FLOWRESOLVE: CONSTRAINING MOTION AND ALLOCATING GUIDANCE IN VIDEO EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-driven editing modifies image or video content from natural language instructions. Inversion-free methods do it without mapping the image or video back to a noise latent, steering the sampling trajectory directly with an editing signal. On images this has shown impressive efficiency and structure preservation. Extending it to video remains challenging, and typically fails in three respects. The subject drifts away from the source motion, the background is perturbed, and the edit is not strong enough on the subject. Motion drift can arise because the backbone's prior does not explicitly enforce consistency with the source motion. The other two failures can result from a misallocated editing signal. Based on this analysis we propose FlowResolve, a training-free framework with three components. Multi-scale temporal self-similarity guidance (MTS) regularises temporal relations against the source. Edit-guided localisation (EGL) spatially attenuates excess target guidance, which can also reduce motion drift. Target attention focusing (TAF) strengthens target-token attention within a source-object mask. Together these mechanisms stabilise the editing signal and add an explicit motion constraint, guiding the flow-based evolution toward the desired edit. Extensive experiments on FiVE-Bench show that FlowResolve improves background preservation, motion fidelity and instruction following.

1 INTRODUCTION

Text-driven editing aims to modify image or video content according to natural language instructions. The rapid progress of this field has been driven by large-scale pretrained diffusion models (Ho et al., 2020; Song et al., 2021). Image generative models such as Stable Diffusion (Rombach et al., 2022; Podell et al., 2024) and FLUX (Labs, 2024) encode visual and semantic priors that let them synthesise high-quality, semantically aligned images from text alone. Video generative models (Wan et al., 2025; Yang et al., 2025b; Kong et al., 2024) extend them from static images to video. *Training-free* video editing (Geyer et al., 2024; Ceylan et al., 2023; Li et al., 2026; Cai et al., 2025) builds on both, using a frozen backbone with the text prompt. These methods divide into two classes: inversion-based and inversion-free.

The first inverts the source into a noise latent before editing, either on a text-to-image model inflated along time (Geyer et al., 2024; Cong et al., 2024; Qi et al., 2023; Kara et al., 2024; Yang et al., 2025a) or on a video backbone (Ku et al., 2024; Wang et al., 2025b). The inversion is only approximate, and its error accumulates across steps, perturbing the background and making the result unfaithful to the source. Recent work reduces this error directly (Jiao et al., 2026; Xie et al., 2025; Wang et al., 2025a), but does not eliminate it. To avoid this explicit inversion stage, a second class of methods edits without inversion. FlowEdit (Kulikov et al., 2025) integrates the difference of two guided velocity fields along a directly constructed path, and later work regularises that trajectory (Kim et al., 2026). Both are derived for single images, and the formulation transfers unchanged to a rectified-flow video backbone (Li et al., 2025). But the transferred update has no explicit source-referenced temporal objective, so temporal behaviour is left largely to the video backbone. In video the edit also reaches well outside the region the prompt describes, perturbing the background. A newer line designs the steering directly in the video domain instead (Li et al., 2026; Cai et al., 2025; Chen et al., 2026). They are fast and structurally faithful. But they do not constrain the result's motion against the source, and the editing signal still spills beyond the region the prompt describes. Editability, background preservation and source-consistent motion are not yet achieved together.

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.

To address two limitations of existing inversion-free methods—preserving source motion and localising the editing signal, we propose FlowResolve, a training-free framework for flow-based video editing. First, motion is constrained by the backbone prior, and that constraint cannot guarantee consistency with the source motion. The standard update does not explicitly compare the temporal relations of the edited and source videos, so motion can drift across sampling steps. We introduce *multi-scale temporal self-similarity guidance* (MTS) to supply the missing constraint. The sampler already forms an estimate of the finished video at every step. We extract motion descriptors from that estimate, largely free of static appearance, and hold their temporal self-similarities to the source's at several temporal strides. Second, the editing signal is misallocated. At each step the editor combines four velocity fields under two guidance scales. Written out, the combination separates into terms that behave differently. One of them subtracts the source's guidance direction from the target's, providing a cue for the edited region. Another references the target alone, with no source direction subtracted from it, and can remain active outside that region. It can therefore perturb the background while the edit stays weak on the subject. We introduce *edit-guided localisation* (EGL), which rescales the second term by a map derived from the first. This confines the edit without a segmentation mask, while attenuating the term computed from the target alone outside the edited region to reduce unintended motion drift. EGL determines where the editing signal acts, while *target attention focusing* (TAF) controls the strength of semantic editing within that region. TAF achieves this by concentrating attention on target-relevant tokens within the region.

In summary, FlowResolve integrates an source-motion constraint with spatially localised guidance in a training-free, inversion-free video editing framework. On FiVE-Bench, it reduces the endpoint error against source motion by 20.5% and improves background PSNR by 1.68 dB over Wan-Edit (Li et al., 2025), while increasing the CLIP-T score in the edited region from 24.81 to 25.47.

2 Related Work

Inversion-based video editing. The dominant early recipe adapts a text-to-image (T2I) diffusion model (Rombach et al., 2022; Podell et al., 2024) to video. Attention is shared, propagated or otherwise inflated across frames. This is done along inter-frame correspondences (Geyer et al., 2024; Wang et al., 2024) or optical flow (Cong et al., 2024), by conditioning each frame on the previous (Ceylan et al., 2023; Yang et al., 2023), or by rearranging and modulating space-time attention (Qi et al., 2023; Kara et al., 2024; Yang et al., 2025a). The resulting temporal coherence is hand-designed, not learned from video. It is also mostly local: adjacent frames agree, but longer spans are not explicitly constrained. Later work keeps the same inversion-then-edit recipe but swaps the image model for a video one (Ku et al., 2024; Wang et al., 2025b), which supplies the missing video prior. On rectified-flow backbones the inversion itself is improved, by a predictor–corrector scheme (Jiao et al., 2026), by noise alignment (Xie et al., 2025), or by a higher-order correction to the ODE solve (Wang et al., 2025a). They all invert first, and the inversion error accumulates across steps (Mokady et al., 2023; Xie et al., 2025).

Inversion-free video editing. Inversion-free methods avoid the inversion altogether. On rectified flow (Liu et al., 2023; Lipman et al., 2023; Esser et al., 2024; Labs, 2024), FlowEdit (Kulikov et al., 2025) integrates the difference of two guided velocity fields along a directly constructed path, and FlowAlign (Kim et al., 2026) regularises that trajectory with a terminal-point similarity to the source. Both work well on images, and the formulation carries over to a rectified-flow video backbone unchanged. FiVE-Bench (Li et al., 2025) does exactly that with Wan2.1, obtaining Wan-Edit, its best overall performer. Other methods instead design the editing update specifically for video. FlowDirector (Li et al., 2026) adds attention-guided masking and direction-aware flow correction, DFVEdit (Cai et al., 2025) operates on a conditional delta flow vector for efficiency, and FlowAnchor (Chen et al., 2026) treats the imprecise localisation and attenuated magnitude of the editing signal with separate modules. However, none of them separates the part of the update that uses the source from the part that ignores it. The resulting update can still extend beyond the region specified by the prompt, causing unintended background drift.

Attention control. Prompt-to-Prompt (Hertz et al., 2023) established the cross-attention maps of a text-conditioned diffusion model as an editable interface. It supports both injection of source maps and reweighting of a chosen token's map. Plug-and-Play (Tumanyan et al., 2023) injects spatial

features to hold structure fixed, and MasaCtrl (Cao et al., 2023) converts self-attention into mutual self-attention so a non-rigid edit can still query the source. Video counterparts extend the same controls across frames (Liu et al., 2024; Jeong & Ye, 2024). FlowAnchor modulates cross-attention inside a mask too, pulling each entry toward the largest or smallest response in its row. A single outlier may then pull the whole row toward it, and since the modulation follows the softmax, the weights need not sum to one. Its own sweep reports fidelity degrading as the interpolation strength grows. TAF instead biases the attention before the softmax, so the map stays a distribution and the backbone is recovered when the bias is off (Section 3.4).

