55	0.79 (+44.5%)
Video Qual.	0.76	0.74	0.74
Transform. Qual.	0.61	0.60	0.60



Figure 5: **Video editing results on a dynamic scene.** Across large scene changes, baseline method shows inconsistent insertion and unintended background modifications, whereas ours maintains a coherent white cat throughout the video.

3.4 Performance on Dynamic Scenes

We further examine performance under dynamic scene changes, where maintaining stable edits is particularly challenging as the content and composition vary substantially over time. We define dynamic videos as those containing shot cuts or substantial changes in scene content and composition. Videos from LAMBDA and AdsQA are categorized as dynamic.

Table 3: **Performance on dynamic and non-dynamic scenes.** HESA improves performance on both dynamic and non-dynamic scenes, with larger gains on dynamic scenes.

Scene Type	VideoCoF	HESA	Gain
Non-dynamic	0.70	0.75	+7.0%
Dynamic	0.66	0.73	+9.7%

For each scene type, we average Temporal Alignment, Structural Fidelity, and Semantic Alignment. As shown in Tab. 3, our method yields a larger improvement on dynamic videos (9.7%) than on non-dynamic videos (7.0%). This trend is more pronounced in Temporal Alignment. VideoCoF drops from 0.62 on non-dynamic videos to 0.48 on dynamic videos, while our method improves the dynamic score to 0.58, corresponding to a 20.8% improvement compared with 8.1% on non-dynamic videos. This suggests that adaptation is particularly effective under substantial scene changes.

Fig. 5 shows the same behavior qualitatively. VideoCoF produces unstable edits and unintended changes to surrounding regions, whereas ours maintains a more coherent and localized edit despite substantial changes in scene composition.

3.5 Analysis of Temporal Coverage and Continuity in the Adaptation Clip

As discussed in Sec. 2, effective clip adaptation requires both temporal continuity and broad temporal coverage. Discontinuous sampling disrupts the input structure, while a single consecutive clip limits temporal coverage. We therefore compare these alternatives with our distributed consecutive design, which provides both local temporal continuity and broad temporal coverage. For efficiency, all analyses and ablations from this section onward use a 24-pair subset of the evaluation set, obtained by sampling one of the three editing prompts for each of the 24 videos. Unless otherwise noted, all subsequent ablations use this same subset.

Table 4: **Ablation on adaptation clip construction.** Distributed consecutive sampling provides broad temporal coverage and local continuity, yielding the strongest source alignment.

Metric Group	Discontinuous	Consecutive	
	Uniform	Single	Distributed
Temporal Align.	0.69	0.68	**0.71**
Structural Fid.	0.80	0.80	**0.81**
Semantic Align.	0.76	0.77	**0.78**
Video Qual.	0.74	0.74	0.74
Transform. Qual.	0.60	0.60	0.60

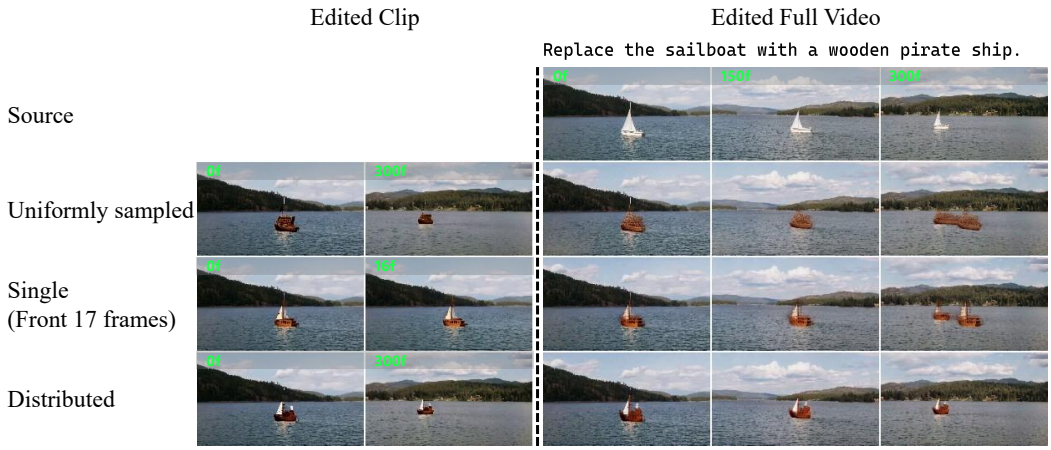

Figure 6: **Effect of adaptation clip construction on full-video editing.** Uniform sampling covers the full temporal range but disrupts local continuity. A single consecutive clip preserves continuity but has limited temporal coverage. Our distributed consecutive design achieves both broad coverage and local continuity, enabling more stable transfer to full-video editing.



