FlowDirector: Training-Free Flow Steering for Precise Text-to-Video Editing

Guangzhao Li[1,2] Yanming Yang[1] Chenxi Song[1] Xiaohong Liu[2] Chi Zhang[1†]
[1]AGI Lab, Westlake University [2]Shanghai Jiao Tong University
gzhaocs@gmail.com, chizhang@westlake.edu.cn

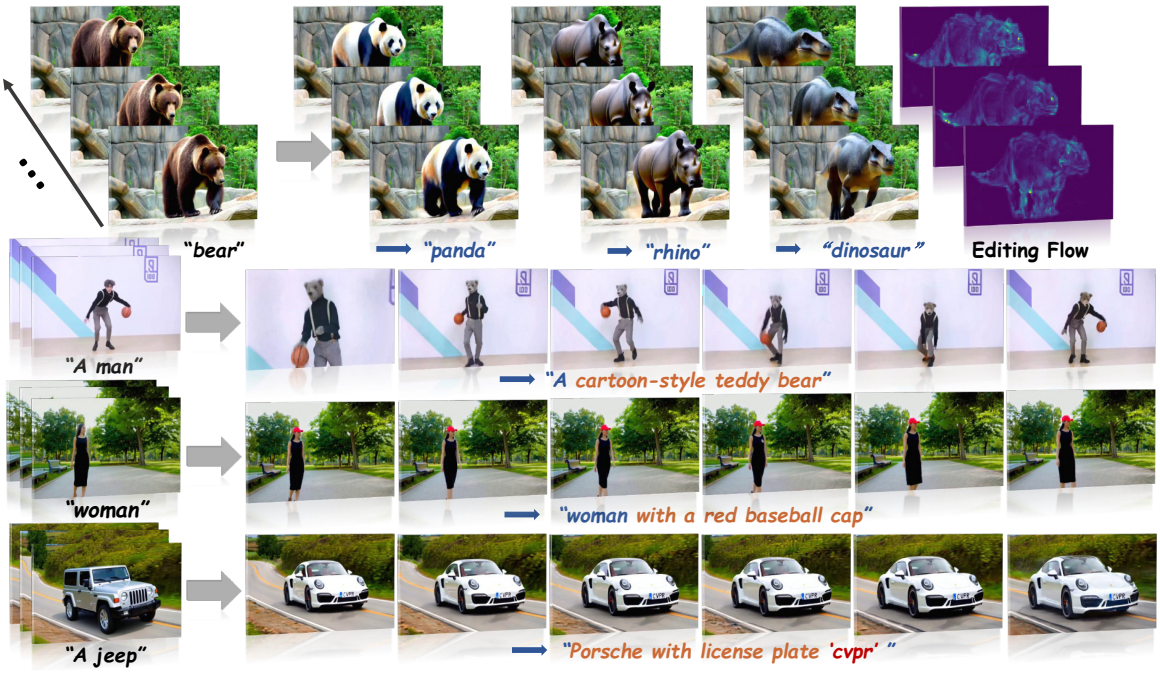

Figure 1. **Editing results.** Our method uses a source video and prompt to generate precise, semantic edits while preserving background content and motion, yielding high visual fidelity and temporal coherence results.

Abstract

Text-driven video editing aims to modify video content based on natural language instructions. While recent training-free methods have leveraged pretrained diffusion models, they often rely on an inversion-editing paradigm. This paradigm maps the video to a latent space before editing. However, the inversion process is not perfectly accurate, often compromising appearance fidelity and motion consistency. To address this, we introduce FlowDirector, a novel training-free and inversion-free video editing framework. Our framework models the editing process as a direct evolution in the data space. It guides the video to transition smoothly along its inherent spatio-temporal manifold using an ordinary differential equation (ODE), thereby avoiding the inaccurate inversion step. From this foundation, we introduce three flow correction strategies for appearance, motion, and stability: 1) Direction-aware flow correction amplifies components that oppose the source direction and removes irrelevant terms, breaking conservative streamlines and enabling stronger structural and textural changes. 2) Motion-appearance decoupling optimizes motion agreement as an energy term at each timestep, significantly improving consistency and motion transfer. 3) Differential averaging guidance strategy leverages differences among multiple candidate flows to approximate a low variance regime at low cost, suppressing artifacts and stabilizing the trajectory. Extensive experiments across various editing tasks and benchmarks demonstrate that FlowDirector achieves state-of-the-art performance in instruction following, temporal consistency, and background preservation, establishing an efficient new paradigm for coherent video editing without inversion.