3 METHOD

Given a source video, a source prompt, and a target prompt, FlowResolve modifies the flow update while keeping the backbone and prompts unchanged. MTS injects source motion into the update to preserve the motion of the source video (Section 3.2), EGL controls the guidance contribution to balance editing strength and source fidelity (Section 3.3), and TAF controls the conditional signal for the target region to enhance the desired edit (Section 3.4). Fig. 1 gives an overview of FlowResolve together with the intermediate signals used by these modules.

3.1 PRELIMINARIES

Rectified flow (Liu et al., 2023; Lipman et al., 2023) connects the data distribution π_0 and the noise distribution $\pi_1 = \mathcal{N}(0, I)$ through a linear interpolation. Let Z_t denote the intermediate state at time t, defined by

$$Z_t \;=\; (1-t)\, X_0 \;+\; t\, X_1, \qquad t \in [0,1], \tag{1}$$

with $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$. The velocity field learned from this interpolation defines the sampling ODE, whose individual trajectories need not follow the straight lines in Eq. 1:

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} \;=\; V(Z_t, t), \tag{2}$$

which defines the continuous transport from the data distribution to the noise distribution. Sampling integrates this ODE from $t = 1$ to $t = 0$. For a text condition $\mathcal{C}$, classifier-free guidance (Ho & Salimans, 2021) combines the conditional and unconditional fields with a guidance scale ω that controls prompt adherence,

$$\tilde{V}(Z, t, \mathcal{C}, \omega) \;=\; V(Z, t, \varnothing) \;+\; \omega\big(V(Z, t, \mathcal{C}) - V(Z, t, \varnothing)\big). \tag{3}$$

FlowEdit (Kulikov et al., 2025) edits without inversion. It maintains an editing state Z^{edit}, initialised at the encoded source X^{src}, and advances it with a velocity difference. At step t_i, it draws $\epsilon_{t_i} \sim \mathcal{N}(0, I)$ and forms a noisy source and a noisy target that share this noise:

$$Z^{\mathrm{src}}_{t_i} = (1-t_i)\, X^{\mathrm{src}} + t_i\, \epsilon_{t_i}, \qquad Z^{\mathrm{tar}}_{t_i} = Z^{\mathrm{edit}}_{t_i} + Z^{\mathrm{src}}_{t_i} - X^{\mathrm{src}}, \tag{4}$$

$$\Delta V_{t_i} = \tilde{V}\big(Z^{\mathrm{tar}}_{t_i}, t_i, \mathcal{P}^*, \omega_{\mathrm{tgt}}\big) - \tilde{V}\big(Z^{\mathrm{src}}_{t_i}, t_i, \mathcal{P}, \omega_{\mathrm{src}}\big), \qquad Z^{\mathrm{edit}}_{t_{i-1}} = Z^{\mathrm{edit}}_{t_i} + (t_{i-1} - t_i)\, \Delta V_{t_i}, \tag{5}$$

where $\mathcal{P}$ and $\mathcal{P}^*$ are the source and target prompts. Because the two branches share ϵ_{t_i}, their noise-dependent components largely cancel. They are guided with different strengths, $\omega_{\mathrm{tgt}} > \omega_{\mathrm{src}}$. Omitting the step index, we write $V^{\mathrm{src,c}}_t, V^{\mathrm{src,u}}_t$ for the conditional and unconditional velocities at Z^{src}_t, and $V^{\mathrm{tar,c}}_t, V^{\mathrm{tar,u}}_t$ for those at Z^{tar}_t. C and F are the numbers of latent channels and frames.

3.2 MULTI-SCALE TEMPORAL SELF-SIMILARITY GUIDANCE

The generative video prior model provides no guarantee that the generated motion matches the source video. In Eq. 5, the update is driven by a velocity difference evaluated at the current latents, with no term enforcing the temporal structure of the source. Since the target branch is conditioned on a different prompt, each update can shift the motion slightly away from the source. These shifts accumulate over the sampling steps, and the motion gradually drifts from the source.

A natural solution is to constrain the motion explicitly during editing. However, directly comparing the edited and source latents also penalises appearance changes, which are the purpose of editing.

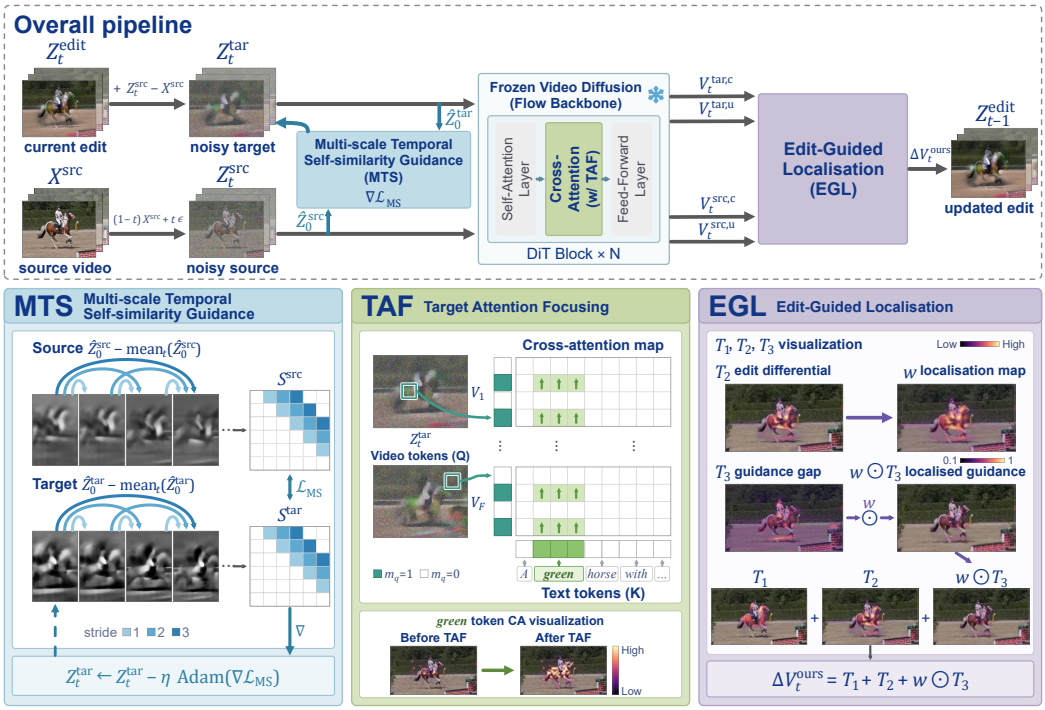


Figure 1: **Overview of FlowResolve.** *Top:* one sampling step with the three components at their points of intervention. MTS corrects the noisy target latent before the velocities of the update are computed, TAF biases cross-attention in the target conditional pass, and EGL rescales the guidance gap in the velocity difference. *Bottom:* MTS aligns the multi-stride temporal self-similarities of the target with those of the source; TAF amplifies the attention of video tokens in the edited region to the target tokens; EGL derives a localisation map w from T_2 and applies it to T_3.

Optical-flow-based warping can become unreliable under large motion, where drift matters most. To address these limitations, we introduce MTS, which preserves motion by matching the temporal relations within the source and target videos rather than their appearance. Specifically, we remove the temporal mean at each spatial location to suppress static appearance while preserving moving content, and then compare frame-to-frame similarities within each video (Fig. 1). The resulting relations therefore mainly reflect how each video changes over time.

