MOVIEDIT: LONG-FORM MULTI-SHOT VIDEO EDITING MADE EASY

Anonymous authors
Paper under double-blind review

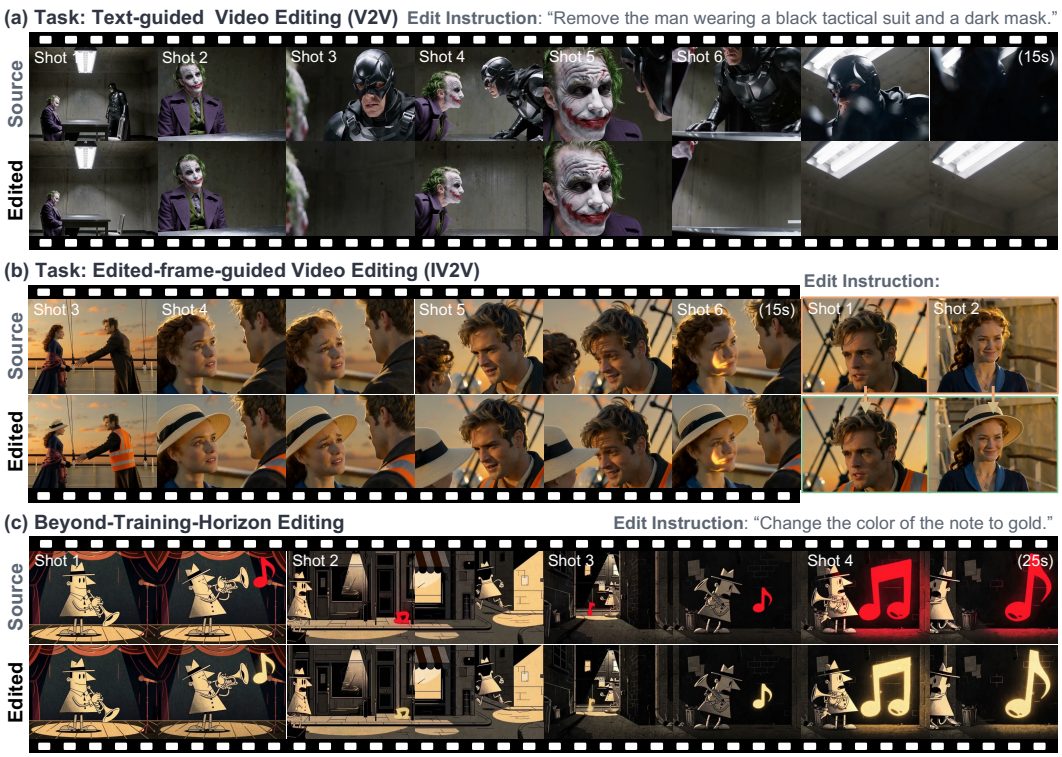

Figure 1: **Long-form multi-shot video editing with MoviEdit.** MoviEdit supports (a) text-guided and (b) edited-frame-guided multi-shot video editing, enabling consistent edits across shots, and (c) naturally generalizes to videos longer than those seen during training.

ABSTRACT

Long-form multi-shot video editing requires an intended edit to remain coherent across abrupt changes in scene, viewpoint, and subject appearance, while processing substantially longer visual context than conventional single-shot editors. We introduce **MoviEdit**, an end-to-end framework for scalable long-form multi-shot video editing that jointly addresses the challenges of data, modeling, and evaluation. To provide cross-shot supervision data at scale, we develop automatic data-generation pipelines spanning synchronized multi-view footage, real-world multi-shot videos, and generated multi-shot content, yielding 87K paired editing videos without manual annotation. Architecturally, MoviEdit employs a cascaded bidirectional–autoregressive design: a lightweight bidirectional stage first establishes a globally coherent editing draft over a spatiotemporally compressed representation, which is then autoregressively propagated and refined over the full video. We further introduce **MoviEdit-Bench**, a dedicated benchmark for systematic evaluation of long-form multi-shot video editing. Extensive experiments show that MoviEdit substantially outperforms existing open-source models and achieves performance comparable to strong proprietary models across text-guided and image-guided editing. Project page: multi-shot-edit-anonymous.github.io

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.

1 INTRODUCTION

Visual narratives are inherently compositional: their meaning emerges not only within individual shots, but also from the relationships among shots over time. Recent advances in generative video modeling have begun to capture this structure, moving beyond short, isolated clips toward long-form, multi-shot video generation (Guo et al., 2025; Zhang et al., 2025; Cai et al., 2026; Meng et al., 2026). Yet existing video editing methods (DecartAI Team, 2025; Jiang et al., 2025; Lee et al., 2026; Liu et al., 2025; Ju et al., 2026) remain predominantly designed for short, single-shot videos, typically spanning only a few seconds. Extending editing to long-form, multi-shot content is not merely a matter of processing more frames: an intended edit must persist coherently across cuts even as the scene, viewpoint, motion, appearance, and surrounding context change.

The difficulty is structural rather than simply one of scale. Editing each shot independently discards the cross-shot dependencies needed to preserve a consistent edit, whereas jointly processing the full source and target sequences quickly becomes computationally prohibitive. Yet this computational dilemma is only one part of the problem. Fundamentally, long-form multi-shot editing exposes three coupled bottlenecks in existing video editing systems: *data*, *modeling*, and *evaluation*. At the *data* level, learning to edit a subject consistently across shots requires source and target video pairs with explicit cross-shot subject correspondence. Such data does not exist in the real world and, to our knowledge, no existing pipeline can synthesize them at the scale and quality required for training (Zi et al., 2025; Bai et al., 2026; Wu et al., 2025; Lin et al., 2026; Long et al., 2026). On the *modeling* side, long-form editing is fundamentally a memory problem. Computationally, editing typically requires processing a source-video representation alongside the target video being synthesized; under common token-concatenation designs, the quadratic cost of global self-attention rapidly becomes prohibitive as duration increases. Third, on the *evaluation* side, there is no dedicated evaluation protocol for this setting. Existing video editing benchmarks (Li et al., 2025; Bian et al., 2025; Ju et al., 2026) focus primarily on short, single-shot clips and therefore do not assess the capabilities that distinguish multi-shot editing, such as whether an edit continues to localize the correct subject after a cut and whether its appearance remains consistent when that subject reappears. Once supervised data and evaluation are in place, the modeling problem becomes tractable with the goal of maintaining temporal and appearance consistency over long horizons.

To this end, we present **MoviEdit**, an end-to-end framework for training and evaluating long-form multi-shot video editing models. MoviEdit addresses the three bottlenecks through a matched set of data, modeling, and evaluation components. First, we develop a scalable data generation pipeline that combines multi-view-aware image editing, single-shot video editing, and VLM-based quality filtering to construct high-quality paired training examples with explicit subject correspondence across shots. Second, we present a cascaded video editing architecture that combines bidirectional global context with causal propagation. A lightweight bidirectional editing stage first constructs a globally coherent edit draft, after which a causal stage autoregressively propagates and refines the edit over the full sequence. This decomposition avoids the prohibitive memory cost of jointly processing the complete source and target videos while preserving temporal and appearance consistency across shots. Lastly, we introduce **MoviEdit-Bench**, a dedicated benchmark that spans multiple difficulty levels and evaluates both local, subject-specific edits and global, scene-level transformations. It provides automated metrics for cross-shot target localization and edit consistency, together with a curated collection of copyright-free multi-shot videos. Together, these components provide the supervision, architecture, and evaluation infrastructure needed to move video editing beyond isolated clips toward coherent, long-form visual narratives. Extensive experiments further demonstrate that a MoviEdit model trained using our framework substantially outperforms existing open-source solutions, and achieves performance on par with strong proprietary systems (Google DeepMind, 2026; Alibaba Cloud, 2026; Runway, 2026) across both text-guided video editing and image-guided video editing.

Our contributions are summarized as follows:

- We introduce a scalable paired-data generation pipeline for supervised long-form multi-shot video editing.
- We propose an efficient long-form video editing architecture based on a cascaded bidirectional–autoregressive diffusion framework.
- We present MoviEdit-Bench, a dedicated benchmark and evaluation suite for long-form multi-shot video editing.

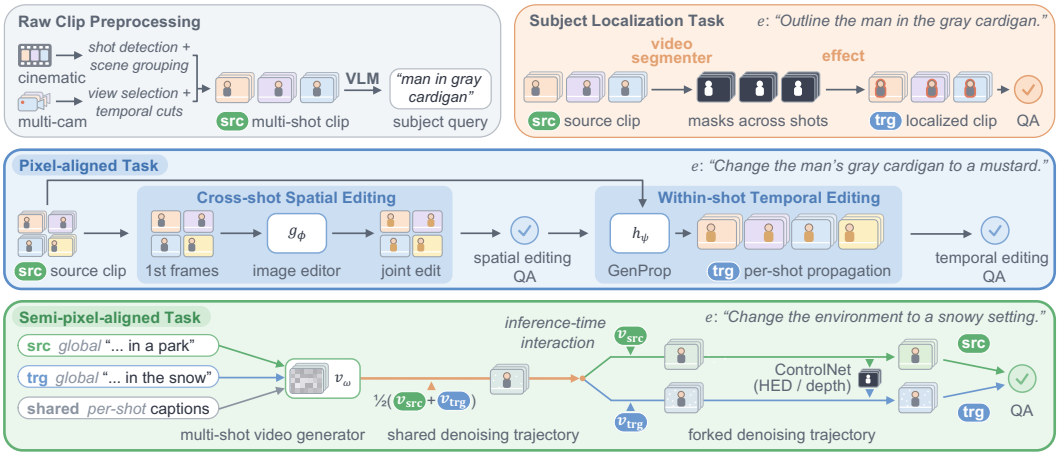

Figure 2: **Scalable paired-data generation for multi-shot video editing.** Three annotation-free pipelines transform diverse video sources into large-scale paired data for consistent multi-shot editing. Pixel-aligned editing uses a spatial–temporal editing decomposition to ensure cross-shot consistency, while semi-pixel-aligned editing uses coupled multi-shot video generation with inference-time interaction to produce structurally consistent source–target pairs.

2 MoviEdit: Long-Form Multi-Shot Video Editing

MoviEdit provides an end-to-end framework that develops the key ingredients for building long-form multi-shot video editors. We first describe our scalable paired-data generation pipelines for multi-shot video editing tasks (Fig. 2), followed by the cascaded architecture for a long-form video editing model (Fig. 4). We then introduce MoviEdit-Bench for systematic evaluation of cross-shot editing capabilities. Finally, we describe several techniques for improving inference efficiency.

2.1 Scalable Paired Data Generation for Multi-Shot Video Editing

We develop three scalable pipelines for constructing paired supervision without human annotation (Fig. 2). We leverage in-house synchronized multi-camera footage (refer to Fig. 9), real-world cinematic clips, or pretrained multi-shot text-to-video diffusion models. These sources offer complementary strengths in cross-view correspondence, cinematic realism, and controllable synthesis, respectively, supporting three supervision regimes: *pixel-aligned editing*, *subject localization*, and *semi-pixel-aligned editing*. Detailed training-data statistics are provided in Appx. B, with data-recipe ablations in Appx. D.2.