Figure 7: **Qualitative effect of source alignment and model adaptation.** VideoCoF modifies background regions, while source alignment better preserves them but can under-edit. Our method balances successful editing and background preservation, although all methods struggle under the large scene change at 300f.

Tab. 4 compares different adaptation-clip constructions. Our design achieves the strongest source alignment with comparable overall quality, confirming the importance of both local temporal continuity and broad temporal coverage for effective full-video adaptation.

Fig. 6 shows the same trend qualitatively. Discontinuous sampling provides broad coverage but lead to less stable transfer, whereas a single consecutive clip lacks temporal diversity. In contrast, our design preserves local input structure while exposing the model to diverse temporal states, resulting in more stable full-video editing.

3.6 Effect of Source Alignment and Model Adaptation

Complementary Effects of Source Alignment and Model Adaptation. As described in Sec. 2, we apply source alignment during adaptation-clip editing to obtain a reliable adaptation target. To disentangle its effect from model adaptation, we compare the original VideoCoF, VideoCoF with source alignment, and our full method. Fig. 7 shows their different behaviors. VideoCoF produces unstable edits across distant frames, whereas source alignment better preserves the source scene but

Table 5: **Effect of source alignment and model adaptation.** Source alignment improves source-related metrics but can weaken edit faithfulness, while adaptation recovers edit strength and remains effective on both VideoCoF and LucyEdit.

Metric Group	VideoCoF			LucyEdit	
	Baseline	+ Source Align.	HESA	Baseline	HESA
Temporal Alignment	0.62	0.73 (+17.8%)	0.71 (+14.2%)	0.72	0.73 (+ 0.7%)
Structural Fidelity	0.73	0.83 (+13.8%)	0.81 (+ 9.9%)	0.81	0.83 (+ 3.6%)
Semantic Alignment	0.48	0.79 (+63.8%)	0.78 (+62.4%)	0.78	0.81 (+ 3.8%)
Video Quality	0.74	0.74	0.74	0.72	0.73 (+ 0.7%)
Transformation Quality	0.60	0.60	0.60	0.60	0.60
Edit Faithfulness (VQAScore)	0.56	0.47 (-16.5%)	**0.60** (+ 5.8%)	0.32	**0.36** (+12.8%)
Self-Consistency	0.86	0.85 (- 1.7%)	0.86	0.86	0.86

tends to yield more conservative edits. Our full method retains source fidelity while producing a more complete and consistent transformation.

As shown in Tab. 5, source alignment improves Temporal, Structural, and Semantic Alignment by 17.8%, 13.8%, and 63.8%, respectively, while slightly reducing Self-Consistency by 1.7%. However, despite the visibly more conservative edits in Fig. 7, Transformation Quality remains unchanged. This suggests that the CLIP-based Transformation Quality metric (Radford et al., 2021) does not fully capture the reduction in edit strength.

We therefore evaluate Edit Faithfulness using a VLM-based generative VQA metric derived from VQAScore (Lin et al., 2024), implemented with Qwen2.5-VL-7B-Instruct (Yang et al., 2024). Unlike the CLIP-based metric, VQAScore directly evaluates whether an edited frame satisfies a declarative description of the target edit. Source alignment reduces Edit Faithfulness by 16.5%, whereas model adaptation improves it by 5.8% over the original VideoCoF. Meanwhile, most source-alignment gains are retained and Self-Consistency is restored to the baseline level. These results show that model adaptation recovers edit faithfulness while preserving the source-fidelity benefits of source alignment.