MTS computes these relations on the clean-sample prediction $\hat{Z}_0 = Z - t\,\tilde{V}$, since the noisy latent retains little recoverable motion information at high noise levels. Let $f_n^{(p)}$ denote the resulting local descriptor at spatial position p of latent frame n (Appendix B). Adjacent latent frames are strongly correlated, so their similarities stay near saturation even when the motion gradually drifts. We therefore compare each frame with the one s frames later, for every stride s in a set $\mathcal{S}$. The multi-scale motion loss is defined as

$$\ell_s = \frac{1}{(F-s)P} \sum_{n=1}^{F-s} \sum_{p=1}^{P} \left| \left\langle f_n^{\mathrm{tar},(p)}, f_{n+s}^{\mathrm{tar},(p)} \right\rangle - \left\langle f_n^{\mathrm{src},(p)}, f_{n+s}^{\mathrm{src},(p)} \right\rangle \right|, \quad \mathcal{L}_{\mathrm{MS}} = \frac{\sum_s \lambda_s \ell_s}{\sum_s \lambda_s}, \tag{6}$$

where P is the number of positions and λ_s weights stride s. The source descriptors come from the source-branch prediction $\hat{Z}_0^{\mathrm{src}}$, which shares the noise and timestep of the target branch. During the early sampling steps, MTS corrects the target latent by gradient descent on this objective,

$$Z_t^{\mathrm{tar}} \leftarrow Z_t^{\mathrm{tar}} - \eta \cdot \mathrm{Adam}\Big(\nabla_{Z_t^{\mathrm{tar}}} \lambda_{\mathrm{MS}} \mathcal{L}_{\mathrm{MS}} \Big), \tag{7}$$

and maps the result back to Z_t^{edit} through Eq. 4. When computing the gradient, the velocity in the clean-sample prediction is treated as a constant, as detailed in Appendix B. MTS therefore uses a stop-gradient approximation without backpropagating through the backbone.

This design addresses the cause identified above. The objective vanishes whenever the target reproduces the temporal self-similarities of the source, encouraging motion preservation without directly

matching the source appearance. Because the latent is corrected before the velocities are computed, the current update already starts from the corrected motion, and deviations are reduced before they accumulate.

3.3 EDIT-GUIDED LOCALISATION OF THE GUIDANCE GAP

The second failure is a perturbed background. Since the two branches share their noise, one might expect their velocities to cancel wherever the two prompts agree. To find what does not cancel, we expand both guided velocities of Eq. 5 with Eq. 3, write $\omega_{\mathrm{tgt}} = \omega_{\mathrm{src}} + \Delta\omega$, and regroup the terms:

$$\begin{aligned}
\Delta V_t &= \tilde{V}\big(Z_t^{\mathrm{tar}}, t, \mathcal{P}^*, \omega_{\mathrm{tgt}}\big) - \tilde{V}\big(Z_t^{\mathrm{src}}, t, \mathcal{P}, \omega_{\mathrm{src}}\big) \\
&= \big[V_t^{\mathrm{tar,u}} + \omega_{\mathrm{tgt}}\big(V_t^{\mathrm{tar,c}} - V_t^{\mathrm{tar,u}}\big)\big] - \big[V_t^{\mathrm{src,u}} + \omega_{\mathrm{src}}\big(V_t^{\mathrm{src,c}} - V_t^{\mathrm{src,u}}\big)\big] \\
&= \big(V_t^{\mathrm{tar,u}} - V_t^{\mathrm{src,u}}\big) + (\omega_{\mathrm{src}} + \Delta\omega)\big(V_t^{\mathrm{tar,c}} - V_t^{\mathrm{tar,u}}\big) - \omega_{\mathrm{src}}\big(V_t^{\mathrm{src,c}} - V_t^{\mathrm{src,u}}\big) \\
&= \underbrace{\big(V_t^{\mathrm{tar,u}} - V_t^{\mathrm{src,u}}\big)}_{T_1\text{: latent gap}} + \underbrace{\omega_{\mathrm{src}}\big[\big(V_t^{\mathrm{tar,c}} - V_t^{\mathrm{tar,u}}\big) - \big(V_t^{\mathrm{src,c}} - V_t^{\mathrm{src,u}}\big)\big]}_{T_2\text{: edit differential}} + \underbrace{\Delta\omega\big(V_t^{\mathrm{tar,c}} - V_t^{\mathrm{tar,u}}\big)}_{T_3\text{: guidance gap}} \\
&= T_1 + T_2 + T_3. \qquad (8)
\end{aligned}$$

The three terms play different roles. T_1 compares the unconditional predictions at the two latents. Without the MTS correction, $Z_t^{\mathrm{tar}} = Z_t^{\mathrm{src}}$ at the first step, so T_1 is exactly zero there, and it is expected to stay small while the edit is local. T_2 contrasts the target and source guidance directions and tends to be small where the two prompts agree, so it encodes *what* should change and *where*. T_3, in contrast, is the target guidance scaled by $\Delta\omega$, with no source term subtracted, so it need not vanish where the prompts agree. As visualised in Fig. 1, T_2 concentrates on the edited object, whereas T_3 also responds in regions the prompt does not ask to change. T_3 is therefore one potential source of background changes. Because it does not reference the source, it may also disturb the motion there. Section 4.4 is consistent with this: localising T_3 lowers the endpoint error.

Suppressing T_3 globally is no remedy, since $\Delta\omega$ also sets the edit strength inside the edited region. Lowering it protects the background only by weakening the edit, and raising it strengthens the edit only by perturbing the background further. One global scale cannot separate the two effects.

We therefore separate them explicitly: EGL controls *where* T_3 acts, and TAF strengthens the edit only inside the edited region. For EGL, T_3 must be confined to the edited region without an external mask. The decomposition already marks this region through T_2. We thus use the model's own edit signal as the localisation cue and rescale T_3 with a spatial map derived from $|T_2|$:

$$\Delta V_t^{\mathrm{ours}} \;=\; T_1 + T_2 + w \odot T_3, \qquad w \;=\; \beta + (1-\beta)\,\min\Big(\frac{\mu}{q_\alpha(\mu)},\, 1\Big), \qquad \mu \;=\; \frac{1}{C}\sum_{c=1}^{C} \big|T_2^{(c)}\big|, \quad (9)$$

where $q_\alpha(\mu)$ is the α-quantile of μ and $\beta \in [0, 1]$ is a floor (Appendix B).

This rescaling is effective for three reasons. First, w follows the edit at every step, so EGL itself needs no segmentation mask. w is high on the edited object and falls to β elsewhere, so $w \odot T_3$ keeps the guidance on the subject while largely suppressing it in the background (Fig. 1). Second, it reduces the source-agnostic push outside the edited region, which also benefits motion fidelity, as discussed above. Third, for fixed velocity predictions, it generalises the unweighted velocity difference: when setting $\beta = 1$, we have $w \equiv 1$, and Eq. 9 reduces to the original update of Eq. 5. EGL does not replace MTS: it only reduces a perturbation, whereas MTS aligns the motion with the source.

3.4 TARGET ATTENTION FOCUSING

The third failure is a weak edit: inside the edited region, the subject changes less than the prompt asks. Since $w \leq 1$, EGL can only control where the target guidance acts, but cannot strengthen it. Increasing ω_{tgt} would instead strengthen the edit globally, affecting the background (Section 3.3).

To analyse this failure, we trace how the target semantics reach the update. The edit is driven by the target guidance direction $V_t^{\mathrm{tar,c}} - V_t^{\mathrm{tar,u}}$, whose two velocities differ only in the text condition. The text enters the backbone through cross-attention, where the target tokens compete with all

other prompt tokens for attention. Inside the edited region, they receive only sparse, weak attention (Fig. 1), so the target semantics are under-represented in $V_t^{\mathrm{tar,c}}$ where the edit should happen.