For pixel-aligned editing and subject localization pipelines, we obtain single-scene, multi-shot clips either from real-world cinematic clips using shot detection (Wang et al., 2026) and scene grouping (Mun et al., 2022), or from synchronized multi-camera footage through camera-view selection and temporal cuts. A vision-language model (VLM) (Yang et al., 2025; Team et al., 2026) proposes candidate subjects as short text queries, merging observations of the same entity across shots into one query. The multi-shot clips and subject queries are used to construct supervision for both tasks.

Pixel-aligned Editing. The pixel-aligned editing task is our primary setting, where the edited video preserves pixel-wise spatial correspondence with the source while modifying only the instructed visual content. For object-level edits, including *add*, *remove*, *swap*, and *appearance change*, we use the subject query to select the target instance to edit in the first frame of each shot. For global edits, including *environment change* and *stylization*, we apply the edit to the full first frame of each shot.

Ensuring cross-shot consistency is a critical challenge in this setting. We address it by decomposing multi-shot editing into two complementary subproblems: *cross-shot spatial editing*, which resolves a consistent edit across representative frames from all shots, and *within-shot temporal editing*, which propagates each resolved edit through its corresponding shot. For the first stage, we arrange the first frames $\boldsymbol{y}_s^1$ from all S source video shots $\{\boldsymbol{y}_s\}_{s=1}^{S}$ into a single image grid $[\boldsymbol{y}_1^1; \ldots; \boldsymbol{y}_S^1]$ and jointly edit them in a single inference pass using off-the-shelf text-guided image editing models g_ϕ. Each edited first frame is then propagated independently within its corresponding shot using a pretrained

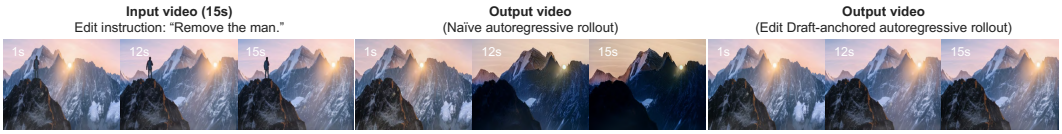

Figure 3: **Long-horizon appearance-drift stress test.** We compare autoregressive editing with and without draft conditioning in a small-scale train setting. Naive autoregressive editing accumulates appearance drift and saturation, while draft-conditioned rollout maintains a stable edited appearance.

first-frame-conditioned single-shot video editing model h_ψ:

$$\tilde{\boldsymbol{y}}^1_{1:S} = [\tilde{\boldsymbol{y}}^1_1; \ldots; \tilde{\boldsymbol{y}}^1_S] = g_\phi\big([\boldsymbol{y}^1_1; \ldots; \boldsymbol{y}^1_S], e\big)\,, \quad \boldsymbol{x}_s = h_\psi\big(\tilde{\boldsymbol{y}}^1_s, \boldsymbol{y}_s, e\big)\,, \quad s = 1, \ldots, S, \tag{1}$$

where e denotes the edit instruction, and $\boldsymbol{x}_s$ is the target edited video shot. This spatial–temporal decomposition avoids bootstrapping from an existing multi-shot video editor. We apply quality filtering at both stages: first-frame QA checks edit success and cross-shot consistency, while propagation QA checks within-shot consistency and overall video consistency after recomposition, rejecting $17\%/38\%$ of multi-camera and $32\%/47\%$ of cinematic pairs at the two stages, respectively.

Subject Localization. This task, also supported by recent video editing models (Ju et al., 2026; Liu et al., 2025), requires identifying a queried subject across multiple shots and applying localized visual transformations, including effects such as *highlighting*, *outlining*, and *color filling*. Given a multi-shot source video and subject query, video segmentation models (Cheng et al., 2024; Sari & Favaro, 2025; Carion et al., 2026) track the subject across all shots where it appears and produce dense per-frame masks. Applying the specified effect to these masks yields paired source–target videos for supervising cross-shot localization. With VLMs, we further filter the generated pairs to ensure consistent subject localization and correct effect application, rejecting 24% of pairs.

Semi-pixel-aligned Editing. This pipeline generates paired source–target videos that preserve spatial structure, temporal composition, and narrative correspondence without requiring strict pixel-level alignment. By relaxing exact correspondence, it supports a broader range of edits while increasing data diversity through coupled generation with a pretrained text-driven multi-shot video generator (e.g., Meng et al., 2026; Zhang et al., 2025). We apply this pipeline to tasks including *environment change*, *medium change*, and *character appearance change*.

For each pair, we couple source and target generation by sharing the same per-shot captions and initial noise while using distinct global captions. An LLM constructs the source and target global captions by modifying only the attribute associated with the intended edit while preserving all other content. To further couple them, we introduce task-specific inference-time interaction mechanisms. For *character appearance change*, the source and target instead share a common denoising trajectory until a fork timestep t_{fork}. For $t > t_{\text{fork}}$, the shared latent is updated using the averaged source- and target-conditioned velocity predictions,

$$\tilde{\boldsymbol{v}}(\boldsymbol{z}_t, t) = \left(\boldsymbol{v}_\omega(\boldsymbol{z}_t, e_{\text{src}}, t) + \boldsymbol{v}_\omega(\boldsymbol{z}_t, e_{\text{tgt}}, t)\right)/2. \tag{2}$$

For $t \le t_{\text{fork}}$, the two generations branch from the shared latent state and are denoised independently under their respective conditions. For environment change and medium change editing, the target generation is additionally conditioned via ControlNet (Zhang et al., 2023) on either an HED edge or a depth map extracted from the source video. These guarantee necessary structural correspondence while allowing the target appearance to diverge.

We further apply VLM-based quality control to generated pairs, filtering them based on prompt alignment, video quality, cross-shot consistency, and source–target distinctiveness, rejecting 66% of pairs. For ControlNet-based pairs, we swap source and target when forming training pairs, using the structurally conditioned generation as input and the original generation as the reconstruction target, since the latter receives higher visual-quality scores.

2.2 Cascaded Architecture for Long-Form Video Editing

A common design for long-form video modeling is chunk-wise autoregressive diffusion (Yin et al., 2025; Liu et al., 2026), where a video is divided into C fixed-size chunks $\boldsymbol{x}^{1:C}$ and modeled with a fixed-size sliding context window over recent chunks. In causal diffusion forcing (CDF) (Chen et al., 2024), each video chunk $\boldsymbol{x}^c$ is independently corrupted at a sampled diffusion timestep t_c, yielding $\boldsymbol{x}^c_{t_c} = \alpha_{t_c}\boldsymbol{x}^c + \sigma_{t_c}\boldsymbol{\epsilon}^c$, and the denoiser is trained with

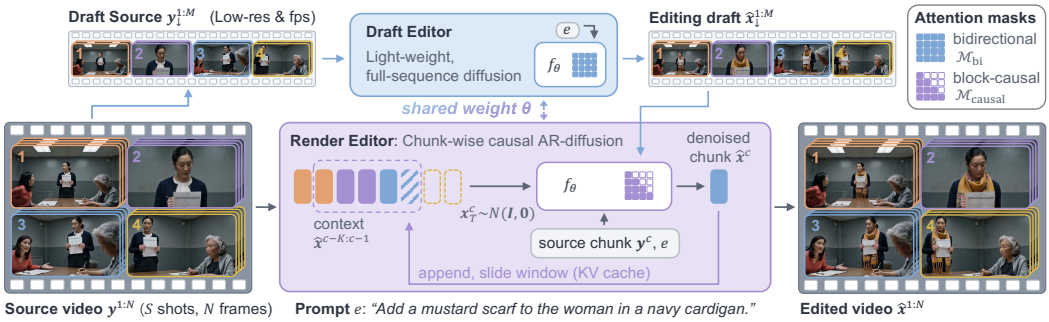

Figure 4: **Cascaded modeling for a long-form video editor.** MoviEdit uses a shared denoiser f_θ to first produce a globally consistent, lightweight edit on a subsampled video sequence (Draft Editor), and then propagates the edit chunk by chunk to the full-resolution video while mitigating long-horizon appearance drift (Render Editor).

$$\mathcal{L}_{\text{CDF}}(\theta) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{t},\boldsymbol{\epsilon}}\left[\sum_{c=1}^{C}\left\|\left[f_\theta\left(\boldsymbol{x}_{\boldsymbol{t}}^{1:C};\mathcal{M}_{\text{causal}}\right)\right]^c - \boldsymbol{v}_{t_c}^c\right\|_2^2\right], \qquad t_c \overset{\text{i.i.d.}}{\sim} p(t), \tag{3}$$

where $\boldsymbol{t} = (t_1, \ldots, t_C)$, $\boldsymbol{x}_{\boldsymbol{t}}^{1:C} = \{\boldsymbol{x}_{t_c}^c\}_{c=1}^C$ and $\boldsymbol{v}_{t_c}^c$ denotes the diffusion target (*e.g.*, velocity) for chunk c at timestep t_c. The causal self-attention mask $\mathcal{M}_{\text{causal}}$ ensures that a standard DiT-based denoiser (Peebles & Xie, 2023) f_θ predicts each chunk c using only itself and the preceding chunks within the retained context window. During inference, the same denoiser f_θ is used within a chunk-wise causal autoregressive diffusion sampler,

$$\hat{\boldsymbol{x}}^c = \mathcal{S}_{f_\theta}\left(\left[\hat{\boldsymbol{x}}^{c-K:c-1};\boldsymbol{x}_T^c\right];\mathcal{M}_{\text{causal}}\right), \qquad \boldsymbol{x}_T^c \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{4}$$

where $\mathcal{S}_{f_\theta}$ denotes a standard diffusion sampler using f_θ, $[\cdot\,;\,\cdot]$ denotes temporal token concatenation, and $\hat{\boldsymbol{x}}^{c-K:c-1}$ denotes the preceding K generated chunks retained as autoregressive context.

However, the limited global context of chunk-wise autoregressive generation leads to cumulative appearance drift and oversaturation over long editing horizons (Fig. 3). To mitigate this issue, we augment the efficient CDF backbone with a lightweight bidirectional diffusion stage that provides global guidance before full-resolution autoregressive editing. The *Draft Editor* first performs full-sequence bidirectional diffusion over a spatially and temporally subsampled video representation t, producing a globally coherent edit draft. This draft serves two purposes: it provides stable appearance anchors that suppress long-horizon drift during autoregressive generation, and, by jointly resolving the edit over the entire sequence, establishes cross-shot consistency even between temporally distant shots. The *Render Editor* then performs full-resolution chunk-wise causal diffusion conditioned on the draft, preserving this global editing structure while efficiently synthesizing local detail from the full-resolution input video chunks (Fig. 4). The *Draft Editor* and *Render Editor* share model weights, differing only in their attention patterns and spatiotemporal resolutions.

Let $\boldsymbol{y}^{1:N}$ denote the N-frame source video and e the edit instruction. We apply a fixed spatiotemporal subsampling operator $\downarrow$ to obtain a shorter, lower-resolution video sequence $\boldsymbol{y}_{\downarrow}^{1:M} = \downarrow(\boldsymbol{y}^{1:N})$, where $M \ll N$. Empirically, we find $M = N/2$ to be effective.