†Corresponding author.

1. Introduction

Text-driven video editing aims to modify video content based on natural language instructions, offering a powerful and intuitive interface for content creation and manipulation. The rapid development of this field has been largely fueled by recent advances in generative artificial intelligence, particularly the emergence of large-scale pre-trained diffusion models [11, 40] such as Stable Diffusion [31, 34]. These models are trained on extensive image-text datasets [3, 23, 36] and encode rich visual and semantic priors, enabling them to synthesize high-quality and semantically aligned images from textual descriptions. Building upon these capabilities, recent research efforts [5, 7, 14, 15, 17, 26, 32, 37, 47] have extended diffusion models from static image generation to dynamic video domains. Such extensions have given rise to training-free video editing [5, 7, 15, 25, 32, 45, 46, 50], which directly manipulates video frames under textual guidance by reusing the knowledge embedded in pre-trained diffusion backbones.

Despite promising progress, text-driven video editing presents unique and significant challenges that distinguish it from image editing. Videos are inherently high-dimensional sequences with rich temporal dynamics, requiring coherence not only within each frame but also across time. Most existing training-free methods [5, 7, 15, 32, 46] rely on an inversion-based [29] paradigm, mapping the input video to a latent trajectory. This approach works effectively for images, but it presents significant challenges for video. Inverting an entire video sequence coherently requires generating a temporally smooth latent trajectory, which existing image-based inversion techniques are not designed to produce. In practice, producing such a coherent latent trajectory is difficult. For instance, [46] shows that the reconstructed results have significant differences in appearance and motion compared to the source video. Small per-frame inversion errors accumulate and cause temporal drift and flicker. Attention misalignment across frames disrupts layout and identity, and mixing motion and appearance leads to style inconsistency and unrealistic motion. Together, these issues substantially degrade edit quality.

In this work, we explore a novel inversion-free framework for text-driven video editing that overcomes the fundamental limitations of prior methods. Inspired by recent progress in flow-based inversion-free image editing [22, 49], we develop an editing paradigm that models the transformation from the original video to the edited result as a direct evolution in data space, governed by a learned ordinary differential equation (ODE). Unlike traditional approaches that attempt to recover precise latent representations for each frame through inversion, the inversion-free paradigm constructs a smooth transformation path that guides the video along its native spatiotemporal manifold, achieving significant improvements in editing fidelity and temporal consistency compared to traditional inversion methods. However, a straightforward application of this paradigm to video editing still presents significant challenges. First, the temporal dimension introduces substantially richer appearance and structural information across frames, making it more difficult to achieve drastic semantic transformations while preserving spatiotemporal plausibility. Second, the lack of explicit strong constraints leads to severe motion distortion and drift in the editing results. Besides, the compounded sampling noise across frames makes the system exceptionally sensitive to perturbations, undermining stability throughout the video sequence. To tackle these challenges, we develop three complementary, training-free flow correction strategies that jointly optimize appearance preservation, motion stability, and spatiotemporal coherence, yielding significantly improved editing quality and robustness.