We therefore strengthen the edit where the target semantics enter the model, rather than through a global guidance scale. Cross-attention links every spatial location to the prompt words (Hertz et al., 2023), so it can be strengthened locally and for the target words only. TAF amplifies the attention that queries inside the edited region pay to the target tokens, by biasing the pre-softmax logits:

$$A \;=\; \mathrm{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d}} + B\right), \qquad B_{q,k} \;=\; \log\gamma \cdot m_q \cdot \mathbb{1}[k \in \mathcal{T}], \tag{10}$$

where $m_q \in \{0,1\}$ indicates whether query q lies in the edited region, $\mathcal{T}$ is the set of target tokens in $\mathcal{P}^*$, $\mathbb{1}[\cdot]$ is the indicator function, and $\gamma > 1$ controls the amplification (Appendix D).

This bias increases attention to target tokens relative to other prompt tokens within the edited region (Fig. 1). Because it acts before the softmax and only in the target conditional pass, attention remains normalised and only $V_t^{\mathrm{tar,c}}$ is directly modified. This velocity enters both T_2 and T_3, so TAF strengthens the semantic signal while also affecting the weights subsequently computed by EGL.

4 Experiments

4.1 Experimental Settings

Implementation. We build FlowResolve on the Wan2.1-T2V-1.3B model (Wan et al., 2025). Following Wan-Edit (Li et al., 2025), we use 50 UniPC steps with editing strength 0.7 and $(\omega_{\mathrm{src}}, \omega_{\mathrm{tgt}}) = (5.0, 10.0)$. For TAF, we use $\gamma = 3.5$ over the normalised-time window $[0.6, 1.0]$. For EGL, we set $\beta = 0.1$ and $\alpha = 0.98$. For MTS, we use latent strides $\mathcal{S} = \{1, 2, 3\}$ with weights $\lambda = (1, 2, 4)$, over $G = 10$ guidance steps with $J = 2$ inner updates per step. The resulting overhead over Wan-Edit is limited to J additional target-branch forwards on each of the first G steps. All experiments for FlowResolve are conducted on eight NVIDIA RTX 3090 GPUs.

Dataset and Baselines. We evaluate on the FiVE-Bench (Li et al., 2025), which contains 419 text-video editing pairs with precise masks, spanning six fine-grained editing types: object replacement (rigid and non-rigid), object addition, object removal, colour editing, and material editing, covering both real and generated videos. We compare against six training-free methods, including four inversion-based methods, TokenFlow (Geyer et al., 2024), RAVE (Kara et al., 2024), UniEdit-Flow (Jiao et al., 2026), and DNAEdit (Xie et al., 2025), and two inversion-free methods, FlowDirector (Li et al., 2026) and Wan-Edit (Li et al., 2025).

Evaluation metrics. We quantitatively compare all methods across four aspects.

(1) *Text Alignment*. We use CLIP-T (Radford et al., 2021) to measure global semantic alignment with the target prompt and Local CLIP-T (L.CLIP-T) to measure alignment within the edited region. Since global alignment can also be influenced by content outside the edited region, we treat L.CLIP-T as the primary measure of edit-specific semantics and report both.

(2) *Background Preservation*. We use PSNR and LPIPS (Zhang et al., 2018) computed outside the instance mask, denoted as M.PSNR and BG.LPIPS, to measure how well the unedited background is preserved and to quantify editing spillover.

(3) *Temporal Consistency*. We distinguish self-consistency from motion fidelity, as they capture different properties. Self-consistency is measured by warping SSIM (Lai et al., 2018) (WarpSSIM) and inter-frame CLIP similarity (CLIP-F), which assess whether adjacent frames remain coherent with one another, but do not compare the edited motion with the source trajectory.

(4) *Motion Fidelity*. We evaluate whether the edited video preserves the *source* motion using two complementary measures: endpoint error against the source optical flow (Teed & Deng, 2020), and the tracklet-correlation motion-fidelity score of Yatim et al. (2024) over point tracks (Karaev et al., 2024). Both are computed over the full frame (EPE, MFS).



Figure 2: **Qualitative comparison of training-free video editing methods.** Our method preserves the background and source motion while performing the desired edit. The prompt above each block indicates the edited word.

4.2 Qualitative Results

We first provide qualitative comparisons with representative training-free video editing methods. As shown in Fig. 2, existing methods can produce visually plausible semantic edits, but they may also introduce unintended changes outside the target region or alter the motion and appearance of the source content. In particular, several baselines either modify background regions or deviate from the original object trajectory. In contrast, FlowResolve more consistently confines the edit to the intended region while preserving the surrounding scene and source motion.

The two examples in Fig. 2 highlight where the baselines fall short. In the first example, RAVE and TokenFlow alter the surrounding scene substantially; for example, the yellow gate pad beside the snowboarders turns into a white column. DNAEdit, FlowDirector, and Wan-Edit preserve the scene appearance but fail to maintain the source trajectory, with the two edited penguins deviating from the original motion. UniEdit-Flow turns the gate pad itself into a penguin, applying the edit outside the region specified by the prompt. In the second example, only FlowResolve and UniEdit-Flow produce the desired green horse, while UniEdit-Flow also introduces green patches on the arena floor. DNAEdit adds flowers to the obstacles and repaints the rider's jacket, RAVE replaces the arena, and TokenFlow tints the entire frame while the horse retains its source colour. FlowDirector also alters the rider and produces a trajectory that deviates substantially from the source. In contrast, FlowResolve performs the intended edit while preserving the surrounding scene and source motion.

These qualitative results highlight the complementary roles of our three components: MTS preserves source motion, EGL localises the edit and suppresses unintended background changes, and TAF strengthens semantic alignment within the target region. Together, they enable FlowResolve to achieve the desired edit while preserving background and motion fidelity.

4.3 Quantitative Results

Table 1 reports results on the full FiVE-Bench. FlowResolve achieves the best performance on M.PSNR, BG.LPIPS, EPE, MFS, and WarpSSIM, while ranking second on L.CLIP-T. Compared with its base sampler Wan-Edit, FlowResolve raises M.PSNR from 26.00 to 27.68 dB and lowers BG.LPIPS by 28.6%. It also reduces EPE by 20.5% and raises MFS from 0.8988 to 0.9082 and L.CLIP-T from 24.81 to 25.47, so the fidelity gains do not come at the expense of the edit.

Full-frame CLIP-T does not follow the same ordering as the source-preservation and motion metrics. RAVE and TokenFlow obtain higher CLIP-T scores, but their substantially lower background

Table 1: Quantitative comparisons on FiVE-Bench. Best results are in **bold** and second-best results are underlined.

	Semantic		Background		Motion fidelity		Temporal consistency	
Method	CLIP-T↑	L.CLIP-T↑	M.PSNR↑	BG.LPIPS↓	EPE↓	MFS↑	WarpSSIM↑	CLIP-F↑
RAVE	**33.89**	25.20	16.30	0.2489	2.145	0.8626	0.9225	0.9706
TokenFlow	33.84	24.82	20.53	0.1572	1.485	0.8953	0.9413	0.9720
UniEdit-Flow	32.32	25.38	24.65	0.1349	2.481	0.8796	0.9505	0.9712
DNAEdit	33.01	25.26	20.90	0.1674	3.414	0.8643	0.9482	**0.9776**
FlowDirector	33.68	**25.59**	23.41	0.0801	2.162	0.8884	0.9530	0.9772
Wan-Edit	32.63	24.81	26.00	0.0695	1.804	0.8988	0.9514	0.9768
FlowResolve	32.81	25.47	**27.68**	**0.0496**	**1.434**	**0.9082**	**0.9544**	0.9756

Table 2: Ablation results for EGL, TAF, and MTS, on the metrics of Table 1. Row 1 is Wan-Edit and row 8 is FlowResolve. The best values for each metric are highlighted in **bold**.

