SOUND SPARKS MOTION: AUDIO AND TEXT TUNING FOR VIDEO MOTION EDITING

Anonymous authors
Paper under double-blind review



Figure 1: Given a source video and an edit prompt, our method tunes the audio and text latent as conditioning parameters, such that the desired motion edit is realized in the edited video.

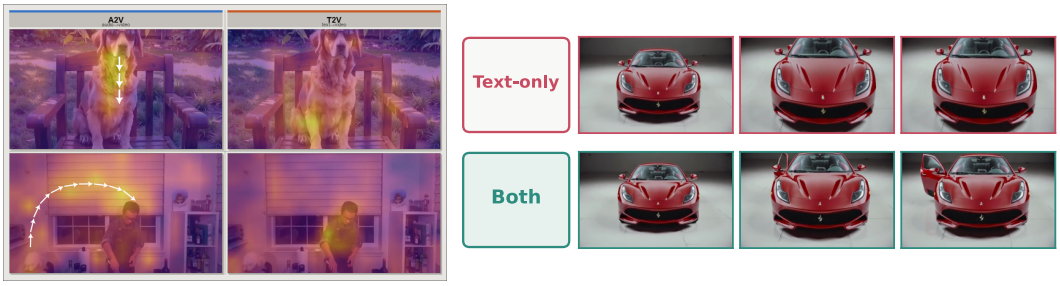
ABSTRACT

Motion-centric video editing remains difficult for large generative video models, which often respond well to appearance changes but struggle to produce specific, localized actions or state transitions in an existing clip. We introduce Sound Sparks Motion, a training-free framework that enables motion editing in an audio-visual video generation model by tuning its internal multimodal conditioning signals at test time. Rather than modifying model weights, our method tunes two lightweight variables: the text conditioning and the audio latent. We find that tuning the text alone is often insufficient, as audio and motion are highly entangled in the underlying model. By also optimizing the audio latent, we can more effectively steer the desired motion, while the resulting audio can simply be discarded after generation. Since there is no direct way to evaluate temporal alignment between text and motion, we guide the tuning process using a vision–language model that provides feedback indicating whether the intended motion appears in the generated video. This simple supervision yields an effective semantic objective for motion editing, while regularization and perceptual-temporal constraints help preserve content and visual quality. Beyond per-video tuning, we show that the learned latent controls are transferable across videos, suggesting that they capture reusable motion-edit directions rather than overfitting to a single example. Our results highlight multimodal conditioning tuning, particularly through the audio pathway, as a promising direction for motion-aware video editing, and suggest that test-time tuning can serve as a lightweight probing mechanism that helps reveal latent motion controls embedded in the model's multimodal conditioning.

1 INTRODUCTION

Video generation models usually do not deliver fully satisfactory results in a single pass. In practice, creating the desired outputs often involves an iterative process in which both visual appearance and motion dynamics are gradually refined. Users frequently need to adjust temporal consistency, object movement, and scene composition across multiple iterations to achieve coherent and realistic results.

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.



(a) (b)

Figure 2: **(a)** A visualization of average attention heat maps for tuned audio and text components, plotted on top of the first frame of the video. **"A2V"** shows "audio-to-video" cross attention and **"T2V"** shows "text-to-video" cross attention. Notice how text focuses more on the semantics and objects, while audio is focusing more on motion trajectory (shown with white arrows). Here the motion is *yawning* (top) and *jumping* (bottom) for the dogs. For an extended view of attention maps on other scenarios refer to Supplementary Fig. 9. **(b)** A comparison between text-only residual tuning and tuning text and audio at the same time on *"a red car door opens."* shows that in some cases, the model fails to capture all the required information to do hard motions while keeping the main subjects in the scene unchanged in text-only mode.

Most current video generation systems are primarily controlled through text-based prompts. While text provides a convenient and flexible interface, it is often insufficient for precisely specifying motion, especially in audio-visual generative models. In such models, motion is tightly entangled with both visual content and audio cues, making it difficult to isolate and control through language alone. As a result, relying solely on text limits the ability to guide motion effectively.

Given an existing source video—whether captured or generated—performing targeted edits, such as modifying an existing motion or introducing a new action, remains challenging. Recent video editing approaches often condition a pretrained generative model on the source video together with a text instruction, preserving the source content while re-synthesizing the desired motion. However, text conditioning alone can struggle to drive substantial motion changes: the model may fail to realize the requested action, produce only weak or incomplete motion, or remain dominated by the dynamics of the source video (see Fig. 1).

In this work, we revisit the role of multimodal conditioning in motion generation and show that audio can provide an additional signal for enabling motion. We observe that motion and audio are strongly coupled within the internal representations of audio-visual generative models, even for actions without an obvious natural sound (see Fig. 2a). Building on this observation, we propose optimizing the audio conditioning at test time to steer the model toward the desired motion. We additionally introduce a lightweight residual adjustment to the text-conditioning space, further aligning the generation with the intended edit without modifying the prompt itself (Fig. 2b).