Standard inversion-free methods often skip early high-noise diffusion steps to maintain structural consistency. This is in contrast to traditional inversion-based paradigms, which actively disrupt these strong structural constraints by mapping data to noise, thereby creating the necessary latitude for significant manipulation. As a result, inversion-free method retains strong content information from the source video, creating a "semantic gravitational pull" that resists drastic alterations. In this regime, the standard editing flow is frequently overwhelmed by the dominant source priors, resulting in overly conservative edits. To overcome this inherent resistance, we introduce **Direction-Aware Flow Correction** (DA-FC). Our key insight is that the raw editing flow contains opposing forces: components aligned with the source that cause redundant drift, and components opposing the source that drive actual semantic change. DA-FC mathematically decouples these forces, actively suppressing the aligned components while significantly amplifying the opposing ones. This strategy effectively breaks the semantic dominance of the source video, enabling profound structural transformations without sacrificing the stability benefits of the inversion-free paradigm.

Another critical challenge in direct ODE editing is the entanglement of motion and appearance flows. In many editing scenarios, we aim for conflicting goals: drastic appearance transformations (*e.g.*, turning a person into a bear) must coexist with strict motion preservation (*e.g.*, maintaining the exact trajectory of a dribbled basketball). Unlike inversion-based approaches that can inject explicit structural guidance (*e.g.*, cross-attention maps) from the source video, inversion-free methods lack such isolated constraints. Consequently, during complex dynamics like occlusions, errors in motion can accumulate and be misinterpreted by the model as necessary appearance changes, leading to severe distortion (*e.g.*, objects shifting incorrectly in depth). Simply enforcing similarity to the source video to

fix this is counterproductive, as it would actively revert the desired appearance edits. To resolve this conflict, we propose **Motion-Appearance Decoupling Flow Correction** (MAD-FC). The core motivation is to establish an orthogonal control plane: by mathematically isolating "pure motion" features from static appearance information, we can formulate an energy function that strictly penalizes motion deviations while remaining completely agnostic to the intended semantic transformations. This allows for continuous, "safe" rectification of temporal drift without compromising the magnitude of the edit.

Finally, direct ODE editing faces the challenge of high-variance velocity estimates, where single-sample noise introduces directional jitter that accumulates into severe temporal flickering. While brute-force averaging [22] can mitigate this in image editing, it is computationally intractable for the high-dimensional video domain. We propose **Differential Averaging Guidance** (DAG) to transcend this limitation. Instead of passively relying on massive sampling to dilute noise—which is too costly for video—DAG adopts an active navigation strategy. By explicitly contrasting a high-quality consensus estimate against a baseline of high-variance outliers, we extract a "noise drift" signal. DAG then uses this signal not just to average out errors, but to actively steer the trajectory away from these noisy deviations, efficiently locking the editing process onto a stable, low-variance manifold with minimal computational overhead.

These three correction strategies, integrated with our direct editing ODE, constitute our framework, FlowDirector. Extensive experiments on standard benchmarks demonstrate that our method consistently outperforms existing training-free baseline methods across multiple key dimensions, including adherence to editing instructions, semantic consistency, motion consistency, and background preservation. Overall, our contributions are summarized as follows:

- We propose **FlowDirector**, a novel training-free video editing framework. It employs a novel inversion-free paradigm and can perform various video editing tasks, providing new understanding and insights for effective video editing.
- We propose three training-free flow correction strategies: direction-aware flow correction, motion-appearance decoupling flow correction, and differential averaging guidance, which significantly improve appearance editing, motion preservation, and editing stability.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance across various editing scenarios and benchmarks. The code will be publicly released to facilitate further research.

2. Related Works

Text-to-Image Editing. Diffusion model-based [11, 24, 40] image editing is primarily classified into two paradigms. Inversion-based methods [2, 10, 16, 18, 28, 29, 48] use an invert-then-denoise strategy: source images are mapped to latent noise via deterministic sampling (*e.g.*, DDIM Inversion [29, 39]), with editing by manipulating internal representations during denoising. Their limitation is that inversion approximation errors degrade reconstruction quality and editing fidelity. To overcome this, inversion-free methods emerged. InfEdit [49] uses Denoising Diffusion Consistent Models for virtual inversion. FlowEdit [22] uses Flow Matching Models to construct a direct ordinary differential equation (ODE) trajectory between the source and target distributions, bypassing inversion. FlowAlign [19] improves editing quality by optimizing the sampling path.