				Semantic		Background		Motion fidelity		Temporal consistency	
#	MTS	EGL	TAF	CLIP-T↑	L.CLIP-T↑	M.PSNR↑	BG.LPIPS↓	EPE↓	MFS↑	WarpSSIM↑	CLIP-F↑
1	—	—	—	32.63	24.81	26.00	0.0695	1.804	0.8988	0.9514	0.9768
2	✓	—	—	32.46	24.55	26.41	0.0642	1.671	0.9043	0.9512	0.9764
3	—	—	✓	**33.12**	**25.72**	26.11	0.0683	2.009	0.9017	0.9510	0.9747
4	✓	—	✓	33.04	25.64	26.05	0.0665	1.841	0.9009	0.9516	0.9749
5	✓	✓	—	32.42	24.58	**28.86**	**0.0461**	**1.302**	**0.9129**	0.9539	0.9767
6	—	✓	✓	32.93	25.58	28.32	0.0481	1.473	0.9092	0.9541	0.9752
7	—	✓	—	32.23	24.37	28.05	0.0479	1.303	0.9098	0.9539	**0.9770**
8	✓	✓	✓	32.81	25.47	27.68	0.0496	1.434	0.9082	**0.9544**	0.9756

PSNR suggests larger deviations from the source content. Thus, a favourable full-frame text score does not necessarily imply source fidelity or motion preservation. Similarly, although FlowDirector achieves the highest L.CLIP-T, it exhibits substantially worse background and motion fidelity than FlowResolve, highlighting the need to evaluate semantics, background, and motion jointly.

4.4 Ablation Study

Table 2 reports all eight combinations of the three components (row 1: Wan-Edit; row 8: FlowResolve). Figs. 3 and 4 compare each ablated variant with the full model on the examples of Fig. 2.

Multi-scale temporal self-similarity guidance (MTS). Fig. 3 illustrates that without MTS the appearance edit still takes place, but the subject departs from the source trajectory (red boxes, w/o MTS row): the two snowboarders are replaced by a single penguin facing the camera, leaving their positions empty, and the horse's head and forehand stay low over the fence. With MTS, the penguins remain where the snowboarders were and the horse follows the upward motion of the source (red boxes, Ours row). Quantitatively, adding MTS consistently reduces EPE, with a maximum improvement of 8.4%, while MTS together with EGL achieves the best overall motion preservation in the table. These results confirm that MTS helps maintain the source trajectory.

Edit-guided localisation (EGL). Fig. 4(a) illustrates that without EGL the editing signal spreads beyond the region described by the prompt (red boxes, w/o row): the yellow gate pad beside the snowboarders, which belongs to the background, grows penguin wings. With EGL, the gate pad is preserved (red boxes, Ours row). Quantitatively, removing EGL from the full model lowers M.PSNR by $1.63\,\mathrm{dB}$ and increases BG.LPIPS, demonstrating its strong effect on background fidelity. It also reduces EPE in the full model, indicating improved motion preservation.

Target attention focusing (TAF). Fig. 4(b) illustrates that without TAF the edit remains within the intended region but is weak (red boxes, w/o row): the horse turns pale olive rather than green. With TAF, the horse is recoloured clearly while the rider and arena are preserved (red boxes, Ours row). Quantitatively, TAF is the only component that consistently improves semantic alignment, increasing L.CLIP-T when added to the other two components. The stronger edit introduces a

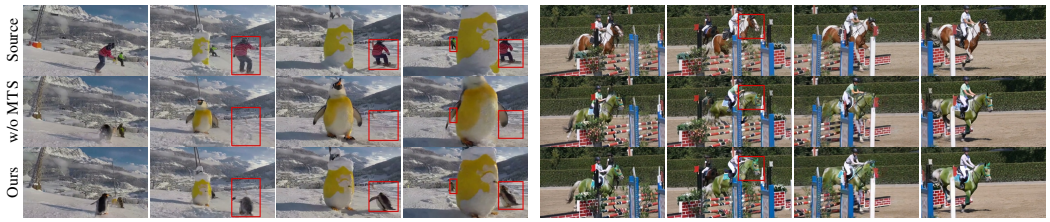


Figure 3: **Ablation study of multi-scale temporal self-similarity guidance (MTS).** Red boxes mark the affected regions.

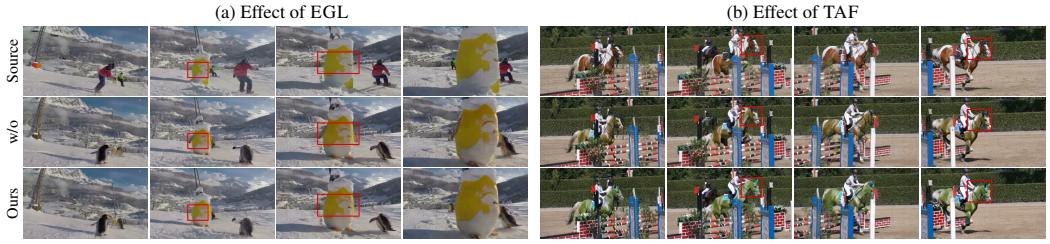


Figure 4: **Ablation study of (a) edit-guided localisation (EGL) and (b) target attention focusing (TAF).** Red boxes mark the affected regions.

modest motion trade-off, but MTS and EGL keep the full model's EPE below the Wan-Edit baseline. Appendix F further discusses how the strength γ controls this trade-off.

Complementarity. The three components address different aspects of the editing problem, and their effects are complementary, as shown in Table 2. MTS and EGL each reduce EPE, and their combination (row 5) achieves the lowest EPE together with the strongest background metrics. Incorporating TAF into MTS +EGL (row 8) further improves L.CLIP-T from 24.58 to 25.47, while increasing M.PSNR by 1.18,dB and EPE by 0.132. Despite this modest reduction in fidelity, the full model maintains better background and motion metrics than Wan-Edit (row 1), indicating that TAF provides a favorable trade-off toward stronger editing quality.

Hyperparameter sweeps. Appendix F sweeps the three settings that decide how much guidance is admitted, and Appendix C the stride set of MTS. The three settings move the same trade-off at different rates. Raising γ improves L.CLIP-T at a small cost in M.PSNR, lowering α gains less for a larger loss, and raising β gains none. This is consistent with the analysis of Section 3.4: α and β set where guidance is admitted, γ sets the edit strength, and only γ controls semantics efficiently.

5 Conclusion

In this paper, we propose FlowResolve, a training-free video editing framework that improves source-motion preservation and instruction following without inversion. We introduce three complementary components to optimise the editing process. MTS introduces a multi-scale temporal constraint to preserve the motion characteristics of the source video while maintaining temporal consistency. EGL decouples editing strength from unintended background changes by correcting the guidance update, improving background preservation during editing. TAF enhances semantic alignment within the target region through attention-based guidance, strengthening the model's ability to follow the editing instruction. Extensive experiments on FiVE-Bench demonstrate that FlowResolve achieves the best background preservation and motion fidelity while remaining competitive in text alignment and temporal consistency. Qualitative results further show that it preserves the source scene and trajectory while performing the desired edits. FlowResolve also has limitations. Its motion constraint assumes that the edited object largely preserves its source motion, which may be less suitable for edits that intentionally change the object's motion; such edits call for relaxing the constraint. In addition, TAF relies on a segmentation mask of the edited object, and MTS adds target-branch forward passes during the first guided steps. Deriving this mask from the model's own edit signal, as EGL does for the guidance gap, is a promising next step.