Our approach requires no modification of the model weights and instead exploits motion-relevant signals already present in the multimodal conditioning space. This provides an additional source of control when text instructions alone are insufficient. We evaluate this principle across existing audio-visual generative models with different architectures and editing mechanisms, showing that the same conditioning-based strategy can improve motion control across different model designs.

To guide this process, we employ a vision-language model that provides binary feedback indicating whether the intended motion or state change appears in the generated video. This feedback defines a simple semantic objective for test-time tuning, while regularization and perceptual-temporal constraints help preserve content and visual quality. Importantly, the tuning operates over a small set of conditioning variables, leaving all model parameters unchanged.

Beyond per-video tuning, we observe that the learned conditioning adjustments can transfer across videos, suggesting that they capture reusable motion-edit directions rather than overfitting to a single instance. Taken together, these findings suggest that test-time tuning can serve as a lightweight

probing mechanism that helps reveal latent motion controls embedded in the model's multimodal conditioning.

2 Related Work

Text-guided video editing. Early video-editing methods extended text-to-image diffusion models to video by adapting inversion, attention control, and feature propagation for temporal consistency (Geyer et al., 2023; Cong et al., 2023; Qi et al., 2023; Liu et al., 2023; Wang et al., 2025). With the emergence of specialized video generators (Zheng et al., 2024; Wan et al., 2025; Kong et al., 2024; HaCohen et al., 2026), subsequent methods increasingly built on native video backbones. These include tuning-free inference-time approaches (Bai et al., 2024; Ku et al., 2024; Xu et al., 2025) and methods based on finetuning (Molad et al., 2023), richer conditioning architectures and adapters (Jiang et al., 2025; Lin et al., 2026; Ye et al., 2025), lightweight LoRA (Hu et al., 2021) for extra conditioning (Ben-Yosef et al., 2026), first-frame edit propagation (Polaczek et al., 2025; Gao et al., 2026), and fully trained instruction-guided editors (Qin et al., 2023; DecartAI, 2025). Despite this progress, current methods remain more effective for appearance, style, and semantic changes than for motion-centric edits, which require temporally grounded instructions and high-quality source-target video pairs that modify motion while preserving content. We address this gap by showing that the audio pathway of a joint audio-video model (HaCohen et al., 2026; MiniMax, 2026) can serve as an internal source of temporal control.

Motion control and editing. Motion in videos arises from camera and object movement. Camera-motion methods condition generation on explicit poses or trajectories (Bahmani et al., 2025a; He et al., 2024; 2025; Bahmani et al., 2025b; Ren et al., 2025), or re-render an existing video under a new camera trajectory (Li et al., 2026; Bai et al., 2025a). Object-motion methods transfer motion from reference videos by manipulating attention (Pondaven et al., 2025) or positional encodings (Gokmen et al., 2025), while others use explicit trajectories (Geng et al., 2024; Namekata et al., 2025; Ma et al., 2023; Namekata et al., 2026), flow fields (Burgert et al., 2025b), or physical forces (Gillman et al., 2025; 2026). Related work also studies trajectory-guided inbetweening between endpoint frames (Wang et al., 2024a; Tanveer et al., 2025) and modifies existing-video dynamics using trajectories (Mou et al., 2024; Burgert et al., 2025a) or anchor-frame paths (Phung et al., 2026; Lee et al., 2025). Our method instead requires neither external motion annotations nor user-specified controls. Unlike audio-conditioned approaches such as Just-Dub-It (Chen et al., 2026), which uses target speech to drive synchronized facial motion for dubbing, we use the internal audio pathway of a joint audio-video model as a latent source of temporal control for general text-guided motion editing.

Reward/VLM-guided prompt alignment. Reward models and VLM feedback have been used to align diffusion models with human preferences and high-level semantic objectives. Image-generation work uses reward modeling and preference optimization (Xu et al., 2023; Prabhudesai et al., 2023; Wallace et al., 2023); EditReward (Wu et al., 2025) applies reward optimization to instruction-guided image editing, FireEdit (Zhou et al., 2025) uses region-aware VLM feedback for fine-grained image editing, and ParetoSlider (Golan et al., 2026) studies post-training for continuous appearance control over competing reward objectives. Video-generation methods similarly use pretrained reward models, human-feedback critics, and preference-based post-training to improve prompt faithfulness and generation quality (Prabhudesai et al., 2024; Wang et al., 2024b; Liu et al., 2025). More closely related, VIVA (Cong et al., 2025) performs VLM-guided reward optimization for instruction-based video editing, while Diffusion-DRF (Wang et al., 2026) shows that frozen off-the-shelf VLMs can act as critics for video-diffusion optimization. In contrast, we neither train a separate reward model nor post-train the generator; a frozen VLM guides per-video optimization of audio latents only at test time.