Text-to-Video Editing. This field is primarily categorized by the underlying generative model. Early approaches adapted T2I models [6, 31, 34, 35], facing temporal consistency challenges due to a lack of temporal understanding. Some are zero-shot (FateZero [32], TokenFlow [7], Flatten [5], RAVE [15]), others require fine-tuning (Tune-A-Video [47], Video-P2P [26]). These T2I-based methods often struggle with temporal consistency. Recent work leverages native T2V models [1, 4, 9, 12, 13, 21, 38, 42–44] and their learned spatiotemporal priors for improved temporal consistency (*e.g.*, VideoSwap [8], VideoDirector [46]). However, all the aforementioned video editing methods are inversion-based. Similar to image editing, their inherent limitations restrict substantial visual transformations, resulting in insufficient editing precision. Furthermore, the accumulation of temporal errors in the two stages of inversion and denoising often causes motion distortion or drift.

3. Method

3.1. Preliminary

Our framework builds on two standard components in flow-based generative modeling.

Flow Matching. Flow Matching learns a time-dependent vector field that transports a simple prior (*e.g.*, $\mathcal{N}(0, I)$) to the data distribution. A neural network $v_\theta(x, t)$ is trained by minimizing

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\big[\, \|v_\theta(x,t) - v_t(x \,|\, x_1)\|^2 \big]\,, \tag{1}$$

where $t \sim \mathcal{U}(0,1)$, $x_1 \sim p_1$, $x \sim p_t(\cdot \,|\, x_1)$, $p_t(x \,|\, x_1)$ is a prescribed path from p_0 to x_1, and $v_t(x \,|\, x_1)$ is its oracle velocity.

Rectified Flow. Rectified Flow [27] specializes Flow Matching to linear paths between p_0 and p_1. Given $x_0 \sim p_0$ and $x_1 \sim p_1$, it considers $x_t = (1-t)x_0 + tx_1$ and learns

$$\mathcal{L}_{\text{RF}} = \mathbb{E}\big[\, \|v_\theta(x_t,t) - (x_1 - x_0)\|^2 \big]\,, \tag{2}$$

where $t \sim \mathcal{U}(0,1)$, $x_0 \sim p_0$, $x_1 \sim p_1$. This linearization simplifies the oracle velocity and enables efficient, stable ODE integration driven by v_θ.

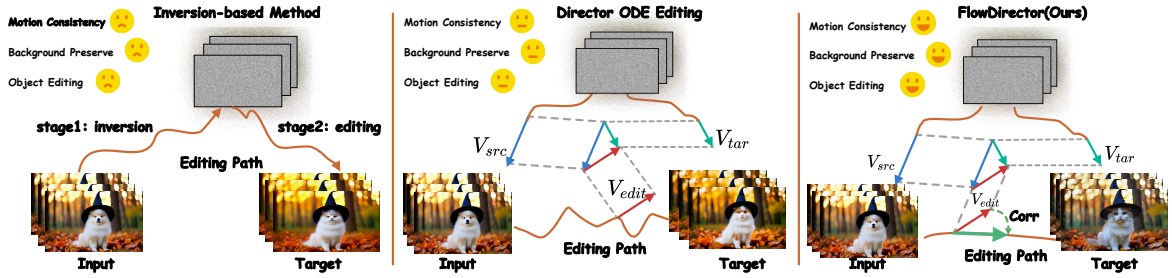

Figure 2. **We compare inversion-based methods, FlowDirector (Direct ODE), and the full FlowDirector. Inversion-based methods** first map the source video into a Gaussian latent space and then use this inverted state as the starting point for the subsequent editing process. **FlowDirector (Direct ODE)** instead constructs source-side and target-side states at each timestep, estimates their velocity fields, and uses the difference between them as an editing flow that drives video editing directly in data space. Building on this formulation, the full **FlowDirector** further corrects the editing flow at every timestep, yielding a shorter and more efficient editing trajectory and substantially improving the final editing quality.