Applicability Across Pretrained Editors. We further evaluate HESA on LucyEdit (DecartAI-Team, 2025) to assess its applicability across pretrained editors. LucyEdit has a lower baseline VQAScore than VideoCoF (0.32 compared with 0.56), indicating weaker editing, which may partly explain its relatively strong source-alignment scores. Since our adaptation relies on the model's own edited clip, weaker initial edits can limit the adaptation gain. Nevertheless, HESA improves Temporal, Structural, and Semantic Alignment by 0.7%, 3.6%, and 3.8%, respectively, while increasing Edit Faithfulness by 12.8%. These results show that the framework remains effective even with a weaker underlying editor.

4 Conclusion

In this paper, we show that the temporal horizon of a pretrained video editing model, limited by short training videos, can be extended without additional training data by adapting the model to its own edit within the horizon. Our method, HESA, constructs a short adaptation clip from consecutive frame groups sampled across the video and obtains a source-aligned edit that covers diverse temporal states while preserving local temporal continuity. Through self-edit adaptation, this edit is incorporated into the model and transferred to the full video using the same source-aligned generation procedure. Although adaptation uses only the short adaptation clip, HESA consistently improves editing performance across two pretrained video editing models, achieving an 8.3% V2V-Bench improvement on inputs more than $9\times$ the training length and showing larger gains under dynamic scene changes. Our analysis further shows that broad temporal coverage and local continuity are both important for effective adaptation, and that source-aligned generation and self-edit adaptation play complementary roles in preserving source alignment and edit faithfulness. More broadly, we believe that a model's own outputs within its reliable temporal range can provide supervision for extending that range in other generative tasks constrained by the length of their training data, such as long video generation.

AI USE STATEMENT

We used generative AI tools to assist with writing and editing, related-work retrieval and discovery, and routine code editing. We also used Claude to generate candidate editing prompts for constructing our evaluation video-prompt pairs; all generated prompts were manually reviewed by the authors. AI-assisted code was reviewed and tested by the authors, and all experimental results and analyses were independently verified against the underlying experiments and data. The authors take full responsibility for the methodology, analyses, claims, and final content of this work.

REFERENCES

Anthropic. Claude fable 5.1. https://www.anthropic.com/claude/fable, 2026.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37971–37981, 2026.

Xiaoyan Cong, Haotian Yang, Angtian Wang, Yizhi Wang, Yiding Yang, Canyu Zhang, and Chongyang Ma. Viva: Vlm-guided instruction-based video editing with reward optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34364–34374, June 2026.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17702–17711. IEEE, 2025.

DecartAI-Team. Lucy edit: Open-weight text-guided video editing. 2025. URL https://d2drjpuinn46lb.cloudfront.net/Lucy_Edit__High_Fidelity_Text_Guided_Video_Editing.pdf.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17603–17612. IEEE, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6507–6516. IEEE, 2024.

Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing. In *International Conference on Learning Representations*, volume 2026, pp. 121814–121846, 2026.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Dohun Lee, Chun-Hao Paul Huang, Xuelin Chen, Jong Chul Ye, Duygu Ceylan, and Hyeonho Jeong. Memory-v2v: Memory-augmented video-to-video diffusion for consistent multi-turn editing. In *European Conference on Computer Vision*, pp. 75–93. Springer, 2026.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Tao Liu, Leela Krishna, Gouti Pavan Kumar, Vishav Garg, et al. V2v-bench: A comprehensive benchmark for video-to-video generation evaluation. *arXiv preprint arXiv:2606.05665*, 2026.

Zichi Liu, Yinggui Wang, Tao Wei, and Chao Ma. Anchorsync: Global consistency optimization for long video editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4494–4503, 2025.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 01–12. IEEE, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Harini Si, Somesh Singh, Yaman Kumar Singla, Aanisha Bhattacharyya, Veeky Baths, Changyou Chen, Rajiv Ratn Shah, and Balaji Krishnamurthy. Long-term ad memorability: Understanding & generating memorable ads. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5707–5718, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7589–7599. IEEE, 2023.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16692–16701. IEEE, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yue Ma, Yan Huang, Min Xu, and Qiang Wu. Videocof: Unified video editing with temporal reasoner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37940–37949, 2026.

Yisu Zhang, Chenjie Cao, Chaohui Yu, and Jianke Zhu. Lion-lora: Rethinking lora fusion to unify controllable spatial and temporal generation for video diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14569–14579. IEEE, 2025.

Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. In *NeurIPS D&B*, 2025.

A Related Work

Model Adaptation for Video Generation and Editing. Model adaptation has been widely used to specialize pretrained video models to specific videos, appearances, or motions. For video editing, Tune-A-Video (Wu et al., 2023) and Dreamix (Molad et al., 2023) adapt models using the input video to improve video-specific editing or source fidelity. For video generation, parameter-efficient methods such as VideoMage (Huang et al., 2025) and LiON-LoRA (Zhang et al., 2025) learn subjects, appearances, motions, or controls from limited references. More recently, Dalal et al. (Dalal et al., 2025) extend long-term temporal modeling by training additional temporal modules in a pretrained video Transformer. Unlike these approaches, which primarily specialize content or motion or introduce additional modules for long-video generation, our work uses adaptation to preserve the editing capability of a pretrained video editor over a substantially longer temporal horizon.

Long-Video Editing. Early video editing methods extend pretrained image diffusion models to video by enforcing temporal consistency. TokenFlow (Geyer et al., 2024) propagates diffusion features across frames, RAVE (Kara et al., 2024) promotes inter-frame interaction through randomized noise shuffling, and AnyV2V (Ku et al., 2024) propagates information from an edited reference frame through temporal feature injection. These methods mainly transfer image-editing capability to video through frame-level propagation. Recent long-video editing methods instead use sparse references or explicit memory. AnchorSync (Liu et al., 2025) edits sparse anchor frames and synthesizes intermediate frames with global consistency, while Memory-V2V (Lee et al., 2026) retrieves previous editing results from external memory to maintain consistency over long segments. In contrast, our method extends the temporal horizon of the pretrained video editor itself. Rather than propagating references or maintaining memory during inference, we adapt the model using an edited adaptation clip that compactly covers diverse temporal states, then apply the learned adaptation directly to the full video.

B Effect of Adaptation-Inference Consistency

Table 6: **Effect of adaptation-inference consistency.** Using standard generation at inference introduces a mismatch with the source-aligned adaptation target and degrades full-video editing performance.

Metric Group	VideoCoF	SA Target + Standard Gen.	HESA
Temporal Alignment	0.65	0.71 (+ 9.4%)	0.75 (+16.2%)
Structural Fidelity	0.75	0.82 (+ 8.9%)	0.86 (+14.9%)
Semantic Alignment	0.48	0.64 (+32.2%)	0.86 (+77.5%)
Video Quality	0.73	0.73	0.73
Transformation Quality	0.60	0.60	0.60
Edit Faithfulness (VQAScore)	0.51	0.60 (+19.0%)	0.60 (+19.4%)

As discussed in the main text, the generation procedure used at inference should remain consistent with that used to construct the adaptation target, since a mismatch can hinder transfer of the learned edit. For efficiency, this analysis uses a balanced half of the 24-pair ablation subset, selecting two samples from each edit-type $\times$ dynamics cell. During full-video editing, we therefore use the same source-aligned generation procedure employed when constructing the adaptation target. To evaluate the importance of this consistency, we compare our full method with a variant that uses the same LoRA adapted to a source-aligned target but performs full-video inference with standard generation.

As shown in Table 6, using standard generation at inference consistently reduces the gains from adaptation. Compared with VideoCoF, the improvements in Temporal, Structural, and Semantic Alignment decrease from 16.2%, 14.9%, and 77.5% with consistent source-aligned generation to 9.4%, 8.9%, and 32.2%, respectively. The largest reduction occurs in Semantic Alignment, indicating that matching the generation procedure between adaptation-target construction and full-video inference is important for transferring the learned adaptation to the full video.

C Additional Qualitative Results

We provide additional qualitative comparisons on long and dynamic videos in Fig. 8-Fig. 10. Across diverse editing instructions, VideoCoF often exhibits unstable edits or unintended modifications to source content as the temporal range increases. In contrast, HESA better preserves the intended edit and source content across distant frames, including videos with substantial changes in subject appearance, viewpoint, and scene composition.