3 Method

Given a source video $\mathbf{x} = \{x_t\}_{t=1}^{T}$, where x_t denotes the RGB frame at time t, and a natural-language edit instruction p, our goal is to generate an edited video $\hat{\mathbf{x}} = \{\hat{x}_t\}_{t=1}^{T}$ that realizes the requested edit while preserving the identity, scene layout, and visual content of the source. We focus

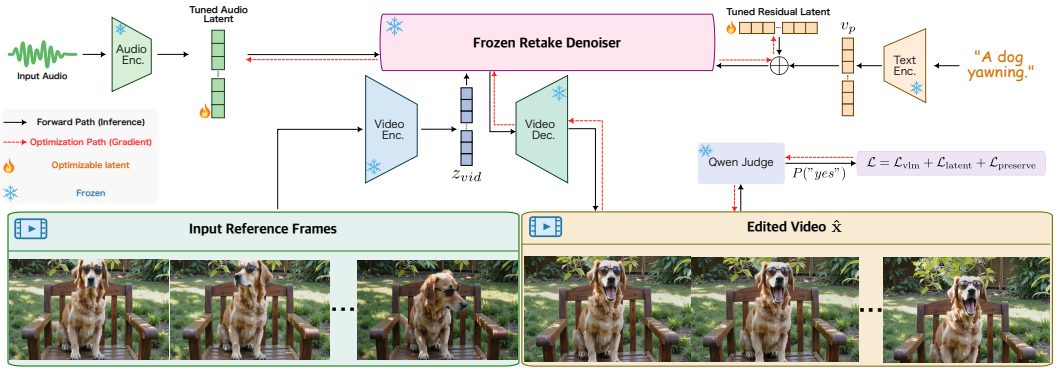


Figure 3: **Overview of Sound Sparks Motion.** Given source audio, reference video frames, and an edit prompt, our method keeps LTX Retake, Text Encoder, Qwen, and the VAE modules frozen, and optimizes only the audio latent and a residual text-conditioning vector. The edited video is temporally sampled and scored by Qwen2.5-VL through a binary motion question, producing a test-time loss that updates only the learnable latents while preserving the frozen generator, making it capable of complex motion edits.

on *motion-centric* edits, in which the requested change is primarily temporal rather than appearance-based, such as making a subject jump, yawn, bloom, or perform a specified action.

We assume access to a pretrained joint audio-video generation model $\mathcal{G}_\theta$ with frozen parameters θ. We instantiate our framework using LTX-2 Retake (HaCohen et al., 2026) and MiniMax-H3(MiniMax, 2026). Although these models support source-conditioned editing through different interfaces, prompt-only conditioning can fail to realize the requested motion, express it only weakly, or reproduce the original dynamics. We therefore tune a lightweight set of internal multimodal conditioning variables at test time while keeping all model weights fixed. The tuning is guided by a semantic motion objective, together with regularization and perceptual-temporal regularizations that preserve the source content. The following subsections describe the tunable conditioning variables, VLM-based motion supervision, and the resulting test-time objective.

3.1 Method Overview

As illustrated in Fig. 3, Sound Sparks Motion turns the frozen audio-video generator into a motion-aware editor by searching its internal conditioning space for controls that realize the requested action. We support both prefix-conditioned and full-video editing. Specifically, let $\mathbf{x}_{i:j} = \{x_t\}_{t=i}^{j}$ denote the editable interval, where $1 \le i \le j \le T$ and $(i, j) = (1, T)$ denotes full-video editing.

We keep the source-video conditioning fixed and tune two lightweight variables: an audio-conditioning representation $\boldsymbol{\alpha}$, initialized from the source-derived representation $\boldsymbol{\alpha}_0$, and a residual text-conditioning variable $\Delta\mathbf{v}$, initialized to zero. Let $\mathbf{v}_p$ denote the original prompt-conditioning representation and $\boldsymbol{\xi}$ the fixed stochastic state used for generation. The default and tuned conditioning states, together with their corresponding outputs, are

$$\begin{aligned} \mathbf{c}_0 &= (\boldsymbol{\alpha}_0, \mathbf{v}_p)\,, & \hat{\mathbf{x}}_{\text{base}} &= \mathcal{G}_\theta\left(\mathbf{x}_{i:j}; \mathbf{c}_0, \boldsymbol{\xi}\right), \\ \mathbf{c} &= (\boldsymbol{\alpha}, \mathbf{v}_p + \Delta\mathbf{v})\,, & \hat{\mathbf{x}} &= \mathcal{G}_\theta\left(\mathbf{x}_{i:j}; \mathbf{c}, \boldsymbol{\xi}\right). \end{aligned} \tag{1}$$

Here, $\hat{\mathbf{x}}_{\text{base}}$ is the untuned, prompt-conditioned output, whereas $\hat{\mathbf{x}}$ is generated using the tuned conditioning state. Only $\boldsymbol{\alpha}$ and $\Delta\mathbf{v}$ are tuned; the generator, conditioning encoders, VLM critic, and all other model parameters remain frozen.

We use different parameterizations for the two conditioning branches. The text branch uses a residual so that the tuned representation remains explicitly anchored to the original prompt semantics. In contrast, the audio representation is already a compact, temporally structured conditioning variable tied to the source clip, so we tune it directly while regularizing it toward $\boldsymbol{\alpha}_0$.