Draft Editing. The Draft Editor applies f_θ bidirectionally to the full subsampled sequence:

$$\hat{\boldsymbol{x}}_{\downarrow}^{1:M} = \mathcal{S}_{f_\theta}\left(\boldsymbol{x}_T^{1:M};\boldsymbol{y}_{\downarrow}^{1:M},e;\mathcal{M}_{\text{bi}}\right), \qquad \boldsymbol{x}_T^{1:M} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{5}$$

where $\mathcal{M}_{\text{bi}} = \mathbf{0}$ removes causal masking, allowing full bidirectional attention across the subsampled sequence. The resulting draft jointly resolves the edit over the entire video, providing a global appearance anchor for subsequent rendering.

Render Editing. Conditioned on this draft edit, the Render Editor applies f_θ causally over fixed-size chunks $c = 1, \ldots, C$ of the full-resolution target video,

$$\hat{\boldsymbol{x}}^c = \mathcal{S}_{f_\theta}\left([\hat{\boldsymbol{x}}^{c-K:c-1};\boldsymbol{x}_T^c];\hat{\boldsymbol{x}}_{\downarrow}^{1:M},\boldsymbol{y}^c,e;\mathcal{M}_{\text{causal}}\right), \qquad \boldsymbol{x}_T^c \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{6}$$

where $\hat{\boldsymbol{x}}^{c-K:c-1}$ denotes the preceding K generated chunks retained as autoregressive context. Each chunk conditions on the recently generated context for temporal continuity, the global draft for long-range appearance and cross-shot consistency, and the corresponding source chunk $\boldsymbol{y}^c$ to preserve content outside the intended edit.

During training, Draft Editor uses a conventional flow-matching objective (Lipman et al., 2022) with full bidirectional attention, while Render Editor is optimized with the CDF objective in Eq. 3.

2.3 Inference Optimization

Our architecture supports complementary techniques for reducing inference latency, including few-step diffusion distillation and KV caching, where efficiency gains are reported in Tab. 4.

Unified Distillation with Asymmetric Step Schedules. After teacher training, we follow velocity moment matching (VMM) (Zhou et al., 2026; Salimans et al., 2024) to jointly distill the bidirectional Draft Editor and causal Render Editor under the shared-weight architecture. Following VMM, timestep pairs are sampled according to a predefined student inference schedule; we assign a separate schedule to each editing stage within the same distillation run. Specifically, the Draft Editor is distilled to 12 denoising steps, while the more computationally expensive Render Editor is distilled to 4 steps. We empirically find that the Render Editor is more robust to aggressive step reduction. This asymmetric allocation preserves sufficient capacity for establishing global edit consistency while aggressively reducing the dominant rendering cost.

KV Caching. During Render Editor inference, the autoregressive formulation and causal attention masking naturally support KV caching. We cache key-value states for all reusable tokens, including text, edit draft, source video, and previously generated chunk tokens, eliminating recomputation.

2.4 MoviEdit-Bench: Evaluating Long-Form Multi-Shot Video Editing

Existing benchmarks for instruction-guided video editing (Bian et al., 2025; Li et al., 2025; Ju et al., 2026) primarily focus on short, single-shot clips and therefore do not explicitly evaluate cross-shot subject localization and editing consistency over long multi-shot sequences. To address this gap, we introduce **MoviEdit-Bench**, a benchmark for long-form cinematic multi-shot video editing under practical production settings, intended for public release. Benchmark construction and evaluation details are provided in Appx. C.

MoviEdit-Bench contains 64 generated multi-shot videos and 128 editing scenarios spanning cinematic and drama-style content. The videos are organized into eight categories with varying cross-shot subject-binding difficulty, ranging from a single recurring subject to crowded multi-subject and mixed human–animal scenes, together with medium-variation and scene-only cases. The benchmark supports both text-guided video editing (V2V) and arbitrary-frame edited-frame-guided propagation (IV2V), with first-frame and non-first-frame reference scenarios. It covers both local edits (e.g., *add*, *remove*, and *appearance change*), and global edits (e.g., *environment change* and *stylization*).

To capture the diverse failure modes of long-form multi-shot editing, our evaluation spans seven complementary dimensions: *non-subject preservation*, *edit localization*, *subject preservation*, *cross-shot edit consistency*, *edit-prompt alignment*, *video quality*, and, for IV2V only, *edited-frame fidelity*. Each dimension is evaluated using complementary signals from pixel-level metrics, embedding-based similarity, and a vision-language model (VLM) judge (Team et al., 2026; Yang et al., 2025), following Ju et al. (2026); Li et al. (2025); Bian et al. (2025). To summarize overall editing success, we further aggregate frame-level VLM judgments across four essential criteria: edit-prompt alignment, edit localization, non-target preservation, and source preservation. Specifically, for a video with T frames, we define

$$\texttt{overall_edit_success} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[a^t_{\text{align}} \wedge a^t_{\text{local}} \wedge a^t_{\text{non-target}} \wedge a^t_{\text{src}}\right], \tag{7}$$

where each a^t denotes the VLM judge's binary verdict for the corresponding criterion at frame t. A frame is counted as successfully edited only when all four criteria are simultaneously satisfied.

3 Experiments

3.1 Experimental Setup

Implementation Details. Using the three data-generation pipelines described above, we construct 87K paired multi-shot editing videos (540p, 15s, 24 fps) from 61,251 raw clips across three data sources (Tab. 3). In addition to this multi-shot dataset, we incorporate a proprietary collection of short, single-shot video editing pairs (540p, 5s, 24fps) during training. Our model uses a DiT-based



Figure 5: Qualitative comparison on MoviEdit-Bench with major baselines.

mixture-of-experts (MoE) backbone with a frozen multimodal text encoder and video VAE. The input sequence concatenates text, source-video, noisy target-video, and task-specific conditioning tokens, including edited-frame and draft-edit conditions. We initialize from a pretrained short-form, single-shot video editing checkpoint *with no prior multi-shot generation or editing capability*. We fully fine-tune all DiT parameters using Fully Sharded Data Parallel (FSDP) across 64 NVIDIA H200 GPUs. Specifically, training proceeds in two stages. We first train the video-to-video (V2V) model for 17K steps. We then specialize the model for IV2V with an additional 8K training steps under the IV2V task formulation. For this stage, edited reference frames are sampled from the ground-truth target video and restricted to shots in which the edited subject is present.

Evaluations. We evaluate MoviEdit and competing methods on MoviEdit-Bench under both text-guided (V2V) and edited-frame-guided (IV2V) settings. V2V models receive only the source video and edit instruction, while IV2V models additionally receive an edited reference frame; neither setting provides object masks. MoviEdit-Bench includes both first-frame and arbitrary-frame IV2V scenarios. For fair comparison, we evaluate all IV2V-capable models on the first-frame subset, and methods supporting arbitrary-frame conditioning on the full benchmark. As no existing, fully

Table 1: Quantitative evaluation on the long-form multi-shot video editing benchmark, MoviEdit-Bench, for text-guided video editing (V2V). (X-shot: cross-shot)

	Model	Non-subject Preservation	Edit Localization		Subject Preservation			X-shot Edit Consistency	Edit Prompt Alignment			Video Quality				Overall Edit Success
		LPIPS↓	Precision↑	Recall↑	MFS↑	LPIPS↓	VLM↑	VLM↑	CLIP↑	ViCLIP↑	VLM↑	DINO↑	NIQE↓	PickS.↑	VLM↑	VLM↑
Open source	SDEdit (HoloCine)	0.311	0.321	0.409	0.535	0.364	0.315	0.894	0.222	0.184	0.491	0.988	5.56	19.4	0.711	0.193
	Memory-V2V	0.179	0.525	0.252	0.630	0.199	0.625	0.452	0.215	0.174	0.437	0.985	**4.82**	19.2	0.289	0.215
	Lucy Edit	0.174	0.729	0.192	0.647	0.185	0.751	0.750	0.213	0.171	0.351	0.984	5.82	19.2	0.508	0.235
	SANA-Streaming	0.141	0.651	0.430	0.651	0.212	0.707	0.490	0.219	0.181	0.592	0.984	5.48	19.3	0.504	0.354
	Wan-Edit	0.0816	0.840	0.355	**0.103**	0.809	0.692	0.217	0.178	0.515	0.983	5.38	19.3	0.568	0.378
Private	Gemini Omni Flash	0.153	0.705	0.834	0.568	0.193	0.684	0.893	0.221	**0.194**	0.834	0.987	4.95	**19.8**	**0.814**	0.586
	HappyHorse-1.0	0.246	0.729	0.788	0.400	0.291	0.728	0.894	0.219	0.193	0.815	0.984	5.12	19.6	0.691	0.623
	Aleph 2.0	0.0916	0.886	0.778	0.671	0.146	0.883	**0.981**	0.225	0.186	0.846	0.988	4.95	19.6	0.801	0.764
	MoviEdit (Ours)	**0.0741**	**0.904**	**0.861**	**0.695**	0.121	**0.911**	0.904	**0.227**	0.190	**0.884**	**0.989**	4.85	19.7	0.762	**0.790**

Table 2: Quantitative evaluation on the long-form multi-shot video editing benchmark, MoviEdit-Bench, for edited-frame-guided propagation (IV2V). (X-shot: cross-shot)

Edited Frame	Model	Non-subject Preservation	Edit Localization		Subject Preservation			X-shot Edit Consistency	Edit Prompt Alignment			Edited-frame Fidelity	Video Quality				Overall Edit Success
		LPIPS↓	Precision↑	Recall↑	MFS↑	LPIPS↓	VLM↑	VLM↑	CLIP↑	ViCLIP↑	VLM↑	VLM↑	DINO↑	NIQE↓	PickS.↑	VLM↑	VLM↑
First	HappyHorse-1.0	0.244	0.751	0.828	0.417	0.304	0.767	0.925	0.220	**0.195**	0.841	0.680	0.985	5.08	19.6	0.707	0.522
	GenProp	0.0788	**0.976**	0.671	0.698	0.104	0.993	0.585	**0.232**	0.186	0.761	0.561	0.989	**3.96**	19.7	0.789	0.581
	Aleph 2.0	**0.0412**	0.962	**0.974**	**0.702**	**0.0895**	0.985	0.906	0.231	0.192	**0.971**	0.819	0.989	4.90	**19.8**	**0.801**	**0.806**
	MoviEdit (Ours)	0.0586	0.965	0.938	0.699	0.107	**1.00**	**0.981**	0.231	0.191	0.938	**0.828**	**0.989**	4.67	19.8	0.797	0.803
Arbitrary	HappyHorse-1.0	0.248	0.742	0.793	0.402	0.307	0.760	0.913	0.219	**0.194**	0.813	0.636	0.984	5.11	19.6	0.709	0.486
	Aleph 2.0	**0.0422**	0.955	**0.952**	**0.705**	**0.0838**	0.981	0.865	0.229	0.193	**0.957**	0.801	0.989	4.94	**19.8**	0.807	0.781
	MoviEdit (Ours)	0.0598	**0.963**	0.920	0.699	0.0996	**0.995**	**0.933**	**0.229**	0.192	0.925	**0.805**	**0.989**	**4.78**	19.7	**0.807**	**0.786**

accessible public benchmark specifically targets long-form multi-shot editing, we additionally evaluate V2V performance on the established short-form benchmarks FiVE-Bench (Li et al., 2025) and VPBench (Bian et al., 2025). For VPBench, which is originally mask-conditioned, we evaluate MoviEdit in a mask-free setting using only the source video and edit instruction. We compare against various public and proprietary systems, including Lucy Edit Dev (DecartAI Team, 2025), Memory-V2V (Lee et al., 2026), SANA-Streaming (Zhao et al., 2026b), Wan-Edit (Li et al., 2025), SDEdit with HoloCine (Meng et al., 2022; 2026), HappyHorse-1.0 (Alibaba Cloud, 2026), Aleph 2.0 (Runway, 2026), Gemini Omni Flash (Google DeepMind, 2026), and GenProp (Liu et al., 2025).