As shown in Fig. 8, HESA maintains consistent object-level edits across distant frames despite changes in subject pose and scale. Fig. 9 further shows that the adapted edit remains stable under larger changes in appearance and scene content, while the baseline introduces unintended modifications or inconsistent transformations. In Fig. 10, HESA similarly preserves localized attribute edits across dynamic scenes without substantially altering unrelated source regions. These results further support the quantitative findings in the main paper.

Failure case. The effectiveness of self-edit adaptation depends on the quality of the edited adaptation clip, since it serves as the supervision target for adaptation. As shown in Fig. 11, when the pretrained editor fails to produce a reliable edit on the adaptation clip, the resulting supervision is weak and adaptation provides only limited improvement. This highlights that HESA inherits, to some extent, the within-horizon editing capability of the underlying pretrained editor.

Results on LucyEdit. As discussed in the main text, LucyEdit exhibits weaker baseline edit faithfulness than VideoCoF, making adaptation more sensitive to the quality of the edited adaptation clip. As shown in Fig. 12, when LucyEdit successfully realizes the target edit on the adaptation clip, HESA transfers the edit consistently across the full video. However, when the adaptation-clip edit is weak or incomplete, the benefit of adaptation is correspondingly limited.

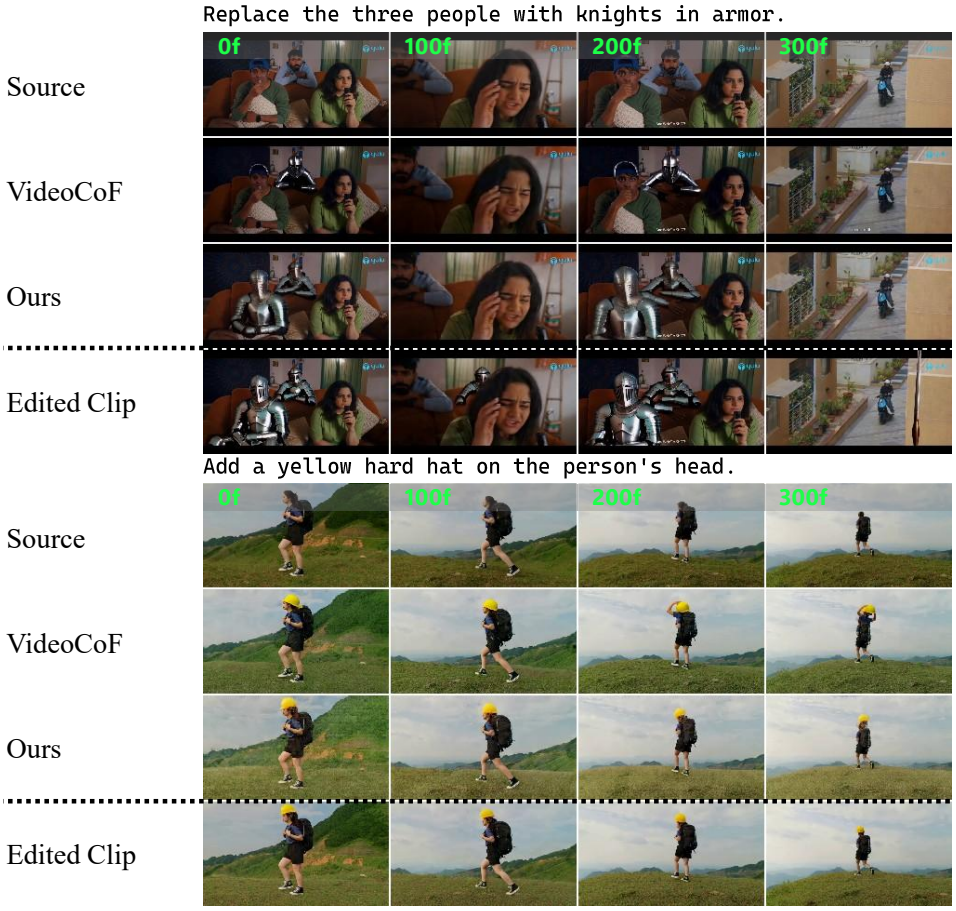

Figure 8: **Additional qualitative comparison on long-horizon object editing.** HESA maintains consistent object-level edits across distant frames while better preserving the source content than VideoCoF. The edited adaptation clip used for self-edit adaptation is shown in the bottom row.