(a) (b)

Figure 4: **(a) Effect of supervision method.** CLIP Radford et al. (2021) guidance primarily optimizes frame-level semantic appearance rather than temporal action, leading to a mostly static yawning state. XCLIP Ni et al. (2022) provides limited temporal awareness, but still introduces global appearance drift and localized artifacts, highlighted by the red box. Qwen-based supervision better captures whether the requested motion unfolds over time, producing a more coherent yawning edit. **(b) Effect of regularization.** Without regularization, VLM-guided tuning produces artifacts and abrupt motion, with representative failures highlighted in red. Latent regularization reduces drift, LPIPS improves frame-level fidelity, and temporal regularization improves motion smoothness. Combining all terms yields the most coherent and stable result.

3.2 Backbone Instantiations

We instantiate $\mathcal{G}_\theta$ using two pretrained joint audio-video models: LTX-2 Retake (HaCohen et al., 2026) and MiniMax-H3 (MiniMax, 2026). Despite their different architectures and editing interfaces, both exhibit the same practical limitation: prompt-only conditioning may omit the requested action or express it only weakly. At the same time, both models expose lightweight audio- and text-conditioning representations that influence video generation and can be tuned while keeping all model weights frozen.

LTX-2 performs source-conditioned editing through its Retake interface, whereas MiniMax-H3 performs prompt-guided video-to-video editing through its omni architecture. In both instantiations, we keep the source-video conditioning and model parameters fixed and tune the corresponding audio representation and residual text-conditioning variable defined in equation 1.

Audio-Conditioning Latent Tuning. The audio pathway provides a temporally structured conditioning signal aligned with video generation. In LTX-2, the attention visualizations in Fig. 2a further show that this pathway correlates spatially and temporally with localized motion regions, motivating its use as a latent handle for motion control. We initialize $\boldsymbol{\alpha}$ from the source-derived representation $\boldsymbol{\alpha}_0$ and tune it to strengthen the requested motion while regularizing it toward its initialization. The audio representation serves only as a control signal; the resulting audio need not be retained after generation.

Residual Text-Conditioning Tuning. The frozen text encoder maps the edit prompt p to a video-conditioning representation $\mathbf{v}_p$. Rather than rewriting the prompt, we tune a zero-initialized residual $\Delta\mathbf{v}$ and condition the generator on $\widetilde{\mathbf{v}}_p = \mathbf{v}_p + \Delta\mathbf{v}$. This residual adjusts the semantic conditioning while remaining anchored to the original prompt representation. Together, the audio representation and text residual provide complementary temporal and semantic control.

3.3 VLM-Guided Motion Supervision

The desired motion is specified semantically by the edit prompt and cannot be captured reliably by pixel-level or framewise perceptual objectives alone. Frame-level image–text models such as CLIP (Radford et al., 2021) also lack the temporal context needed to distinguish an action from a static appearance. We therefore use the frozen Qwen2.5-VL model (Bai et al., 2025b) as a differentiable edit critic. Fig. 4 compares this supervision with CLIP and XCLIP.

For an edit prompt p, we ask Qwen whether the generated video clearly shows the action or state change described by p. Let ℓ_{yes} and ℓ_{no} denote the logits assigned to the answer tokens `yes` and `no`, respectively. We define

$$P_{\text{yes}} = \frac{\exp(\ell_{\text{yes}})}{\exp(\ell_{\text{yes}}) + \exp(\ell_{\text{no}})}, \qquad \mathcal{L}_{\text{vlm}} = -\log\left(P_{\text{yes}} + \epsilon\right), \tag{2}$$

where ϵ is a small constant for numerical stability. Minimizing $\mathcal{L}_{\text{vlm}}$ increases the likelihood that the generated video contains the requested motion. Gradients pass through the frozen critic and generator to tune only $\boldsymbol{\alpha}$ and $\Delta\mathbf{v}$.

For efficiency, Qwen evaluates a temporally sampled subset of the generated frames. The sampling schedules and their effects on the temporal emphasis and apparent motion speed are detailed in the supplementary material.

3.4 Regularized Test-Time Tuning

The VLM objective alone can produce latent drift, visual artifacts, or solutions that exploit the critic, as shown in Fig. 4. We therefore regularize both lightweight conditioning variables and preserve the appearance and temporal consistency of the prompt-only output $\hat{\mathbf{x}}_{\text{base}}$. The weights of regularizers are backbone-specific and are reported in the supplementary material.

Let $D_{\text{temp}}(\hat{\mathbf{x}})$ denote the mean LPIPS (Zhang et al., 2018) distance between consecutive sampled frame pairs. The temporal term penalizes frame-to-frame variation beyond that of the prompt-only output, reducing flicker and abrupt changes without suppressing motion already produced by the backbone. We tune only $\boldsymbol{\alpha}$ and $\Delta\mathbf{v}$, while keeping all pretrained model parameters fixed. The complete test-time objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{vlm}} &+ \lambda_{\alpha}\left\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\right\|_2^2 + \lambda_v \left\|\Delta\mathbf{v}\right\|_2^2 \\ &+ \lambda_{\text{lpips}}\mathcal{L}_{\text{LPIPS}}\left(\hat{\mathbf{x}}, \hat{\mathbf{x}}_{\text{base}}\right) \\ &+ \lambda_{\text{temp}} \max\left(0,\ D_{\text{temp}}(\hat{\mathbf{x}}) - D_{\text{temp}}(\hat{\mathbf{x}}_{\text{base}})\right).\end{aligned} \tag{3}$$