3.2 Results

Comparisons. We compare MoviEdit with publicly available and proprietary systems on 15-second multi-shot editing using MoviEdit-Bench. For V2V, MoviEdit achieves the strongest overall performance among the evaluated systems, surpassing Runway Aleph 2.0 in overall edit success and leading on the majority of evaluation metrics (Tab. 1). For IV2V, MoviEdit achieves the highest overall success on the full benchmark and performs on par with Aleph 2.0 across the individual evaluation dimensions, with complementary strengths across different metrics (Tab. 2). The radar plot in Fig. 6 further highlights the balanced performance of MoviEdit across the major benchmark criteria. We further evaluate MoviEdit on the established FiVE-Bench and VPBench as external validation (Tab. 5), obtaining results consistent with those observed on MoviEdit-Bench.

Fig. 5 shows qualitative V2V and IV2V comparisons. Public baselines exhibit pronounced failure modes, including artifacts propagated from previous shots, incorrect cross-shot edit localization, edit failure, and poor source consistency. HappyHorse and Gemini Omni generally perform the requested

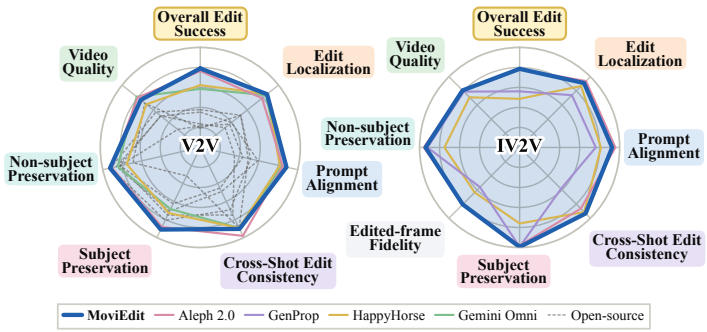

Figure 6: MoviEdit-Bench radar chart showing that MoviEdit achieves performance on par with proprietary systems.

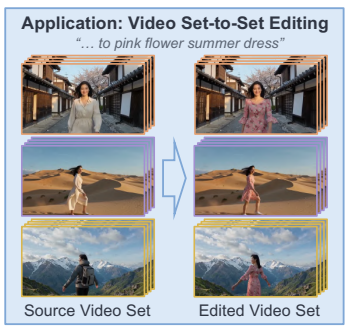

Figure 7: MoviEdit extends to downstream applications.

Table 3: Training dataset statistics at 540p, 10-15s, and 4-6 shots per video.

Task	Source	# raw pairs	# filtered pairs
Localization	Multi-cam	13,801	10,500
Pixel-aligned	Multi-cam	38,979	21,008
	Cinematic	138,633	50,405
Semi-pixel-aligned	Generative	13,815	4,706
Total		**205,228**	**86,619**

Table 4: Inference efficiency ablation at 540p and 361 frames on 8 H100 GPUs (seconds).

Variant	Draft Editor	Render Editor	Total
Full temporal attn.	25.1	339.0	364.1 (×1.0)
+ Sliding-window attn.	-	233.8	258.9 (×1.4)
+ KV cache	-	141.7	166.8 (×2.2)
+ Distillation	**3.6**	**8.2**	**11.8 (×30.9)**

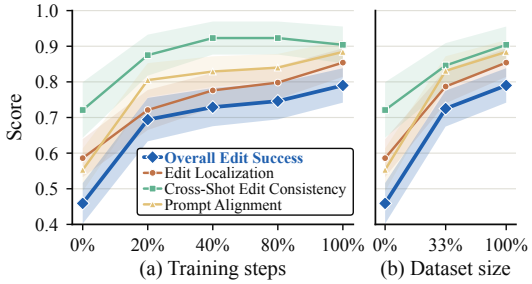

Figure 8: Scaling study on (a) training steps and (b) dataset size. X-axes show fractions of the full schedule (17K steps) / full dataset (87K pairs) used. Shading denotes 95% confidence intervals.

edits well but systematically deviate from the source, failing to preserve shot transitions and timing and hallucinating motion absent from the source. Aleph 2.0 provides the strongest baseline results, while MoviEdit shows better source preservation, edit localization, and prompt alignment. These trends are consistent with quantitative results on MoviEdit-Bench.

Scaling and Efficiency Analysis. We study MoviEdit's scaling behavior and inference efficiency. For scaling, we separately vary training duration and dataset size: 0%, 20%, 40%, 80%, and 100% of the 17K-step schedule, and 0%, 33%, and 100% of the 87K-pair dataset, respectively. Increasing either dataset size or training duration consistently improves performance across metrics (Fig. 8), suggesting further gains from scaling. We next ablate inference efficiency at 540p and 361 frames on 8 H100 GPUs (Tab. 4). Starting from full temporal attention (364.1 s), sliding-window attention reduces denoising latency to 258.9 s, and KV caching further reduces it to 166.8 s. Combined with few-step distillation, the full system runs in 11.8 s, yielding a 30.9× speedup over the full-attention baseline. Together, these optimizations reduce the denoising latency to 11.8 s, below the 15 s input-video duration, substantially improving the practicality of long-form multi-shot editing.

Applications. We explore applications beyond standard V2V and IV2V multi-shot editing, including settings not seen during training. First, MoviEdit supports *multi-reference-frame editing* (Fig. 1(b)); although trained with a single edited reference frame, it can condition on multiple edited references at inference time, combine edits to different subjects, and propagate them consistently. Second, we extend image set-to-set editing (Feingold et al., 2026) to videos with *video set-to-set editing* (Fig. 7). We concatenate multiple source videos into a single multi-shot sequence and edit them jointly, yielding consistent transformations despite variations in appearance and background. Finally, we evaluate *beyond-training-horizon editing* (Fig. 1(c)). Although trained on videos up to 15 seconds, MoviEdit generalizes to 25-second inputs with up to 17 shots, as well as long-take single-shot videos, extending both temporal duration and shot count beyond the training regime.

4 Conclusion

We presented MoviEdit, a unified framework for long-form multi-shot video editing that jointly addresses scalable supervision, efficient modeling, and dedicated evaluation. MoviEdit combines scalable pipelines for generating paired multi-shot editing data, cascaded bidirectional–autoregressive modeling, and MoviEdit-Bench for evaluating cross-shot localization, consistency, preservation, and edit alignment. Across text- and image-guided editing, MoviEdit substantially outperforms existing open-source models and matches strong proprietary systems, while generalizing to set-to-set editing and videos beyond the training horizon. Regarding limitations, MoviEdit can struggle with target localization under strong identity ambiguity, such as when multiple visually similar subjects are present or the target is only partially visible (e.g., body part, or object fragment). These limitations motivate several directions for future work. Agentic editing workflows could further improve edit fidelity and computational efficiency, for example by selectively processing only the shots or subjects that require editing. On the data side, controllable rendering engines (Engine, 2018) offer another scalable source of supervision. Finally, MoviEdit focuses on correspondence-preserving edits that retain the source motion and narrative structure; extending multi-shot long video editing to free-form transformations of motion and narrative structure remains an important future direction.

AI USE STATEMENT

In this work, we used generative AI tools for generating and processing synthetic data used in our experiments, including data generation and quality assurance, and for assisting with the implementation of research methods and experimental code. We did not use generative AI tools for other tasks requiring disclosure, such as formulating hypotheses or mathematical claims, developing proofs, designing the research methodology or experiments, conducting qualitative analysis, or interpreting experimental results; the remaining required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to aid in polishing the writing and improving the clarity and readability of the manuscript, and for literature retrieval and discovery. We reviewed all AI-assisted work. In particular, potentially relevant literature identified with AI assistance was manually checked by the authors for relevance and accuracy, AI-assisted implementations and code were reviewed and tested for correctness, and AI-generated or AI-processed synthetic data were inspected and validated before use in our experiments.

We take responsibility for the final content of this work, including text, claims, code, data, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work focuses on advancing long-form video editing and does not involve human-subject studies. The data used in this work are drawn from publicly available sources, internally collected or acquired datasets, and synthetically generated materials. All videos presented in this manuscript are either synthetically generated or have undergone facial blurring where real individuals appear. As with other generative video technologies, the proposed method could potentially be misused to create misleading or manipulated content. We therefore encourage responsible use of the technology and appropriate consideration of privacy, consent, copyright, and content authenticity when applying or deploying such systems. We also acknowledge that biases present in the training data may be reflected in generated outputs and leave further investigation of these issues to future work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the MoviEdit architecture, data-generation pipelines, training setup, and evaluation protocol in the main paper and supplementary materials. To facilitate reproducibility of the data construction process, we will release the code for our paired multi-shot data-generation pipelines. We will also release the evaluation data for MoviEdit-Bench together with the corresponding metric computation code, enabling standardized evaluation of cross-shot editing performance.

REFERENCES

Alibaba Cloud. Happyhorse video editing api reference. https://www.alibabacloud.com/help/en/model-studio/happyhorse-video-edit-api-reference, 2026.

Jianhong Bai, Menghan Xia, Xintao WANG, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di ZHANG. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9232d474be0a4e5f1e1bcb0765f17f9a-Paper-Conference.pdf.

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.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 92075–92093, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23206–23217. IEEE, 2023.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Zhou Daquan, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. In *International Conference on Learning Representations*, volume 2026, pp. 38973–38997, 2026.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161. IEEE, 2024.

DecartAI Team. Lucy edit: Open-weight text-guided video editing. Technical report, 2025. URL https://huggingface.co/decart-ai/Lucy-Edit-1.1-Dev. Model: Lucy-Edit-1.1-Dev.

Unreal Engine. Unreal engine. *Retrieved from Unreal Engine: https://www. unrealengine.com/en-US/what-is-unreal-engine-4*, 2018.

Kate Feingold, Omri Kaduri, and Tali Dekel. Match-and-fuse: Consistent generation from unstructured image sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30490–30499, 2026.

Kunyu Feng, Yue Ma, Bingyuan Wang, Yuefeng Wang, Zhiyuan Qin, Hao Cheng, Hao Li, Qifeng Chen, and Zeyu Wang. Mseditor: Toward consistent multi-shot video editing. *arXiv preprint arXiv:2608.17559*, 2026.

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.