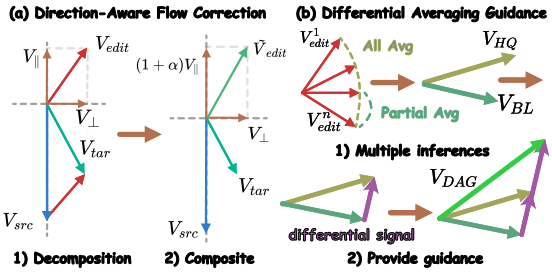

Figure 3. **Illustration of proposed Direction-Aware Flow Correction (DA-FC, a) and Differential Averaging Guidance (DAG, b).** DA-FC strengthens editing through orthogonal decomposition of the editing flow, amplifying parallel components that oppose the source semantics while eliminating co-directional components. (a) shows the case of amplifying the inverse component. DAG guides the editing flow toward a stable state by generating differential signals.

3.2. Editing Flow Generation

To avoid the inversion stage that often harms visual fidelity and motion consistency, we directly construct an editing trajectory in the data space using a pre-trained text-to-video (T2V) model v_θ. Given a source video X_{src} with prompt c_{src} (or an automatically inferred description by an LLM if c_{src} is not provided) and a target prompt c_{tar}. We define the editing state at time $t \in [0, 1]$ as Z_t^{edit}. Following FlowEdit [22], this state is constructed via the unified editing equation:

$$Z_t^{\text{edit}} = X_{\text{src}} - Z_t^{\text{src}} + Z_t^{\text{tar}}, \tag{3}$$

where Z_t^{src} and Z_t^{tar} are the perturbed source and target states at time t. The evolution of Z_t^{edit} is governed by an ODE whose velocity field is the difference between target and source velocities:

$$\frac{dZ_t^{\text{edit}}}{dt} = V_{\text{edit}}(t) = v_\theta(Z_t^{\text{tar}}, t, c_{\text{tar}}) - v_\theta(Z_t^{\text{src}}, t, c_{\text{src}}). \tag{4}$$

The trajectory starts from the source video $Z_1^{\text{edit}} = X_{\text{src}}$ and converges to the edited result Z_0^{edit} as $t \to 0$.

To compute $V_{\text{edit}}(t)$, we need the intermediate states Z_t^{src} and Z_t^{tar}. Following Rectified Flow [27], we obtain the source state by linear interpolation $Z_t^{\text{src}} = (1-t)X_{\text{src}} + tN_t$, where $N_t \sim \mathcal{N}(0, I)$. Since the target video is not observed, we derive Z_t^{tar} by rearranging Eq. (3): $Z_t^{\text{tar}} = Z_t^{\text{edit}} + Z_t^{\text{src}} - X_{\text{src}}$. This construction ensures that Z_t^{src} and Z_t^{tar} share the same noise N_t, so noise and other random perturbations cancel in the difference $V_{\text{edit}}(t)$, and the resulting editing flow primarily captures the semantic change induced by the prompts.

3.3. Direction-Aware Flow Correction

To promote thorough edits while improving stability, we introduce *Direction-Aware Flow Correction* (DA-FC). At each timestep we obtain the source and target velocities $V_{\text{src}} = v_\theta(Z_t^{\text{src}}, t, c_{\text{src}})$ and $V_{\text{tar}} = v_\theta(Z_t^{\text{tar}}, t, c_{\text{tar}})$, and define the editing velocity $V_{\text{edit}} = V_{\text{tar}} - V_{\text{src}}$ (we omit t below for brevity). We perform a token-wise orthogonal decomposition of V_{edit} along V_{src}:

$$V_{\parallel} = \frac{\langle V_{\text{edit}}, V_{\text{src}} \rangle}{\|V_{\text{src}}\|^2 + \varepsilon} V_{\text{src}}, \quad V_{\perp} = V_{\text{edit}} - V_{\parallel}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the Euclidean inner product and ℓ_2 norm, and $\varepsilon > 0$ is a small constant for numerical stability. We then apply the direction-aware correction:

$$\tilde{V}_{\text{edit}} = \begin{cases} V_{\perp}, & \text{if } \langle V_{\parallel}, V_{\text{src}} \rangle \geq 0, \\ (1+\alpha)\, V_{\parallel} + V_{\perp}, & \text{if } \langle V_{\parallel}, V_{\text{src}} \rangle < 0, \end{cases} \tag{6}$$

where $\alpha > 0$ is an amplification factor. Thus, components aligned with the source direction are suppressed to reduce redundant drift, while anti-aligned components are amplified to encourage meaningful structural changes, yielding a more decisive and stable editing trajectory.

The corrected flow still acts on the full spatial extent and may alter irrelevant regions. To localize edits, we derive an editing mask from cross-attention maps of the source and target prompts. We aggregate cross-attention maps over

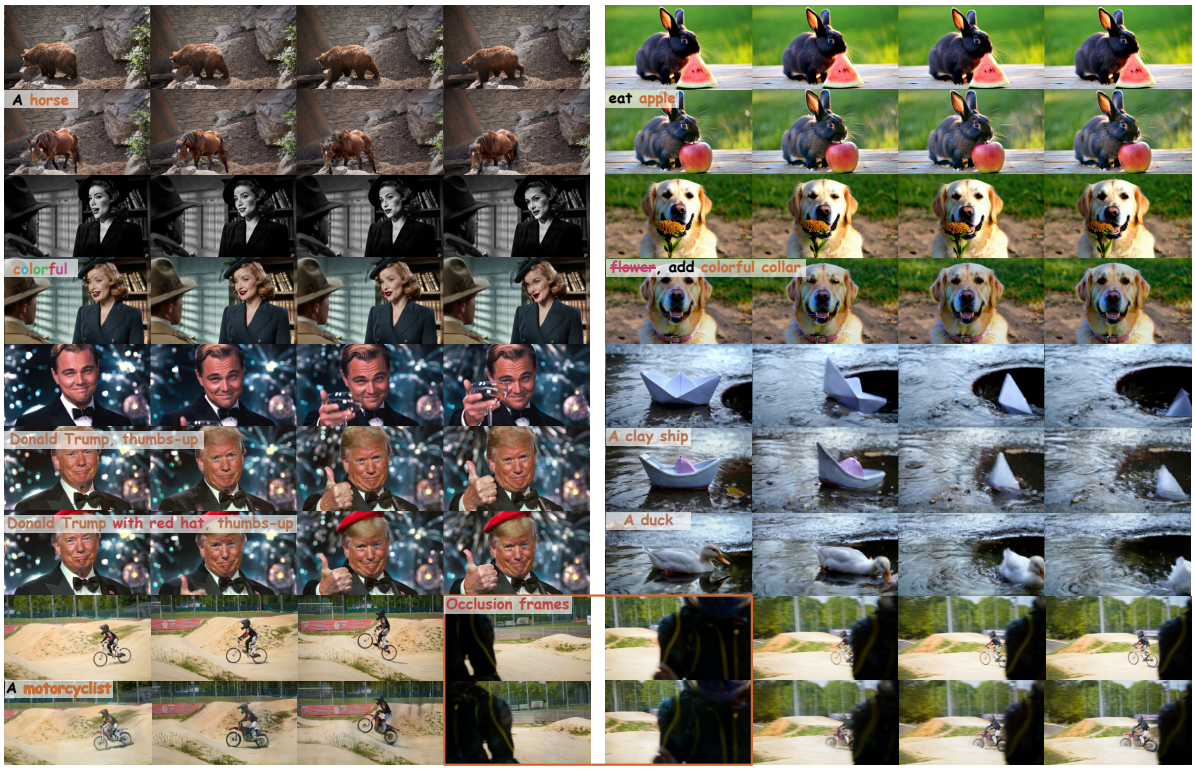

Figure 4. **Qualitative results of FlowDirector.** Our method can perform various types of editing tasks. The results demonstrate high fidelity to prompts, preservation of unedited regions and motion, temporal coherence, and visual plausibility.