4 Experiments

4.1 Experimental Setup

We evaluate Sound Sparks Motion on a controlled benchmark of 25 motion-centric edits, each pairing a source video with a prompt describing an action or state change. It spans motions of varying complexity and spatial extent, from localized state changes (like a rose blooming) to broader actions (like a dog jumping), and includes both generated and real videos. Some related actions appear in different subjects and environments, allowing us to test whether performance depends on a particular scene. We apply Sound Sparks Motion to all 25 examples (including six transfer examples). Because running every baseline on every example is costly in computation and commercial-model usage, we conduct the full baseline comparison on a 12-example subset spanning distinct motions and scenes; the selection criteria and complete scenario list are provided in the supplementary material.

All methods use the same source videos, prompts, and matched generation settings whenever supported; baseline-specific inputs follow the corresponding official setups. All results use a single fixed generation seed $\boldsymbol{\xi}$, shared across all scenarios and with the prompt-only backbone baseline (Eq. 1), and each edit is optimized in a single run. The final output is selected automatically as the checkpoint with the lowest total objective $\mathcal{L}$ (Eq. 3), i.e., the VLM loss plus the regularizers; no manual selection over seeds, runs, or checkpoints is performed. More details are provided in the supplementary material.

Baselines. We compare against prompt-only **LTX-2 Retake**; tuning-free **UniEdit**; instruction-guided **Kiwi-Edit**; **VACE** with Wan2.1-1.3B and LTX-Video-2B HaCohen et al. (2024) backbones; first-frame-guided **LoRA-Edit** Gao et al. (2026); and the commercial editor **Runway-Aleph** Runway (2025). Baseline-specific configurations and auxiliary inputs are detailed in our released code.

Evaluation Protocol. We use both VLM-based evaluation and a human study. GPT-5.5 Thinking receives the source video, edit prompt, and candidate edits, and evaluates task success (TS), edit

Table 1: **Human evaluation across 12 scenarios.** Win and Top-3 denote first-place and top-three rates; Score assigns 3/2/1 points to first/second/third place; We report task achievement (*Ach.*) –an optional checkbox in the evaluation– as unfiltered (*all*) and filtered to only those who marked at least two checkboxes (*filtered*).

Method	**Win↑**	**Top-3↑**	**Score↑**	**Ach. (all)↑**	**Ach. (filtered)↑**
VACE-Wan	0.0%	8.3%	0.10	7.1%	10.8%
VACE-LTX	0.0%	1.2%	0.02	1.6%	2.5%
Kiwi-Edit	0.0%	10.7%	0.12	6.7%	10.8%
UniEdit	0.8%	4.0%	0.06	3.6%	5.7%
LTX-2 Retake	10.3%	63.1%	1.08	24.6%	39.2%
Runway-Aleph	11.1%	58.3%	0.98	26.2%	41.8%
LoRA-Edit	21.4%	59.1%	1.28	31.0%	48.7%
Sound Sparks Motion	**56.3%**	**95.2%**	**2.37**	**50.4%**	**79.1%**

Method	**TS↑**	**VLM↑**	**EA↑**	**MQ↑**	**Rank↓**
VACE-Wan	8.3%	3.95	2.33	1.92	5.58
VACE-LTX	8.3%	3.51	1.58	1.50	6.00
Kiwi-Edit	16.7%	4.20	2.46	2.17	5.33
UniEdit	33.3%	2.88	3.12	2.08	7.33
LTX-2 Retake	50.0%	5.77	4.67	4.29	3.00
Runway-Aleph	50.0%	5.56	5.46	4.88	3.58
LoRA-Edit	33.3%	5.37	4.33	3.88	3.58
Sound Sparks Motion	**83.3%**	**7.16**	**7.42**	**6.48**	**1.58**

Table 2: **GPT-5.5 Thinking evaluation across 12 scenarios.** TS is task success rate; VLM Score is the weighted 0–10 score; EA and MQ denote edit alignment and motion quality; VLM Rank is the average comparative rank, lower is better.

alignment (EA), motion quality (MQ), source preservation (SP), and visual quality (VQ). We emphasize motion-centric criteria using $S_{\text{vlm}} = 0.35S_{\text{EA}} + 0.25S_{\text{MQ}} + 0.20S_{\text{SP}} + 0.20S_{\text{VQ}}$. TS is the percentage of scenarios judged to clearly realize the requested action or state change.

For human evaluation, 21 participants evaluate the same 12 scenarios in an anonymous randomized survey. For each scenario, participants view the source, prompt, and eight anonymized outputs, select their top three, and mark outputs that clearly achieve the requested edit. The reported metrics are defined in Tab. 1.

Table 3: Ablation of guidance modality (VLM evaluation, 10 scenarios). Scores on a 1–10 scale. *Alignment* reflects semantic fidelity to the text prompt.