Google DeepMind. Gemini omni flash. https://deepmind.google/models/model-cards/gemini-omni-flash/, 2026.

Xu Guo, Zhengxuan Wei, Xinghui Li, Hanzhuo Huang, Xinyu Liu, Xiangyang Luo, Min Wei, Yiran Zhu, Qiulin Wang, Yulong Xu, et al. Contextmaster: Interactive multi-shot video creation via fixed-budget sparse context routing. *arXiv preprint arXiv:2608.04956*, 2026.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17281–17291, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167283–167308. Curran Associates, Inc., 2025a. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f4823f831af67a3ef15e41a85434422a-Paper-Conference.pdf.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y Wang, Joan Lasenby, and Chun-Hao Huang. Spacetimepilot: Generative rendering of dynamic scenes across space and time. *arXiv preprint arXiv:2512.25075*, 2025b.

Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023.

Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *European Conference on Computer Vision*, pp. 358–376. Springer, 2024a.

Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9212–9221. IEEE, 2024b.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11164–11175. IEEE, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Cherry Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, et al. Editverse: Unifying image and video editing and generation with in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 137234–137255, 2026.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022. IEEE, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

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.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. Five-bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16672–16681. IEEE, 2025.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026.

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

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8599–8608. IEEE, 2024.

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, et al. Generative video propagation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17712–17722. IEEE, 2025.

Fuchen Long, Cong Wang, Zitao Gao, Wenhao Zhong, Yu Cheng, Xiaolu Hou, Yan Li, Xiao Cao, Xinlong Sun, Xi Chen, et al. Coinve-200k: A large-scale high-quality dataset for compositional instruction-guided video editing. *arXiv preprint arXiv:2608.17566*, 2026.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *10th International Conference on Learning Representations (ICLR)*, 2022.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 461–471, 2026.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2013.

Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Bassl: Boundary-aware self-supervised learning for video scene segmentation. In *Asian Conference on Computer Vision*, pp. 485–501. Springer, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205. IEEE, 2023.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15932–15942. IEEE, 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.

Runway. Runway aleph 2.0. `https://runway.com/news/introducing-aleph-2-and-edit-studio`, 2026.

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36046–36070. Curran Associates, Inc., 2024. doi: 10.52202/079017-1136. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/3f66d5cdbe032bb750f2dc523357b7a5-Paper-Conference.pdf`.

Alp Eren Sari and Paolo Favaro. Flowcut: Unsupervised video instance segmentation via temporal mask matching. *arXiv preprint arXiv:2505.13174*, 2025.

Wonyong Seo, Jaeho Moon, Jaehyup Lee, Soo Ye Kim, and Munchurl Kim. Propfly: Learning to propagate via on-the-fly supervision from pre-trained video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43228–43238, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Boyang Wang, Guangyi Xu, Jiahui Zhang, Zhipeng Tang, and Zezhou Cheng. Omnishotcut: Holistic relational shot boundary detection with shot-query transformer. *arXiv preprint arXiv:2604.24762*, 2026.

Wen Wang, Yan Jiang, 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.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024.

Chenyang Wu, Fuchen Long, Binyuan Huang, Xinlong Sun, Xi Chen, Chun-Le Guo, and Chongyi Li. Thinking on shots: Consistent multi-shot video editing with agentic reasoning. *arXiv preprint arXiv:2608.26809*, 2026.

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. 7623–7633. 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, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hao Yang, Zhiyu Tan, Jia Gong, Luozheng Qin, Hesen Chen, Xiaomeng Yang, Yuqing Sun, Yuetan Lin, Mengping Yang, and Hao Li. Omni-video 2: Scaling mllm-conditioned diffusion for unified video generation and editing. *arXiv preprint arXiv:2602.08820*, 2026a.

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, 2026b.

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8476. IEEE, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026a.

Yuyang Zhao, Yicheng Pan, Qiyuan He, Jincheng Yu, Junsong Chen, Tian Ye, Haozhe Liu, Enze Xie, and Song Han. Sana-streaming: Real-time streaming video editing with hybrid diffusion transformer. *arXiv preprint arXiv:2605.30409*, 2026b.

Yuqian Zhou, Zhenghong Zhou, Zongze Wu, Cameron Smith, Richard Zhang, Jiebo Luo, Eli Shechtman, and Zhe Lin. Editstream: A unified autoregressive framework for interactive video generation and editing. *arXiv preprint arXiv:2608.21424*, 2026.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0506. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/160e9ef249b9a836b7a82d342698af7a-Paper-Datasets_and_Benchmarks_Track.pdf.

SUPPLEMENTARY MATERIAL

This supplementary material provides additional details and results to complement the main paper. We first review related work on video editing and long-form video synthesis (Sec. A). We then provide additional details on our training dataset construction (Sec. B) and MoviEdit-Bench (Sec. C), including its design and evaluation metrics. Finally, we present additional experimental results on beyond-training-horizon editing (Sec. D.1), dataset recipe ablations (Sec. D.2), and comparisons on external benchmarks (Tab. 5), followed by a discussion of limitations and failure cases (Sec. D.3).

A RELATED WORK

Diffusion-based Video Editing. Early diffusion-based video editing methods largely repurpose pretrained generative models as editing priors through inference-time adaptation. Typical strategies include inversion (Qi et al., 2023; Wang et al., 2023), attention or feature manipulation (Qi et al., 2023; Liu et al., 2024; Geyer et al., 2024; Ceylan et al., 2023; Jeong et al., 2024a), auxiliary visual conditioning (Zhang et al., 2023; Jeong & Ye, 2023), and per-video optimization (Wu et al., 2023; Jeong et al., 2024b; 2025), enabling edits without requiring large-scale paired training data. While effective in many settings, these methods rely primarily on manipulating pretrained generative priors at inference time rather than learning generalizable editing behavior from data, which can limit robustness, scalability, and editing flexibility while introducing substantial inference overhead.

More recently, video editing has shifted toward a data-driven paradigm, where models learn editing behavior from paired source–target videos (DecartAI Team, 2025; Ju et al., 2026; Jiang et al., 2025; Liu et al., 2025; Polyak et al., 2024; Yang et al., 2026a; Seo et al., 2026; Lin et al., 2026; Lee et al., 2026; Huang et al., 2025b). These methods typically build on pretrained video generators and differ mainly in how source-video information is incorporated, ranging from channel-level (DecartAI Team, 2025) or token-level concatenation (Ju et al., 2026; Lee et al., 2026) to adapter-based additive conditioning (Jiang et al., 2025). They also vary in how the edit is specified; for example, GenProp (Liu et al., 2025) conditions on an edited reference frame and propagates the corresponding transformation through the video. Following the same data-driven philosophy, our models use token-level source conditioning and are trained separately for text-guided and edited-image-guided editing. Unlike prior reference-image-guided methods (Ouyang et al., 2024; Ku et al., 2024; Liu et al., 2025) that typically condition on an edited first frame, our image-guided model supports edited references from arbitrary frames. Complementary approaches introduce explicit reasoning or agentic control into the editing process, from latent reasoning over edit regions (Yang et al., 2026b) to LLM/VLM-driven shot-level planning for multi-shot editing (Wu et al., 2026).

More broadly, while most existing end-to-end video editing models remain focused on short, single-shot settings, our work systematically addresses long-form multi-shot editing through cross-shot supervision, long-context modeling, and dedicated evaluation.

Long-Form and Multi-Shot Video Synthesis. Recent advances have made long-form video generation increasingly tractable through autoregressive generation and efficient long-context modeling. Methods such as CausVid (Yin et al., 2025) and Self-Forcing (Huang et al., 2025a) generate videos autoregressively in chunks, using windowed attention or rolling KV caches to bound the context and computational cost over long horizons. In parallel, SANA-Video (Chen et al., 2026) and SANA-Streaming (Zhao et al., 2026b) replace softmax attention with linear attention (Katharopoulos et al.,

Table 5: Quantitative evaluation on the short-form single-shot video editing benchmarks FiVE-Bench and VPBench for text-guided video editing (V2V).

	Model	FiVE-Bench										VPBench						
		Structure	Background Preservation				Text Alignment		IQA	Motion	FiVE-Acc	Background Preservation				Text Alignment		Quality
		Dist.↓	PSNR↑	LPIPS↓	MSE↓	SSIM↑	CLIPS.↑	CLIPS.$_{\text{edit}}$↑	NIQE↓	Fidelity↑	VLM↑	PSNR↑	SSIM↑	LPIPS↓	MSE↓	CLIPS.↑	CLIPS.$_{\text{edit}}$↑	FVID↓
Open	Lucy Edit	15.49	<u>25.66</u>	99.65	50.52	82.15	25.41	20.33	6.30	87.95	0.3532	26.53	0.856	0.091	0.0095	11.11	<u>13.89</u>	<u>0.236</u>
Open	SANA-Streaming	<u>14.63</u>	24.25	111.66	56.80	80.52	24.82	19.78	6.22	86.44	0.1921	23.34	0.825	0.108	0.0100	<u>11.96</u>	13.86	0.252
Open	Wan-Edit	**11.36**	25.38	<u>96.88</u>	<u>34.92</u>	81.71	26.64	21.18	6.16	88.20	0.4010	<u>27.09</u>	<u>0.877</u>	**0.076**	**0.0023**	11.00	13.54	0.238
Private	Gemini Omni Flash	31.29	20.45	132.49	129.43	67.96	**28.02**	<u>22.24</u>	<u>6.09</u>	82.27	**0.8007**	23.44	0.789	0.102	0.0114	11.46	13.30	0.270
Private	HappyHorse-1.0	34.36	19.71	145.69	142.78	64.12	<u>27.85</u>	**22.30**	**5.88**	83.67	<u>0.7864</u>	21.19	0.695	0.126	0.0138	10.22	12.92	0.293
Private	Aleph 2.0	23.85	25.00	104.51	85.82	<u>83.68</u>	27.56	21.97	6.12	<u>88.69</u>	0.7784	26.41	<u>0.877</u>	0.094	0.0121	10.74	13.36	0.248
	MoviEdit (Ours)	16.86	**26.59**	**88.23**	**31.73**	**84.67**	26.95	21.57	6.32	**89.07**	0.6623	**28.30**	**0.885**	<u>0.082</u>	<u>0.0043</u>	**12.76**	**14.00**	**0.197**

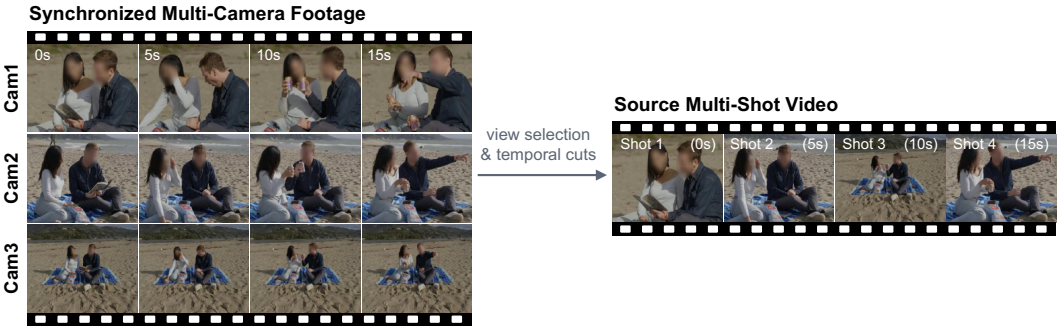

Figure 9: **Examples of synchronized multi-camera footage used as raw data for MoviEdit.** Each scene is captured simultaneously from 2–4 cameras, showing the same real-world scene from different viewpoints while actors perform natural interactions. *Note: Faces of real actors in the displayed examples are blurred to protect their identity2.*

2020), enabling efficient long-context modeling and constant-memory, constant-compute streaming generation. Few-step diffusion distillation has also been explored to reduce autoregressive sampling latency, including DMD- and consistency-based methods (Huang et al., 2025a; Zhao et al., 2026a) and, more recently, velocity moment matching (VMM) (Zhou et al., 2026).