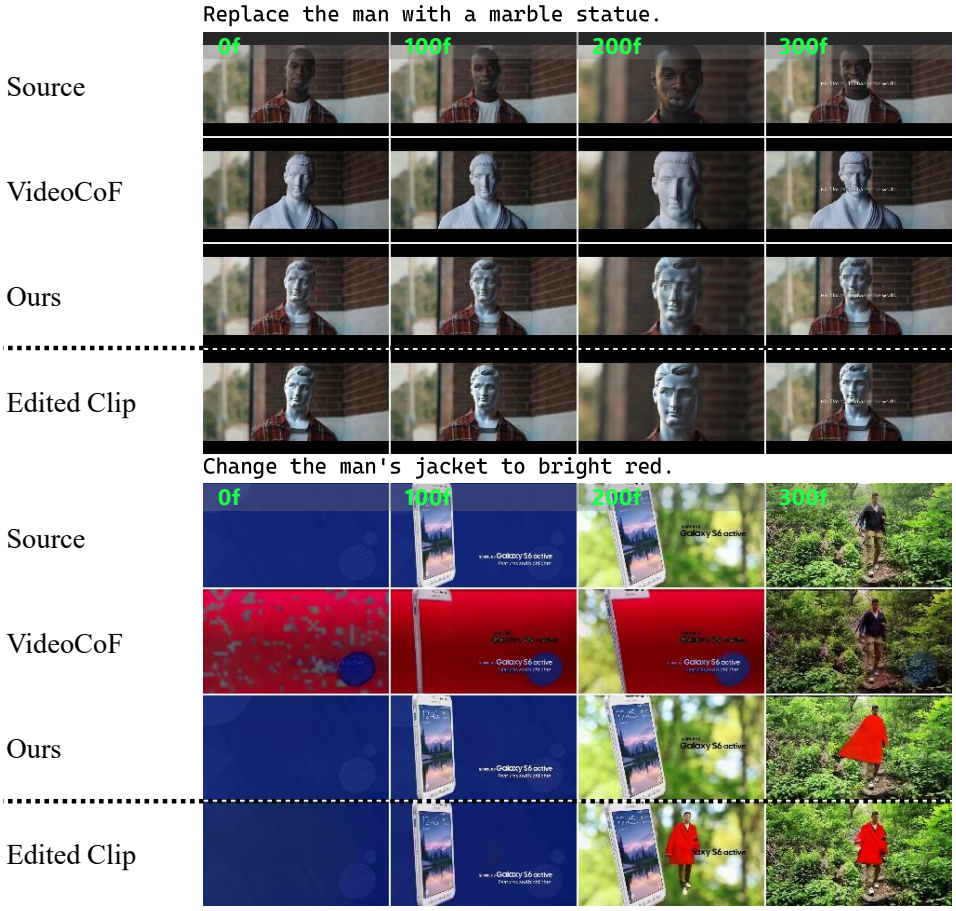

Figure 9: **Additional qualitative comparison under large appearance and scene changes.** HESA preserves the intended edit across distant frames and substantial scene changes, whereas VideoCoF exhibits inconsistent transformations or unintended modifications to the source content.