Metric	**Text**	**Audio**	**Both**
Alignment↑ *(text role)*	7.3	7.1	**7.5**
Naturalness↑ *(audio role)*	7.0	7.3	**7.9**
Preservation↑	7.6	7.8	**8.4**
Overall↑	7.25	7.24	**7.72**

4.2 Main Results

Qualitative Comparison. Supplementary Fig. 6 compares Sound Sparks Motion with representative baselines. Prompt-only Retake often preserves the source scene but misses the requested motion, while several general-purpose editors introduce visible changes at the cost of identity drift, layout changes, or abrupt object insertions. Our method more consistently realizes the requested temporal change while preserving the source subject and scene.

Quantitative Comparison. Tables 1 and 2 summarize human and VLM-based comparisons over the 12-scenario benchmark. In the human study, Sound Sparks Motion receives 56.3% of all first-place votes, more than twice the win rate of the strongest baseline, LoRA-Edit at 21.4%. It also achieves the highest Top-3 Rate, AVG Score, and user-marked task achievement, indicating that participants prefer our outputs not only as the best result, but also consistently rank them among the

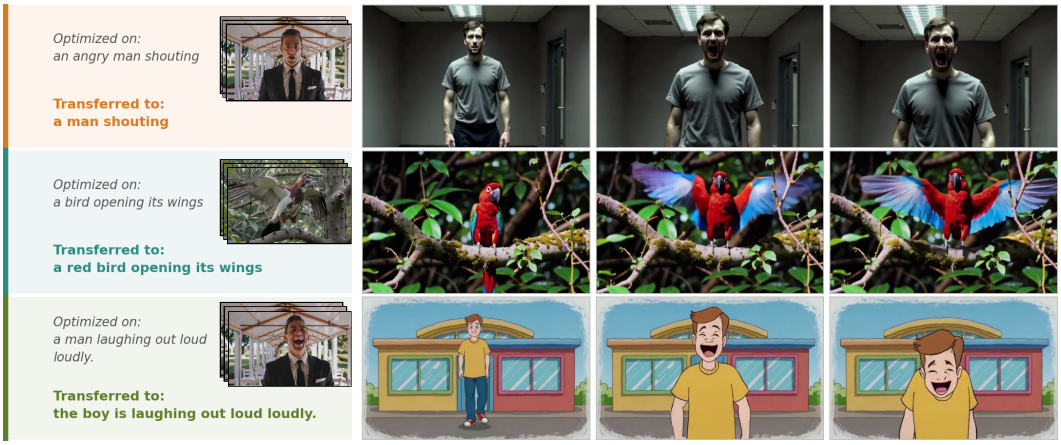


Figure 5: **Transferability of optimized latents.** Learned audio and residual text latents are reused on related target prompts without additional tuning, transferring motions such as shouting, wing opening, and laughing across different subjects and scenes. Note that in the last row although the latents are initially learned for a realistic video, they can be transferred to a cartoonish video, which shows the generality of the learned latents.

strongest candidates. GPT-5.5 Thinking evaluation confirms the same trend: Sound Sparks Motion achieves the highest scores across all motion-related metrics and best average rank. Over LTX-2 Retake, our method improves task success from 50.0% to 83.3% and motion quality from 4.29 to 6.48, and outperforms general-purpose editors in edit alignment and rank, confirming that test-time latent tuning is particularly effective for motion-centric edits.

Runtime. Test-time latent tuning for LTX-2 takes approximately 10 minutes; the turtle example in Fig. 6 required 589 seconds end-to-end. Timing for other models, per-iteration timing and hardware details are provided in supplementary Sec. C.8.

4.3 Ablation Studies

Effect of Joint Tuning. Figs. 2a and 2b show that text and audio provide complementary controls: text-residual tuning improves semantic alignment, while audio-latent tuning supports smooth and natural temporal motion, particularly when text-only tuning is insufficient. Tab. 3 confirms these roles through the *Alignment* and *Naturalness* scores, respectively, with joint tuning ranking first in **8 out of 10** scenarios, including one tie.

Effect of Motion Supervision. Fig. 4a shows that frame-level CLIP guidance produces a mostly static yawning state without a clear temporal transition. XCLIP provides limited temporal awareness but can miss longer actions and introduce global changes in lighting and appearance, whereas Qwen better captures whether an action unfolds over time, producing a more natural and temporally coherent edit.

Table 4: **Human evaluation of Sound Sparks Motion on MiniMax-H3 across 25 scenarios.** H3 + Ours is MiniMax-H3 equipped with Sound Sparks Motion. Preference rows report response percentages; task accomplishment reports an independent checkbox rate, for which ties do not apply.

Metric	H3	H3 + Ours	Tie
Task accomplished↑	46.67%	**68.27%**	–
Better overall↑	14.93%	**61.33%**	23.73%
More natural motion↑	20.00%	**51.73%**	28.27%
Stronger motion↑	16.80%	**58.67%**	24.53%

Effect of Regularization. Fig. 4b shows that unregularized VLM guidance can produce blur, inconsistent colors, abrupt motion, and latent drift. Latent regularization reduces drift, LPIPS improves frame-level fidelity and color consistency, and temporal regularization smooths frame-to-frame evolution. Combining all three provides the best trade-off between cleaner frames and coherent motion.