Beyond a continuous sequence, recent work has begun to address multi-shot video generation by extending conventional short-form video models with two key modifications: training on multi-shot video data and introducing shot-aware sparse attention for efficient cross-shot context modeling. HoloCine (Meng et al., 2026) employs dense intra-shot attention while using the first-frame tokens of each shot as compact summary tokens for cross-shot information exchange. Mixture-of-Contexts (Cai et al., 2026) further introduces content-adaptive sparse attention that dynamically retrieves a small set of relevant historical chunks for each query, while ContextMaster (Guo et al., 2026) adopts fixed-budget context routing that combines mandatory task-specific context with dynamically selected historical information. MSEditor (Feng et al., 2026) instead leverages synchronized multi-view data to adapt a first-frame-conditioned video model for consistent appearance propagation across multiple shots. Despite their different implementations, these approaches improve scalability by limiting dense global interactions across the video, making it difficult to maintain a consistent appearance over long temporal spans.

B Training Dataset Construction

Raw Data Sources. We collect raw video clips from three sources: synchronized multi-camera footage (1,147 clips), cinematic footage (32,474 clips), and generative videos (27,630 clips). In particular, the synchronized multi-camera footage captures the same real-world scene simultaneously from 2–4 cameras, with actors performing natural interactions. This setup is conceptually similar to the synthetic SynCamMaster setup (Bai et al., 2025), but uses real-world footage. Representative examples are shown in Fig. 9.

Editing Pair Generation. Using these raw data sources, we construct the MoviEdit training dataset through three data-generation pipelines: localization, pixel-aligned, and semi-pixel-aligned. The resulting dataset contains 86,619 paired multi-shot videos at 540p resolution, approximately 15 seconds in duration, and 24 fps. The detailed breakdown by pipeline, data source, and editing task is provided in Table 6.

C MoviEdit-Bench

C.1 Source Multi-Shot Videos

MoviEdit-Bench comprises 64 source videos, primarily covering cinematic, story-driven multi-shot scenes with coherent space, time, and subject identity across shots. The scenes span diverse settings, including single-subject activities, interpersonal interactions, larger ensembles, and human–animal interactions, with deliberate variation in pose, viewpoint, visibility, and camera framing to support cross-shot editing evaluation.

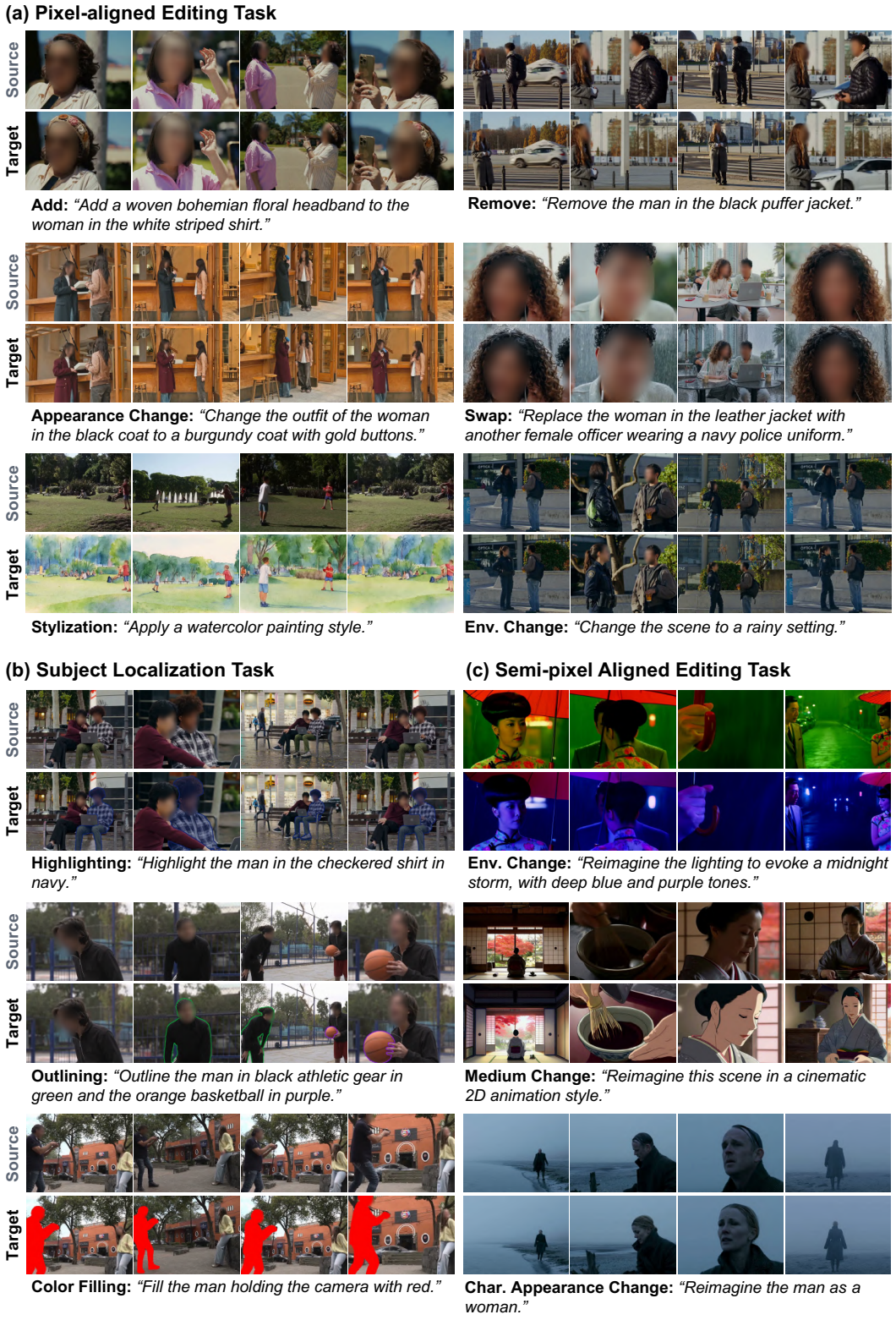

Figure 10: **Representative paired training data across editing tasks.** We show source and target video examples from our three data-generation pipelines. The pixel-aligned pipeline covers add, remove, appearance change, swap, stylization, and environment change; the localization pipeline covers highlighting, outlining, and color filling; and the semi-pixel-aligned pipeline covers environment change, medium change, and character appearance change. *Note: Faces of real actors in the displayed examples are blurred to protect their privacy.*

Pipeline	Data source	Editing task	# Pairs
Localization	Multi-camera	Highlighting	3,500
		Outlining	3,500
		Color filling	3,500
		Subtotal	10,500
Pixel-aligned	Multi-camera	Add	2,162
		Remove	3,209
		Appearance change	4,530
		Swap	2,926
		Stylization	3,841
		Environment change	4,340
		Subtotal	21,008
	Cinematic	Add	6,787
		Remove	11,921
		Appearance change	14,206
		Swap	8,742
		Stylization	1,606
		Environment change	7,143
		Subtotal	50,405
Semi-pixel-aligned	Generative	Environment change	1,364
		Medium change	1,368
		Character appearance change	1,974
		Subtotal	4,706
		Total	**86,619**

Table 6: Detailed composition of the MoviEdit training dataset by data-generation pipeline, data source, and editing task.

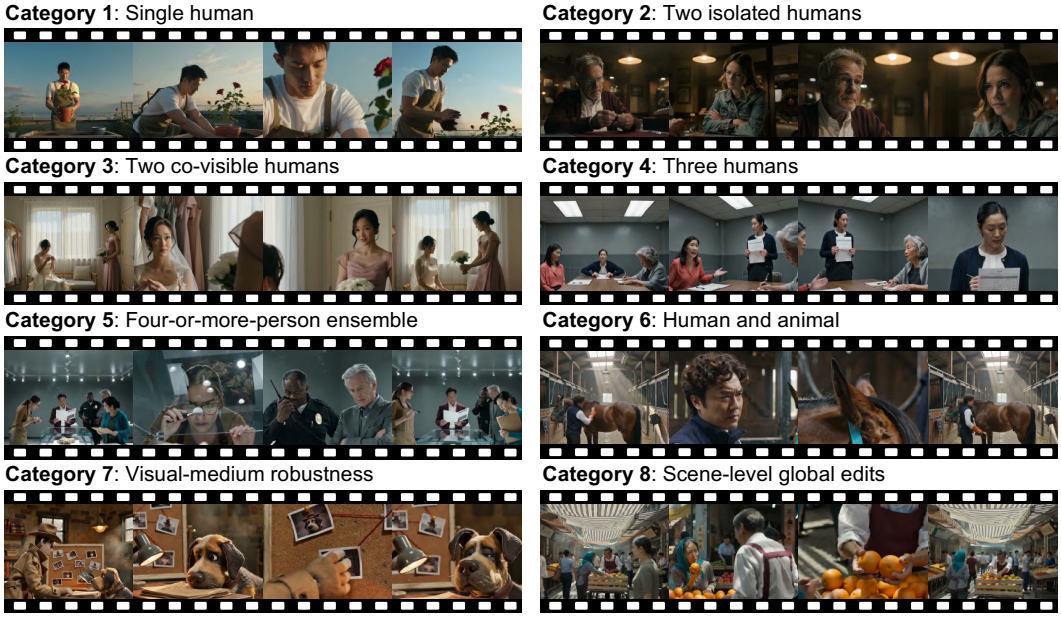

Figure 11: **Representative source videos from MoviEdit-Bench.** The examples highlight the diversity of subject composition, cross-shot grounding difficulty, and visual appearance in the benchmark. Full videos are available on the project page.