USE OF AI

In this work, we used generative AI tools to improve the readability and clarity of the manuscript, including polishing and rephrasing existing text. We did not use generative AI tools to generate synthetic data, develop or implement the method, design or provide feedback on the methodology or experiments, formulate mathematical claims, assist with proofs, interpret experimental results, translate content, or clean and reformat datasets.

REPRODUCIBILITY STATEMENT

The method is training-free and introduces no additional parameters. Reproduction therefore requires the public Wan2.1-T2V-1.3B checkpoint and the public FiVE-Bench videos and prompt pairs. Section 3 and Appendix B specify the guidance decomposition and its localisation (Eqs. 8–9), the descriptor construction (Eq. 11), the multi-scale temporal objective (Eq. 6), the mask-to-token mapping (Eq. 12), and the attention bias (Eq. 10). Algorithm 1 gives the complete editing procedure. The appendix provides the additional implementation details, ablations, and evaluation results, including the settings used for the reported experiments. An anonymised source code package implementing the guidance loop and attention bias is provided in the supplementary material.

ETHICS STATEMENT

This work develops a video editing method using publicly released models and benchmark videos. As with other video editing methods, the technology could be used to produce misleading or deceptive footage of real people or events. Our method is designed to preserve the motion of a source video while modifying its visual content, which may improve the temporal plausibility of such edits. We therefore acknowledge the potential for misuse and release the implementation for research purposes only. Our experiments use publicly available benchmark data and do not involve human subjects or the collection of new personal data.

REFERENCES

Lingling Cai, Kang Zhao, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Kejie Huang. Dfvedit: Conditional delta flow vector for zero-shot video editing. *arXiv preprint arXiv:2506.20967*, 2025.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22560–22570, 2023.

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

Ze Chen, Lan Chen, Yuanhang Li, and Qi Mao. Flowanchor: Stabilizing the editing signal for inversion-free video editing. *arXiv preprint arXiv:2604.22586*, 2026.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: Optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633. PMLR, 2024.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshops: DGMs_Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *International Conference on Learning Representations*, 2024.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Jackson Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ArU2CeB7Tm.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6507–6516, 2024.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pp. 18–35. Springer, 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*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 179–195. Springer, 2018. doi: 10.1007/978-3-030-01267-0_11.

Guangzhao Li, Yanming Yang, Chenxi Song, Xiaohong Liu, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7805–7815, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16672–16681, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15932–15942, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 402–419. Springer, 2020.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. In *Advances in Neural Information Processing Systems*, volume 37, pp. 96541–96565, 2024.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 64044–64058. PMLR, 2025a.

Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2589–2598, 2025b.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 137937–137955, 2025.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023.

Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *International Conference on Learning Representations*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025b.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

A FULL ALGORITHM

Algorithm 1 Video editing with FlowResolve

Require: source video, prompts $\mathcal{P}, \mathcal{P}^*$, target object phrase, timesteps $t_N > \cdots > t_1 > t_0 = 0$

$X^{\mathrm{src}} \leftarrow \mathrm{VAE.encode}(\text{source video})$; $\quad Z^{\mathrm{edit}} \leftarrow X^{\mathrm{src}}$
$m \leftarrow$ token mask from SAM 2 masklet ▷ Eq. 12
$\mathcal{T} \leftarrow$ target phrase sub-token indices in $\mathcal{P}^*$
for $i = N, N-1, \ldots, 1$ **do**
  $t \leftarrow t_i$; $\epsilon_t \sim \mathcal{N}(0, I)$; form $Z_t^{\mathrm{src}}, Z_t^{\mathrm{tar}}$ ▷ Eq. 4
  if $i > N - G$ **then** ▷ first G steps
    $\hat{Z}_0^{\mathrm{src}} \leftarrow Z_t^{\mathrm{src}} - t\,\tilde{V}(Z_t^{\mathrm{src}}, t, \mathcal{P}, \omega_{\mathrm{src}})$; $f^{\mathrm{src}} \leftarrow f(\hat{Z}_0^{\mathrm{src}})$ ▷ Eq. 11
    for J Adam steps **do**
      $\hat{Z}_0^{\mathrm{tar}} \leftarrow Z_t^{\mathrm{tar}} - t\,\tilde{V}(Z_t^{\mathrm{tar}}, t, \mathcal{P}^*, \omega_{\mathrm{tgt}})$ ▷ no gradient through $\tilde{V}$
      $Z_t^{\mathrm{tar}} \leftarrow Z_t^{\mathrm{tar}} - \eta\,\mathrm{Adam}\big(\nabla_{Z_t^{\mathrm{tar}}} \lambda_{\mathrm{MS}} \mathcal{L}_{\mathrm{MS}}(f^{\mathrm{src}}, f(\hat{Z}_0^{\mathrm{tar}}))\big)$ ▷ Eqs. 6–7
    end for
    $Z^{\mathrm{edit}} \leftarrow Z_t^{\mathrm{tar}} - Z_t^{\mathrm{src}} + X^{\mathrm{src}}$
  end if
  $V_t^{\mathrm{src,c}}, V_t^{\mathrm{src,u}} \leftarrow V(Z_t^{\mathrm{src}}, t, \mathcal{P}),\ V(Z_t^{\mathrm{src}}, t, \varnothing)$
  $V_t^{\mathrm{tar,c}} \leftarrow V(Z_t^{\mathrm{tar}}, t, \mathcal{P}^*)$ **with** TAF bias ▷ if t in the attention window, Eq. 10
  $V_t^{\mathrm{tar,u}} \leftarrow V(Z_t^{\mathrm{tar}}, t, \varnothing)$
  $T_1, T_2, T_3 \leftarrow$ Eq. 8; $\quad w \leftarrow \beta + (1-\beta)\min(\mu / q_\alpha(\mu), 1)$ ▷ Eq. 9
  $Z^{\mathrm{edit}} \leftarrow Z^{\mathrm{edit}} + (t_{i-1} - t_i)\,(T_1 + T_2 + w \odot T_3)$ ▷ Eq. 5
end for
return $\mathrm{VAE.decode}(Z^{\mathrm{edit}})$

B IMPLEMENTATION DETAILS OF MTS AND EGL

Motion descriptors. MTS builds its descriptors from the clean-sample prediction $\hat{Z}_0 \in \mathbb{R}^{C \times F \times H \times W}$. We first remove the temporal mean at each spatial location, which cancels static appearance, and then describe every pooled position by its normalised local neighbourhood:

$$R_{:,n} = \hat{Z}_{0,:,n} - \tfrac{1}{F}\textstyle\sum_\tau \hat{Z}_{0,:,\tau}, \qquad f_n = \mathrm{Normalize}\Big(\mathrm{Unfold}_\kappa\big(\mathrm{AvgPool}_\rho(R_{:,n})\big)\Big) \in \mathbb{R}^{D \times P}. \quad (11)$$

Here $\mathrm{AvgPool}_\rho$ is spatial average pooling with factor ρ, which suppresses high-frequency noise in $\hat{Z}_0$ and reduces the number of positions; Unfold_κ gathers the $\kappa \times \kappa$ neighbourhood of each pooled position, so that every descriptor has dimension $D = C\kappa^2$ and P is the number of pooled positions; and $\mathrm{Normalize}$ scales each descriptor to unit ℓ_2 norm, so the inner products in Eq. 6 are cosine similarities. The source descriptors are computed in the same way from the source-branch prediction $\hat{Z}_0^{\mathrm{src}}$ at each guided step; Appendix G compares this choice with the clean latent X^{src}. In Eq. 6, the $1/(F-s)$ factor averages each stride over its own frame pairs, so long strides, which contain fewer pairs, are not outweighed by short ones.

Latent update. Eq. 7 is applied at each of the first G sampling steps with J Adam updates (Section 4.1). In Eq. 7, the velocity $\tilde{V}$ in $\hat{Z}_0^{\mathrm{tar}}$ is treated as a constant when differentiating. The source descriptors do not depend on Z_t^{tar}, so no gradient flows into them.