4.4 Transferability of Optimized Latents

To test generalization beyond a single clip, we tune the audio representation and text residual on a source video and prompt, then reuse both unchanged on a different video with a related prompt. As shown in Fig. 5, the transferred controls induces related motions on new target clips without additional tuning, suggesting that they capture reusable motion-edit directions rather than memorized frame-level changes.

4.5 Generalization Across Audio-Video Backbones

Our framework applies to pretrained audio-video generators that support video editing and expose audio- and text-conditioning representations. Beyond LTX-2 Retake, we instantiate Sound Sparks Motion in MiniMax-H3's prompt-guided video-to-video interface (MiniMax, 2026). We keep its model parameters and video-reference conditioning fixed and tune only two lightweight variables—the audio-reference representation and a residual on the prompt conditioning—using the same training-free objective, with H3-specific weights, critic aggregation, and update rules detailed in the supplementary material.

We compare untuned H3 with H3 equipped with Sound Sparks Motion across a separate 25 motion-centric set of scenarios. In an anonymous paired study, 15 participants judge whether each output achieves the edit and select the better result based on motion realization, source preservation, and overall quality.

Table 5: **Human evaluation of conditioning modalities on MiniMax-H3.** Best Pick is the percentage of trials ranked first; Edit Success and Naturalness are mean user ratings on a 1–5 scale. The full variant tunes both audio and residual text conditioning.

Tuned conditioning	Best Pick↑	Edit Success↑	Naturalness↑
Audio only	28.89%	3.62	3.61
Text only	12.22%	3.33	3.08
Audio + Text (full)	**58.89%**	**4.08**	**3.63**

As shown in Tab. 4, Sound Sparks Motion increases task achievement from 46.67% to 68.27% and is preferred in 61.33% of paired comparisons. We did an ablation study and human evaluation on H3 as well, with 15 participants on 12 scenarios to show that text-residual tuning or audio latent tuning alone are not sufficient and the best approach is to tune them both (Tab. 5). These result show that our lightweight multimodal conditioning tuning is not specific to LTX-2 and can activate or enhance motion in a substantially different omni-model architecture. Implementation details, runtime, ablations, qualitative examples, and per-scenario results are provided in the supplementary material and our supplementary website [1].

5 Conclusions, Limitations and Future Work

We presented a training-free approach for motion-centric video editing by tuning audio and residual text conditioning in a frozen audio-visual generator. More broadly, stronger multimodal conditioning does not necessarily improve controllability: motion emerges from interactions across modalities, and increased cross-modal attention can further entangle them. However, these pathways also contain latent motion controls that can be exploited through test-time tuning without modifying model weights, shifting the focus from introducing new controls to leveraging those already encoded by the model. Our approach was ranked the best with a meaningful gap in comparison with other baselines in the human evaluations and improved activating motion (task achievement) to 79.1% on top of LTX-2 and to 68.27% on top of H3.

Our limitations include additional test-time computation, which may restrict interactivity; some failure cases on complex or ambiguous actions; and coarse VLM-based binary feedback in the absence of a general text-motion alignment model. Future work will develop more capable text-motion correspondence objectives, reduce tuning costs by incorporating feedback earlier in the denoising process, and evaluate additional audio-visual editors as they become available.

[1] https://sound-sparks-motion.supplementary-site.workers.dev

AI USE STATEMENT

We used generative AI tools to assist with literature discovery; improve the grammar, clarity, and concision of portions of the paper; and support coding and experimental execution, including adapting code to different GPU configurations. AI tools also assisted in preparing the demonstration video, including its animations and transitions. The authors reviewed the AI-assisted material and take responsibility for the final text, implementation, results, and presentation artifacts.

ETHICS STATEMENT

Motion editing can be misused to fabricate actions in real footage or falsely attribute them to individuals. This risk is relevant even though most of our evaluation uses generated videos, as the method can also be applied to real clips. Edited outputs should be identified as such when shared, and deployment would require safeguards for provenance and identity-sensitive content. Source-video use must respect applicable licensing and consent requirements. Our evaluation also includes human judgments of edited clips; the study interface and evaluation procedure are described in the supplementary material.

For our human evaluation, we did not request or store participants' personal information. Before starting, participants were informed that their choices would be used only to calculate per-scenario and overall evaluation metrics, and proceeded by clicking "Consent and start the user study." We report these responses only in aggregate.

REPRODUCIBILITY STATEMENT

Section 3 specifies the test-time tuning procedure and its instantiations on LTX-2 Retake and MiniMax-H3. Section 4 describes the benchmark, baselines, and evaluation protocol, while the supplementary material provides additional implementation settings, evaluation prompts, and human-study details. We plan to release the implementation publicly upon acceptance to support independent reproduction of our results.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *Proc. CVPR*, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Vd3d: Taming large video diffusion transformers for 3d camera control. *Proc. ICLR*, 2025b.

Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Matan Ben-Yosef, Tavi Halperin, Naomi Ken Korem, Mohammad Salama, Harel Cain, Asaf Joseph, Anthony Chen, Urska Jelercic, and Ofir Bibi. Avcontrol: Efficient framework for training audio-visual controls. *arXiv preprint arXiv:2603.24793*, 2026.

Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz. Motionv2v: Editing motion in a video, 2025a. URL https://arxiv.org/abs/2511.20640.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, 2025b.

Anthony Chen, Naomi Ken Korem, Gal Zeevi, Tavi Halperin, Matan Ben Yosef, Urska Jelercic, Ofir Bibi, Or Patashnik, and Daniel Cohen-Or. Just-dub-it: Video dubbing via joint audio-visual diffusion, 2026. URL https://arxiv.org/abs/2601.22143.

Xiaoyan Cong, Haotian Yang, Angtian Wang, Yizhi Wang, Yiding Yang, Canyu Zhang, and Chongyang Ma. Viva: Vlm-guided instruction-based video editing with reward optimization. *arXiv preprint arXiv:2512.16906*, 2025.

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. *arXiv preprint arXiv:2310.05922*, 2023.

DecartAI. Lucy edit: Open-weight text-guided video editing. Online, 2025. Manuscript / project page.

Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. Controllable first-frame-guided video editing via mask-aware loRA fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xkRMJ1Y7Um.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals, 2025. URL https://arxiv.org/abs/2505.19386.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals, 2026. URL https://arxiv.org/abs/2601.05848.

Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. Ropecraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers, 2025. URL https://arxiv.org/abs/2505.13344.

Shelly Golan, Michael Finkelson, Ariel Bereslavsky, Yotam Nitzan, and Or Patashnik. ParetoSlider: Diffusion models post-training for continuous reward control. *arXiv preprint arXiv:2604.20816*, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025. doi: 10.48550/arXiv.2503.10592.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. doi: 10.48550/arXiv.2106.09685.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 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.

Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. Generative video motion editing with 3d point tracks. *arXiv preprint arXiv:2512.02015*, 2025.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026. doi: 10.48550/arXiv.2602.08068.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5521–5531, 2022.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance, 2026. URL https://arxiv.org/abs/2603.02175.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv:2303.04761*, 2023.

Wan-Duo Kurt Ma, J. P. Lewis, and W. Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation, 2023.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

MiniMax. Minimax h3: An open model breaking the boundaries between tasks and modalities. MiniMax Research, July 2026. URL https://www.minimax.io/blog/minimax-h3. Accessed: 2026-09-07.

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

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *arXiv preprint arXiv:2405.13865*, 2024. doi: 10.48550/arXiv.2405.13865.

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B. Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uQjySppU9x.

Koichi Namekata, Yash Kant, Zhizheng Liu, Ryan D. Burgert, Yuancheng Xu, Kuan Heng Lin, Emmett Steven, Julien Philip, Li Ma, Andrea Vedaldi, Paul Debevec, and Ning Yu. Go-with-the-track: Video compositing and motion control with point tracking. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA*, SIGGRAPH Conference Papers '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 979-8-4007-2554-8/2026/07. doi: 10.1145/3799902.3811093. URL https://doi.org/10.1145/3799902.3811093.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

Quynh Phung, Long Mai, Cusuh Ham, Feng Liu, Jia-Bin Huang, and Aniruddha Mahapatra. Trace: Object motion editing in videos with first-frame trajectory guidance. *arXiv preprint arXiv:2603.25707*, 2026.

Sagi Polaczek, Or Patashnik, Ali Mahdavi-Amiri, and Daniel Cohen-Or. In-context sync-lora for portrait video editing. *arXiv preprint arXiv:2512.03013*, 2025.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *CVPR*, 2025.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation, 2023.

Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients, 2024. URL https://arxiv.org/abs/2407.08737.

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. *arXiv:2303.09535*, 2023.

Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv preprint arXiv:2305.12328*, 2023. doi: 10.48550/arXiv.2305.12328.

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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Runway. Introducing Runway Aleph. https://runwayml.com/research/introducing-runway-aleph, July 2025. Runway Research. Accessed: 2026-05-10.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Maham Tanveer, Yang Zhou, Simon Niklaus, Ali Mahdavi Amiri, Hao Zhang, Krishna Kumar Singh, and Nanxuan Zhao. Multicoin: Multi-modal controllable video inbetweening, 2025. URL https://arxiv.org/abs/2510.08561.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 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.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. *arXiv preprint https://arxiv.org/abs/2410.18978*, 2024a.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024b.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-drf: Free, rich, and differentiable reward for video diffusion fine-tuning, 2026. URL https://arxiv.org/abs/2601.04153.

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 Computer Vision and Pattern Recognition Conference*, 2025.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. IEEE, 2003.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025.

Jiacong Xu, Yiqun Mei, Ke Zhang, and Vishal M Patel. Freevis: Training-free video stylization with inconsistent references. *arXiv preprint arXiv:2510.01686*, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2630–2640, 2025.

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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model, 2025. URL https://arxiv.org/abs/2503.19839.