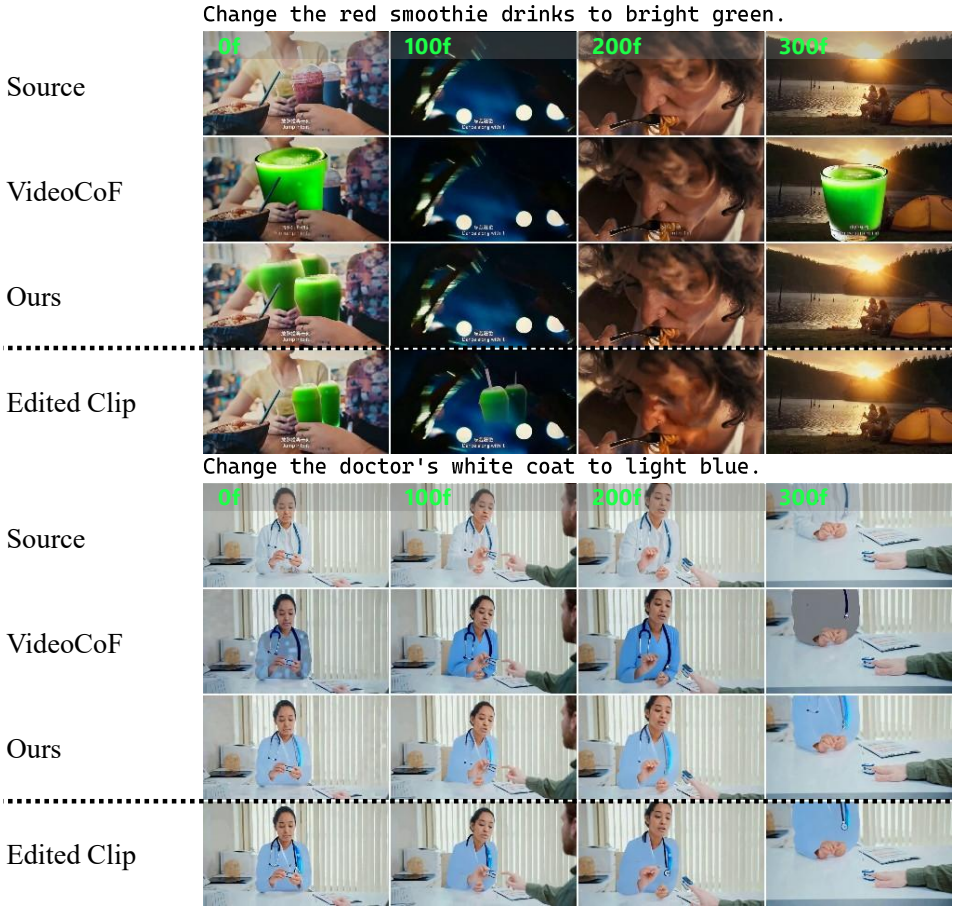

Figure 10: **Additional qualitative comparison on dynamic videos.** Across substantial changes in scene content and composition, HESA maintains localized attribute edits while better preserving unrelated source regions.