Category	Composition	# Videos	Source-video design	Primary evaluation focus
C1	Single human	6	One persistent subject observed across varying actions, poses, framings, and viewpoints.	Cross-shot localization under pose and viewpoint changes.
C2	Two isolated humans	6	Two humans belong to the same scene, while each shot contains only one of them.	Target selection across shots and correct exclusion when the target is absent.
C3A	Two co-visible humans, same gender	6	Same-gender subjects are distinguished by roles, clothing, hairstyles, accessories, positions, or object ownership.	Subject disambiguation between visually and semantically competing humans.
C3B	Two co-visible humans, different gender	6	Different-gender subjects co-occur within shots and are distinguished by roles and visual attributes.	Cross-shot subject binding with multiple co-visible candidates.
C4	Three humans	6	Three text-addressable characters share a continuous scene with distinct visual or semantic identifiers.	Localization under increased multi-person competition.
C5	Four-or-more-person ensemble	6	Larger ensembles contain distinguishable roles, appearance, props, and spatial positions.	Target localization in crowded multi-person scenes.
C6	Human and animal	6	A human and an animal coexist across a continuous multi-shot scene.	Cross-species target grounding and identity persistence.
C7	Visual-medium robustness	6	Scenes span non-photorealistic media such as stylized 3D, stop-motion, watercolor, claymation, anime, and paper-cut styles.	Robustness of localization and editing across visual media.
C8	Scene-level global edits	16	Scenes are designed for global transformations such as stylization, lighting, weather, and time-of-day changes.	Global editing without subject-localization constraints.
	Total	**64**		

Table 7: **Overview of the MoviEdit-Bench source-video categories.** C1–C6 progressively vary cross-shot subject-grounding difficulty, C7 evaluates robustness across visual media, and C8 targets scene-level global editing.

The source videos are organized into eight high-level categories (Tab. 7), with representative examples shown in Fig. 11. Categories C1–C6 progressively increase the difficulty of grounding an edit instruction to the correct subject across shots, ranging from a single persistent subject to crowded multi-subject and human–animal scenes. C3 is further divided into same-gender and different-gender two-person settings to examine different levels of subject ambiguity. C7 evaluates robustness across non-photorealistic visual media, while C8 targets scene-level global edits without entity localization.

Each scenario is authored by an LLM as a structured specification containing a global scene description, per-shot descriptions, and associated metadata. We generate each source video from scratch using a multi-shot text-to-video model, avoiding reliance on copyrighted source footage. Each video contains four to six shots and 361 frames at 24 fps, corresponding to approximately 15 seconds. The prompts, metadata, and generated videos are then manually reviewed for scene coherence, subject identity, shot correspondence, and suitability for editing evaluation.

C.2 Editing Scenarios

We construct two editing scenarios for each source video, yielding 128 scenarios in total. The 48 entity-centric videos in C1–C7 provide local edits spanning *add*, *remove*, *appearance change*, and *swap*, with a 3:1 ratio between subject targets (characters or animals) and object targets. The 16 videos in C8 provide scene-level global edits covering *environment change* and *stylization*. Overall, local and global edits account for 75% and 25% of the benchmark, respectively; the detailed task distribution is given in Tab. 8.

Editing scenarios and instructions are authored by an LLM from the source-video descriptions and structured entity metadata, following the predefined task distribution. All scenarios and edit instructions are subsequently manually reviewed for target correctness, instruction clarity, and suitability for evaluation.

Edit task	Scope	# Scenarios	Ratio
Add	Local	24	18.75%
Remove	Local	24	18.75%
Appearance change	Local	24	18.75%
Swap	Local	24	18.75%
Environment change	Global	24	18.75%
Stylization	Global	8	6.25%
Total		**128**	**100%**

Table 8: Distribution of the 128 MoviEdit-Bench editing scenarios across local and global edit categories.

For edited-frame-guided propagation, MoviEdit-Bench supports arbitrary edited-frame conditioning. We divide the 128 scenarios evenly into 64 first-frame and 64 non-first-frame reference settings. For each scenario, a suitable reference frame is manually selected, ensuring visibility of the edited target when applicable, and edited with an off-the-shelf text-guided image editor using the corresponding edit instruction.

C.3 Evaluation Metadata

Beyond the inputs required for benchmark evaluation, including the source video, edit instruction, and edited reference frame for the IV2V task, MoviEdit-Bench provides additional metadata for metric computation. For local edits, we provide per-frame masks of the editing target, obtained using video segmentation models (Cheng et al., 2024; Sari & Favaro, 2025; Carion et al., 2026) followed by manual verification and correction. The masks are temporally aligned with the source video and paired with the corresponding edit instruction. We also provide shot-boundary metadata for each source video, represented as the start and end frame indices of each detected shot using OmniShotCut (Wang et al., 2026) for shot-level metric computation.

Each editing scenario is further accompanied by a target-video prompt. Following the same format as the source-video generation prompt, it contains a global scene description together with per-shot descriptions. The target prompt is authored as a self-contained description of the fully edited target video, rather than as a relative modification of the source prompt, and is manually reviewed for consistency with the intended edit. These target prompts are also used to evaluate baselines formulated with target-video conditioning rather than standard edit instructions, such as Wan-Edit (Li et al., 2025) and SDEdit (Meng et al., 2022).

C.4 Evaluation Metrics

We evaluate each edited video along seven complementary dimensions: *non-subject preservation*, *edit localization*, *subject preservation*, *cross-shot edit consistency*, *edit-prompt alignment*, *video quality*, and *edited-frame fidelity*. We additionally report an *overall success* score that jointly considers multiple editing criteria. Each metric produces one score per editing scenario, and model-level results are averaged over applicable scenarios. Metric applicability varies by edit type, as summarized in Tab. 9.

Evaluation protocol. Methods may differ in output resolution and duration. We resize the source video, edited video, and target masks to a common 960×544 resolution and temporally align them to the 361-frame source video. When a method produces fewer frames, we evaluate only the overlapping range. The shot boundaries provided with MoviEdit-Bench define the units for shot-level evaluation, and temporal computations are performed independently within each shot. For mask-based metrics, we dilate the target masks by 9 pixels to exclude ambiguous boundary regions and omit frames or shots in which the target is not visible.

VLM judge. We use a VLM (Team et al., 2026; Yang et al., 2025) to evaluate semantic properties that are difficult to capture with pixel- or embedding-level similarity, such as whether an edit is applied to the correct subject. Unless otherwise specified, VLM-based metrics share a common frame-pair protocol. We uniformly sample four frames from each shot, including the first and last frames. For each sample, the judge receives the source frame, edited frame, edit instruction, and

target description, and independently determines whether: (i) the target is visible in the source frame, (ii) the requested edit is present, (iii) the edit is applied to the intended target, (iv) unrelated content is preserved, (v) target attributes unrelated to the instruction are preserved, and (vi) the edit completely covers the target. Frames are sampled from all shots, including those in which the target is absent, enabling the judge to detect edits that incorrectly leak to other subjects or shots. For global edits, the corresponding checks assess whether the requested scene-level change is present, covers the entire scene, and preserves the underlying scene content. All VLM metrics below reuse these judge outputs unless otherwise noted. Each binary VLM verdict is mapped to 1 for `true` and 0 for `false`. Frame-level VLM metrics are averaged over all applicable sampled frames, while scenario-level judgments, such as cross-shot consistency, directly yield a binary scenario score.

Non-subject preservation. This dimension measures preservation outside the intended edit region. For local edits, we compute LPIPS (Zhang et al., 2018) between corresponding source and edited frames after masking out the dilated target region, and average the score over frames.

Edit localization. This dimension measures whether the edit is confined to the intended target (*Precision*) and sufficiently covers it (*Recall*). Precision is the fraction of sampled frames passing VLM check (iv), penalizing unintended changes to competing subjects or to shots in which the target is absent. Recall is the fraction passing check (vi), measuring whether the intended edit is completed rather than only partially applied. For global edits, Recall instead measures whether the requested change covers the entire scene, while Precision is undefined.

Subject preservation. This dimension measures preservation of target properties that are not requested to change. *MFS* follows the Motion Fidelity Score (Yatim et al., 2024) used in FiVE-Bench (Li et al., 2025), adapted to multi-shot videos. We compute it independently for each target-bearing shot and average across shots. Within each shot, a 20×20 grid of points inside the source target bounding box is tracked with CoTracker3 (Karaev et al., 2025) in both source and edited videos, and the normalized frame-to-frame displacement directions are compared. Since MFS evaluates motion rather than appearance, it also applies to swaps and global edits. *LPIPS* (Zhang et al., 2018) compares the target regions between source and edited frames and is used for add and appearance-change edits, where the original subject identity should remain. *VLM* reports the fraction of frames passing check (v); for global edits, it measures preservation of the underlying scene content.

Cross-shot edit consistency. This dimension measures whether the requested edit remains consistent across shots. We jointly provide the VLM judge with edited frames sampled from all target-bearing shots and ask whether visible instances exhibit the same edit, such as the same added object, modified appearance, or substituted identity. The judge is instructed to assess consistency independently of edit correctness, which is evaluated by other metrics. This produces one binary score per scenario and is defined when at least two relevant shots are available. It does not apply to removal edits. For global edits, the metric evaluates whether the resulting environment or visual style remains consistent across shots.

Edited-frame fidelity. For the IV2V task, this dimension measures whether the propagated edit remains faithful to the edited reference frame. We include the reference frame in each frame-pair VLM query and evaluate whether the edited subject or scene preserves its reference identity and attributes while allowing natural variations in viewpoint, pose, lighting, and motion. The score is the fraction of positive judgments and is not defined for removal edits.

Edit prompt alignment. This dimension measures alignment between the edited video and the per-shot target descriptions provided in the evaluation metadata. Each shot is evaluated against its corresponding target description, including shots that should remain unchanged. For *CLIP*, we uniformly sample eight frames per shot, compute cosine similarity between each frame and its target description using CLIP ViT-L/14 (Radford et al., 2021), and average the scores. For *ViCLIP* (Wang et al., 2024), the same eight frames are encoded jointly as a video clip and compared with the text embedding of the target description. *VLM* reports the fraction of sampled frames passing check (ii).

Video quality. This dimension evaluates the edited video independently of edit correctness. *DINO* measures temporal coherence as the mean cosine similarity between DINOv2 (Oquab et al., 2023) features of adjacent frames within each shot. *NIQE* (Mittal et al., 2013) measures no-reference image quality and is averaged over frames, with lower values indicating better quality. *PickScore* (Kirstain et al., 2023) scores each frame against the corresponding per-shot target description. For *VLM*, we jointly provide the middle frame of each shot and ask the judge to rate visual quality from 1 to 5

Dimension	Metric	Local				Global	
		Add	Remove	App.	Swap	Env.	Style
Non-subject preservation	LPIPS	✓	✓	✓	✓	–	–
Edit localization	Precision (VLM)	✓	✓	✓	✓	–	–
	Recall (VLM)	✓	✓	✓	✓	✓	✓
Subject preservation	MFS	✓	–	✓	✓	✓	✓
	LPIPS	✓	–	✓	–	–	–
	VLM	✓	–	✓	–	✓	✓
Cross-shot edit consistency	VLM	✓	–	✓	✓	✓	✓
Edited-frame fidelity†	VLM	✓	–	✓	✓	✓	✓
Edit prompt alignment	CLIP, ViCLIP, VLM	✓	✓	✓	✓	✓	✓
Video quality	DINO, NIQE, PickS., VLM	✓	✓	✓	✓	✓	✓
Overall edit success	VLM	✓	✓	✓	✓	✓	✓

Table 9: **Applicability of the evaluation metrics across edit types.** A metric is undefined (–) when its underlying premise does not apply. †Reported only for edited-frame-guided methods.

based on naturalness and artifacts such as warping, blur, malformed regions, unnatural lighting, and geometry. We normalize the resulting score s to $[0, 1]$ as $(s - 1)/4$.