Localisation map. The magnitude of T_2 varies across timesteps and videos, so a fixed threshold would not transfer between them. Normalising μ by its α-quantile, taken over all latent positions of the current step, makes w in Eq. 9 depend on the spatial shape of the edit differential rather than on its scale. The floor β accounts for non-rigid edits: the region the subject will occupy then differs from the region it occupies now, so T_2 is weak exactly where new content must appear, and β keeps part of the target guidance there. The map w is shared across all channels of the latent and is computed from T_2 with gradients detached, since it serves only as a spatial weighting.

C STRIDE SET OF THE MULTI-SCALE OBJECTIVE

Table 3 evaluates different stride sets and weights with all other components fixed at the operating point. Using only stride 1, which constrains neighbouring frames, barely improves endpoint error over removing MTS. In contrast, including non-adjacent strides consistently reduces endpoint error. Among the evaluated configurations, the full set $\mathcal{S} = \{1, 2, 3\}$ with weights $\lambda = (1, 2, 4)$ achieves the lowest endpoint error (1.434) and the highest MFS among the stride settings (0.9082). We therefore use this configuration throughout the paper.

Table 3: Stride set and weights of MTS on all 419 pairs, with everything else at the operating point ($*$). The first row repeats row 6 of Table 2, i.e., the full model without MTS. The best value in each column among the stride settings is in **bold**.

	CLIP-T↑	L.CLIP-T↑	M.PSNR↑	BG.LPIPS↓	EPE↓	MFS↑
no MTS	32.93	25.58	28.32	0.0481	1.473	0.9092
$\mathcal{S} = \{1\}$	**32.84**	**25.47**	27.69	0.0498	1.472	0.9067
$\mathcal{S} = \{2\}$	32.81	25.46	27.77	0.0494	1.459	0.9080
$\mathcal{S} = \{3\}$	32.81	25.44	**27.83**	**0.0490**	1.448	**0.9082**
$\mathcal{S} = \{1, 2\}$, equal weights	32.83	25.45	27.64	0.0497	1.453	0.9070
$\mathcal{S} = \{1, 3\}$, equal weights	32.80	25.42	27.68	0.0495	1.445	0.9071
$\mathcal{S} = \{1, 2, 3\}$, equal weights	32.80	25.41	27.62	0.0497	1.441	0.9074
$\mathcal{S} = \{1, 2, 3\}$, $\lambda = (1, 2, 4)^*$	32.81	**25.47**	27.68	0.0496	**1.434**	**0.9082**

D MASK-TO-TOKEN MAPPING

We obtain a temporally consistent mask M of the edited object with SAM 2 (Ravi et al., 2025). The mask is initialised from a box produced by an open-vocabulary detector prompted with the target noun phrase and then propagated through the video as a masklet. We map the resulting mask to the DiT token grid following the temporal chunking of the causal VAE: latent frame 1 is encoded from pixel frame 1 alone, while latent frame $i > 1$ aggregates pixel frames $\{4i - 6, \ldots, 4i - 3\}$. We first take the union within each chunk $\mathcal{C}(i)$ and then apply inclusive pooling to the patch grid,

$$M_i^{\text{lat}} = \max_{f \in \mathcal{C}(i)} M_f, \qquad m = \text{MaxPool3D}\big(M^{\text{lat}}\big) \in \{0, 1\}^L, \tag{12}$$

where L is the number of DiT query tokens. Max pooling marks a patch as foreground whenever any of its pixels belongs to the mask, preserving thin structures under the $16\times$ spatial downsampling.

E COMPLETE RESULTS

Table 4 reports the metrics omitted from Table 1: warping error and whole-frame reconstruction quality. All numbers are computed with the same evaluation protocol on the same 419 pairs.

Table 4: Metrics omitted from Table 1 on FiVE-Bench. Best in **bold**, second best <u>underlined</u>.

	Temporal	Whole-frame reconstruction			
Method	Warp-Err↓	PSNR↑	LPIPS↓	MSE↓	SSIM↑
RAVE	7.377	15.18	0.2932	0.0333	0.5813
TokenFlow	4.524	19.54	0.1894	0.0125	0.6657
UniEdit-Flow	4.193	22.48	0.1696	0.0073	0.7562
DNAEdit	4.836	19.22	0.2076	0.0137	0.6568
FlowDirector	4.117	20.29	0.1251	0.0113	0.7836
Wan-Edit	<u>4.024</u>	<u>23.30</u>	<u>0.0983</u>	<u>0.0056</u>	<u>0.8073</u>
FlowResolve	**3.813**	**24.39**	**0.0771**	**0.0044**	**0.8345**

FlowResolve achieves the best value in every column of Table 4. Compared with Wan-Edit, it raises whole-frame PSNR from 23.30 to 24.39 dB and lowers LPIPS from 0.0983 to 0.0771.

F SENSITIVITY TO γ, α, AND β

Table 5: The three settings that determine how much guidance is admitted, swept one at a time on FiVE-Bench with everything else at the operating point ($*$), i.e., the configuration of Tables 1 and 2. The best value in each column within each block is in **bold**.

	CLIP-T↑	L.CLIP-T↑	M.PSNR↑	BG.LPIPS↓	EPE↓	MFS↑
TAF strength $\gamma = 1.5$	32.48	24.81	**27.96**	**0.0484**	**1.332**	**0.9105**
$\gamma = 2.0$	32.62	25.06	27.88	0.0487	1.359	0.9095
$\gamma = 2.5$	32.72	25.24	27.81	0.0490	1.380	0.9090
$\gamma = 3.0$	32.77	25.38	27.73	0.0494	1.405	0.9080
$\gamma = 3.5^*$	32.81	25.47	27.68	0.0496	1.434	0.9082
$\gamma = 4.0$	32.84	25.54	27.61	0.0499	1.452	0.9076
$\gamma = 4.5$	32.87	25.61	27.56	0.0501	1.479	0.9069
$\gamma = 5.0$	**32.89**	**25.66**	27.50	0.0503	1.509	0.9063
EGL quantile $\alpha = 0.90$	**32.89**	**25.54**	27.03	0.0548	1.587	0.9051
$\alpha = 0.95$	32.86	25.50	27.37	0.0519	1.506	0.9073
$\alpha = 0.98^*$	32.81	25.47	27.68	0.0496	1.434	0.9082
$\alpha = 0.99$	32.75	25.33	**27.89**	**0.0481**	**1.393**	**0.9086**
EGL floor $\beta = 0.05$	32.78	25.40	**27.77**	**0.0487**	**1.415**	**0.9086**
$\beta = 0.10^*$	32.81	**25.47**	27.68	0.0496	1.434	0.9082
$\beta = 0.15$	32.82	25.42	27.62	0.0501	1.457	0.9076
$\beta = 0.20$	32.82	25.44	27.54	0.0509	1.468	0.9079
$\beta = 0.25$	**32.83**	25.45	27.46	0.0516	1.493	0.9075

Table 5 varies γ, α, and β one at a time, keeping all other settings at the operating point. The trade-off becomes increasingly imbalanced as γ increases. Each 0.5 increment in γ results in a relatively consistent fidelity degradation, with M.PSNR decreasing by about 0.05–0.08,dB and endpoint error increasing by 0.02–0.03, while the semantic improvement becomes progressively smaller. Specifically, the L.CLIP-T gain per step decreases from 0.25 at $\gamma = 1.5$ to below 0.1 for $\gamma > 3.5$. On average, a 0.5 increase in γ below the operating point yields a 0.17 gain in L.CLIP-T, whereas the average gain above the operating point is only 0.06. This diminishing return indicates that increasing γ beyond the operating point provides limited additional semantic improvement while incurring a comparable loss in fidelity. Even at $\gamma = 1.5$, where L.CLIP-T matches that of Wan-Edit (24.81), FlowResolve maintains its fidelity advantage, with a 1.96,dB higher M.PSNR and a 26% lower endpoint error.