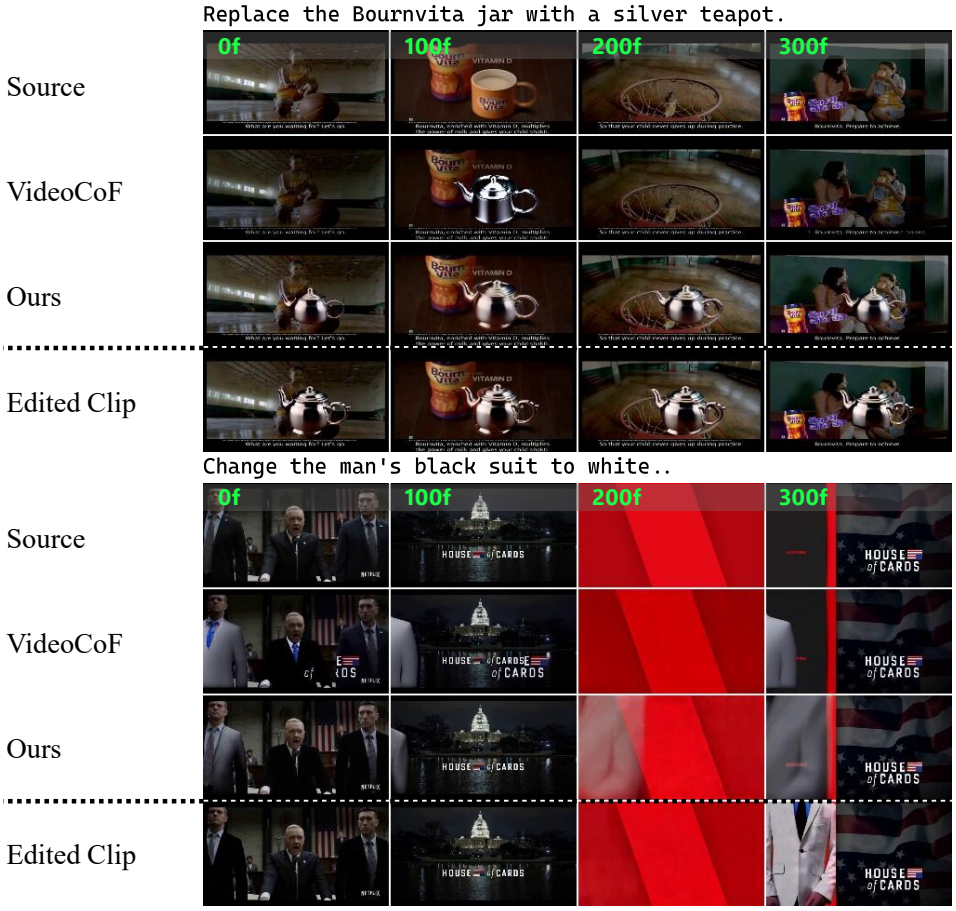

Figure 11: **Failure case caused by an unreliable adaptation target.** When the pretrained editor fails to generate a reliable edit on the adaptation clip, self-edit adaptation receives weak supervision and yields only limited improvement on full-video editing.

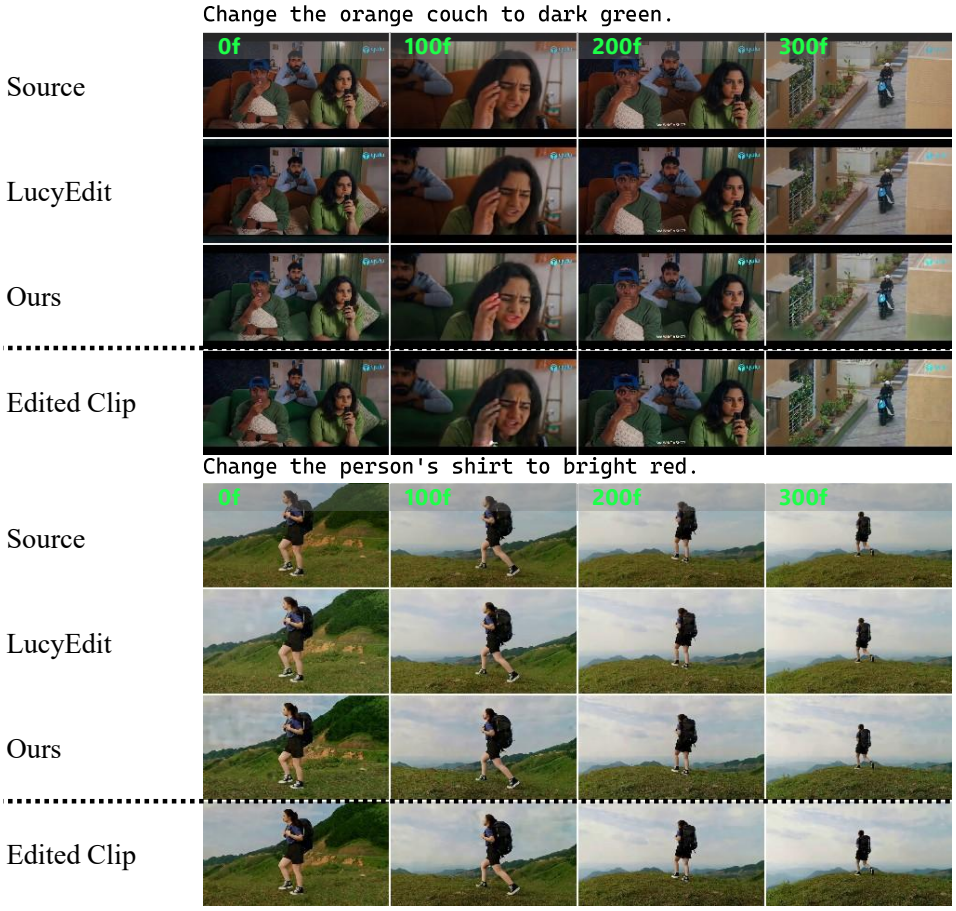

Figure 12: **Qualitative results with LucyEdit.** When LucyEdit produces a successful edit on the adaptation clip, HESA transfers it consistently across the full video. When the adaptation-clip edit is weak, however, the gain from adaptation is limited.