Overall edit success. Overall edit success provides a strict frame-level measure that combines the applicable VLM criteria. A sampled frame is considered successful only when the requested edit is present, applied to the intended target, and unrelated content is preserved. When applicable, it must additionally preserve non-edited target attributes, fully cover the target, and remain consistent with the edited reference frame. The final score is the fraction of successful frames. Since it reuses the frame-pair judgments above, no additional VLM queries are required.

We visualize editing results from multiple models across different score ranges (Fig. 12). The examples show a progression from unsuccessful edits with incorrect subject localization or incomplete edit execution to successful edits with consistent appearance across shots.

D Additional Results

D.1 Beyond-Training-Horizon Editing

We further investigate the ability of MoviEdit to generalize beyond its training horizon. Although trained on videos of up to 15 seconds, MoviEdit can extend to substantially longer sequences with diverse shot structures. Fig. 13 presents qualitative results on two 25-second videos: (a) a multi-shot video comprising 17 shots and (b) a continuous single-shot video. The former demonstrates consistent editing across frequent shot transitions, substantially exceeding both the duration and shot count seen during training, while the latter demonstrates sustained editing consistency over an extended continuous sequence. These results illustrate the ability of MoviEdit to handle longer editing horizons across different temporal structures.

D.2 Data Recipe Ablations

To examine the effect of our task-specific data recipes, we conduct small-scale ablation training by selectively removing different supervision pipelines. Fig. 15 shows qualitative comparisons for subject-localization and semi-pixel-aligned training data.

Semi-pixel-aligned data improves generalization to edits that require relatively larger changes in structure or shape. As shown in the top part of Fig. 15, the model trained with semi-pixel-aligned data better handles such edits. This complements our pixel-aligned data pipeline, which has difficulty constructing source–target editing pairs involving large structural or shape changes. Adding subject localization data improves the model's ability to identify the intended edit target. As shown in the bottom part of Fig. 15, the model trained with subject localization supervision exhibits improved edit localization compared with the variant trained without it. This observation is consistent

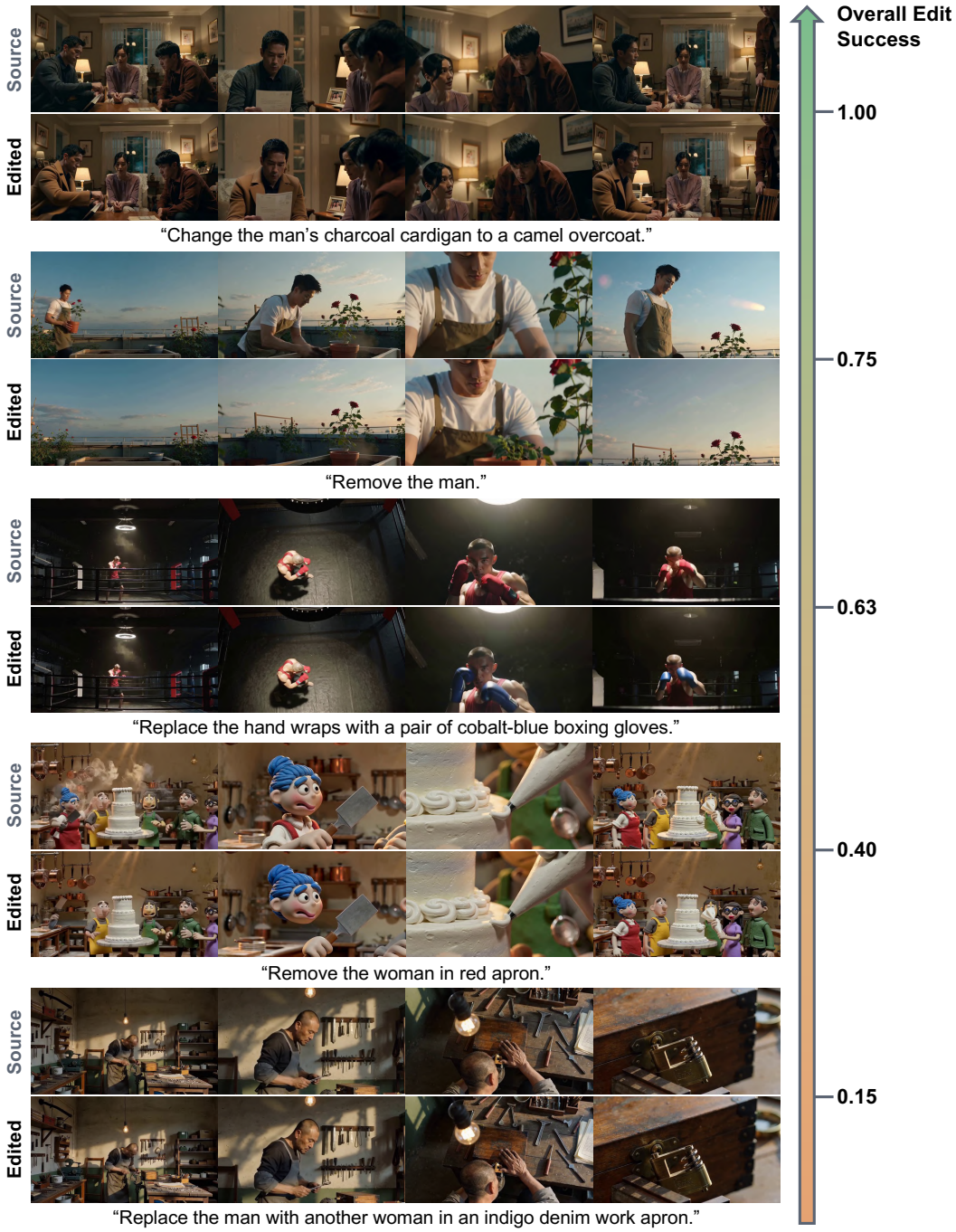

Figure 12: **Qualitative comparison across overall edit success scores.** Representative V2V results from multiple models, ordered by overall edit success. Examples progress from incorrect localization or incomplete edits to successful, cross-shot-consistent editing.

with prior findings in video editing (Liu et al., 2025), where task-specific tracking supervision has been shown to benefit editing performance.

D.3 Limitations and Failure Cases

Fig. 14 highlights two challenging target-localization failure modes. In Fig. 14(a), multiple visually similar subjects create identity ambiguity, which can cause the edit to be applied to the wrong target.

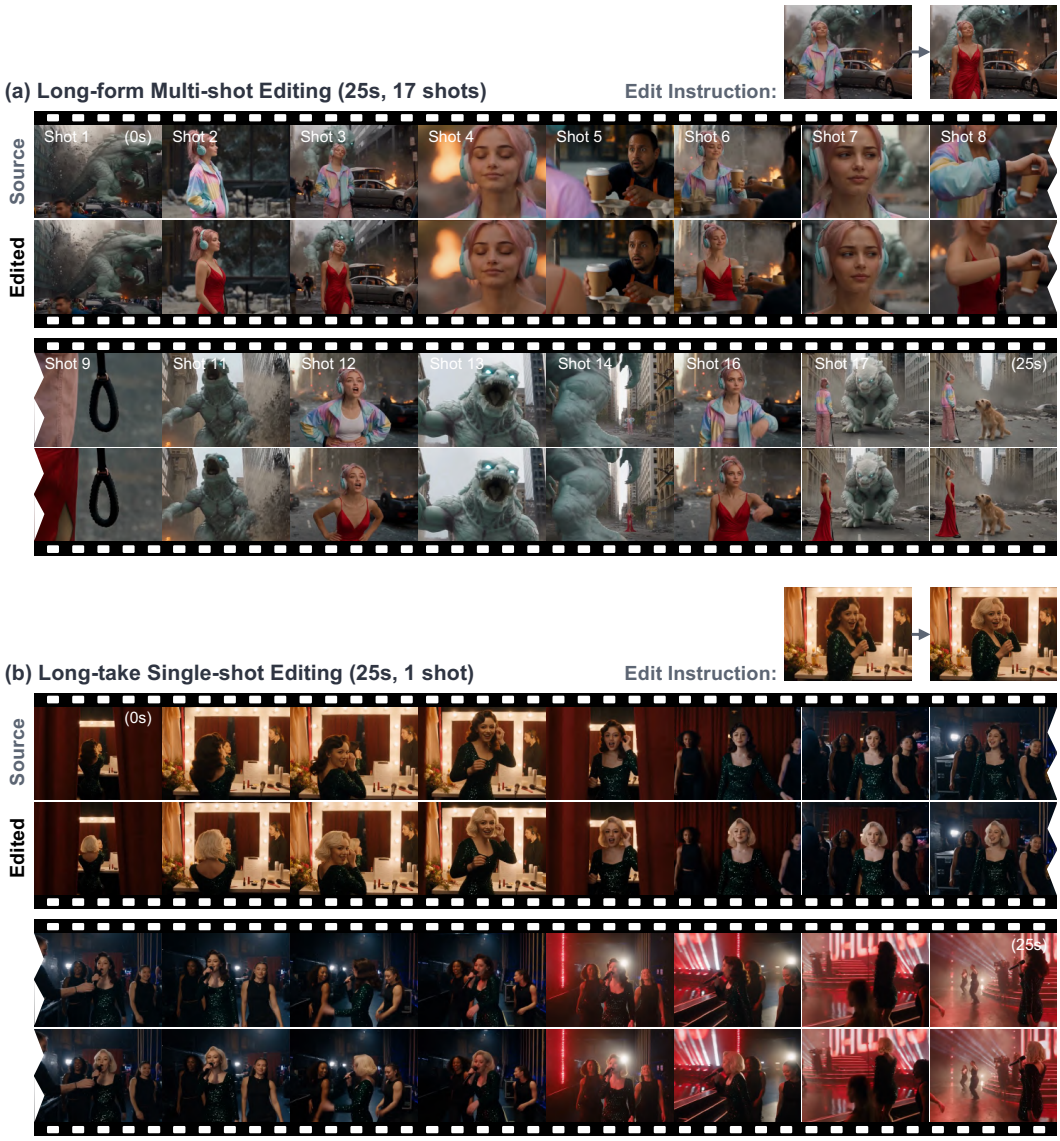

Figure 13: **Beyond-training-horizon editing.** Qualitative results on 25-second videos: (a) a 17-shot multi-shot sequence and (b) a continuous single-shot sequence. Both exceed the 15-second training horizon, demonstrating generalization to longer multi-shot and long-take videos.



Figure 14: **Failure cases of MoviEdit.** (a) **Inter-subject identity ambiguity.** When multiple visually similar subjects are present, the model may mistakenly apply the edit to an unintended subject rather than the specified target. (b) **Partial-observation ambiguity.** When the target is only partially visible, the model may struggle to accurately localize and edit the intended subject.



Figure 15: **Ablation of task-specific data recipes. Top:** Adding semi-pixel-aligned training data improves generalization to edits involving relatively larger changes in structure or shape. **Bottom:** Adding subject localization training data improves the model's ability to identify the intended edit target, leading to improved edit localization.

In Fig. 14(b), partial target visibility (e.g., only a body part or object fragment is visible) can make cross-shot association difficult, leading to incomplete or incorrect edits.