Lowering α from 0.98 to 0.90 increases L.CLIP-T only slightly, from 25.47 to 25.54, whereas M.PSNR drops from 27.68 to 27.03 and endpoint error rises from 1.434 to 1.587. Raising β above 0.10 does not improve L.CLIP-T and consistently reduces M.PSNR. Based on these results, we use $\gamma = 3.5$, $\alpha = 0.98$, and $\beta = 0.10$ as the operating point.

G SOURCE REFERENCE

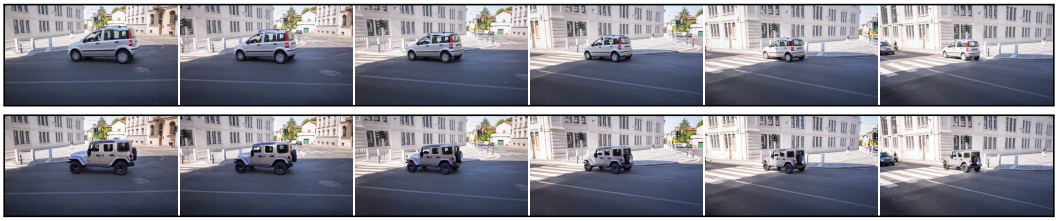
The source descriptors can be computed from either of two latents. FlowResolve uses the source-branch prediction $\hat{Z}_0^{\mathrm{src}}$, obtained from the noisy source latent Z_t^{src} at the current step in the same way as $\hat{Z}_0^{\mathrm{tar}}$ is obtained from Z_t^{tar}. The alternative is the clean latent X^{src}, the encoded source video, which is sharp and fixed across steps. A one-step prediction is blurred at high noise levels, which alters its temporal self-similarity, so a sharp X^{src} and a blurred $\hat{Z}_0^{\mathrm{tar}}$ can yield different self-similarities even when their motion agrees. $\hat{Z}_0^{\mathrm{src}}$ carries the same blur as $\hat{Z}_0^{\mathrm{tar}}$, so the comparison in Eq. 6 is between predictions of equal sharpness. Replacing it with the clean latent (Table 6) lowers M.PSNR from 27.68 to 27.48, raises endpoint error from 1.434 to 1.538, and lowers MFS from 0.9082 to 0.9058, while CLIP-T and L.CLIP-T rise by only 0.03 and 0.02.

Table 6: Source-reference ablation on FiVE-Bench. The main experiment uses the predicted source latent $\hat{Z}_0^{\mathrm{src}}$, while the alternative uses the clean source latent X^{src}. The better value in each column is in **bold**.

	Semantic		Background		Motion fidelity		Temporal consistency	
Method	CLIP-T↑	L.CLIP-T↑	M.PSNR↑	BG.LPIPS↓	EPE↓	MFS↑	WarpSSIM↑	CLIP-F↑
FlowResolve ($\hat{Z}_0^{\mathrm{src}}$)	32.81	25.47	**27.68**	**0.0496**	**1.434**	**0.9082**	**0.9544**	**0.9756**
Clean latent (X^{src})	**32.84**	**25.49**	27.48	0.0517	1.538	0.9058	0.9541	0.9751

H MORE QUALITATIVE RESULTS

Figs. 5–7 show further results of FlowResolve on FiVE-Bench, covering object replacement, colour editing, and object addition. For each example, the top row shows six uniformly sampled frames of the source video and the bottom row the corresponding frames of the edited video. In the prompt below each example, replaced words are struck through and new words are highlighted; colour words are shown in the colour they name.



A silver ~~car~~ jeep is making a left turn at an urban intersection, surrounded by modern buildings and a pedestrian walking on the sidewalk.



A ~~soccer ball~~ basketball is rolling quickly across a grassy area, passing between two tree trunks in a garden with a wire fence in the background.



A purple city bus is moving steadily along a street lined with trees and construction barriers.



Two ~~white~~ orange ibises are walking along the sandy shore, pecking at the ground as gentle waves roll in the background.

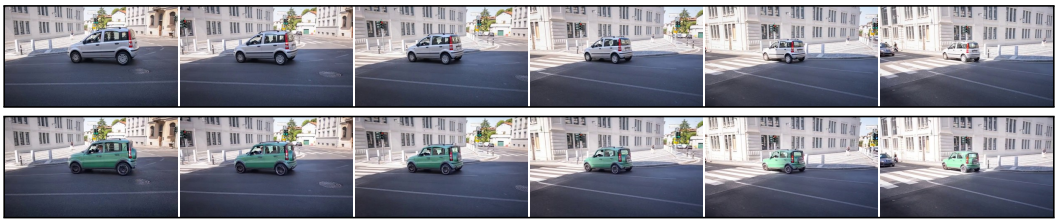
Figure 5: More qualitative results of FlowResolve on FiVE-Bench. Top: source video; bottom: edited video.



A pink paraglider is soaring gracefully over a lush, green landscape with a quaint village nestled among the hills.



A ~~black~~ pink car is performing a burnout, creating thick clouds of smoke on a track surrounded by a crowd of spectators.

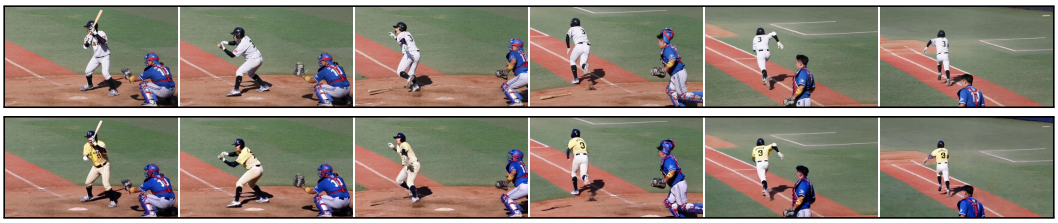


A ~~silver~~ green car is making a left turn at an urban intersection, surrounded by modern buildings and a pedestrian walking on the sidewalk.

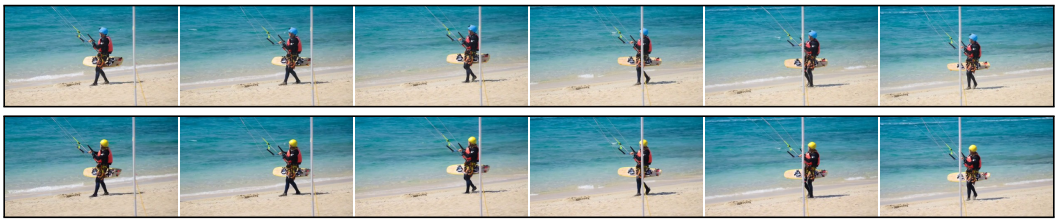


A blue horse with a rider is gracefully jumping over a series of obstacles in an outdoor equestrian arena.

Figure 6: More qualitative results of FlowResolve on FiVE-Bench (continued). Top: source video; bottom: edited video.



A baseball player in yellow in a striped uniform is preparing to bunt the ball on a sunny day at the baseball field.



A person wearing a ~~blue~~ yellow helmet and a red life vest is walking along a sandy beach while holding a kiteboard and controlling the kite lines.



A purple soccer ball is rolling quickly across a grassy area, passing between two tree trunks in a garden with a wire fence in the background.



A dog wearing sunglasses is wagging its tail excitedly while sitting on a sandy beach with waves crashing in the background.

Figure 7: More qualitative results of FlowResolve on FiVE-Bench (continued). Top: source video; bottom: edited video.