heads and token dimensions, apply 2D average pooling and normalization per frame, and then threshold by the mean to obtain $\mathbf{M}_{\text{src}}$. The same procedure on target-prompt keywords yields $\mathbf{M}_{\text{tar}}$, we define $\mathbf{M} := \mathbf{M}_{\text{src}} \cup \mathbf{M}_{\text{tar}}$. To allow smooth transitions between edited regions and background, we soften this mask by applying a Euclidean distance transform d to the background and an exponential decay:

$$\widetilde{\mathbf{M}}_{c,t}(x,y) = \mathbf{M}_{c,t}(x,y) + \big(1 - \mathbf{M}_{c,t}(x,y)\big)\, e^{-\lambda\, d_{c,t}(x,y)}, \tag{7}$$

where (x, y) indexes spatial locations, c and t denote channel and time, and $\lambda > 0$ controls the decay rate. Finally, we apply the softened mask to the corrected editing flow:

$$\hat{V}_{\text{edit}} \;=\; \tilde{V}_{\text{edit}} \odot \widetilde{\mathbf{M}}, \tag{8}$$

where $\odot$ denotes element-wise multiplication.

3.4. Motion-Appearance Decoupling Flow Correction

To maintain motion consistency in an inversion-free setting, we introduce *Motion-Appearance Decoupling Flow Correction* (MAD-FC), which enforces a dynamic motion constraint at each denoising step. At time t, we first estimate the noiseless source and target states using the Tweedie estimator, $Z_0^{\text{src}} = Z_t^{\text{src}} - t\, V_{\text{src}}$ and $Z_0^{\text{tar}} = Z_t^{\text{tar}} - t\, V_{\text{tar}}$, and compute temporal averages on each side to obtain anchor frames $(A_h^{\text{src}}, A_h^{\text{tar}})$ as appearance references. Motion representations are then defined by subtracting the anchors, $G_{\text{src}} = Z_0^{\text{src}} - A_h^{\text{src}}$ and $G_{\text{tar}} = Z_0^{\text{tar}} - A_h^{\text{tar}}$, which suppress static appearance and emphasize motion. Finally, we measure motion disagreement between source and target using the energy:

$$J_t(Z) = \tfrac{1}{2}\, \|G_{\text{tar}} - G_{\text{src}}\|_2^2. \tag{9}$$

After the standard ODE update $Z_t^{\text{edit}} \leftarrow Z_t^{\text{edit}} + \Delta t\, V_{\text{edit}}$, we further minimize J_t with respect to Z_t by performing a single gradient-descent step. Using the first-order approximation $\nabla_{Z_t} J_t \approx G_{\text{tar}} - G_{\text{src}}$, we update the edited state in the negative gradient direction:

$$Z_t^{\text{edit}} \leftarrow Z_t^{\text{edit}} - \zeta\, (G_{\text{tar}} - G_{\text{src}}). \tag{10}$$

Rewriting this in terms of Z_0 and the anchors yields the practical update rule:

$$Z_t^{\text{edit}} \leftarrow Z_t^{\text{edit}} - \zeta \Big[\underbrace{(Z_0^{\text{tar}} - Z_0^{\text{src}})}_{\text{motion}} - \phi\, \underbrace{(A_h^{\text{tar}} - A_h^{\text{src}})}_{\text{appearance}} \Big]. \tag{11}$$

Here Δt is the ODE integration step size, $\zeta > 0$ controls the overall correction strength, and $\phi \in [0, 1]$ modulates how strongly appearance alignment via the anchors is enforced.

In effect, MAD-FC reduces the motion energy between source and target at every timestep, transferring source motion to the edited video while allowing a tunable trade-off between appearance correction and motion preservation

through ϕ. This yields stable edits that remain consistent even under occlusions and complex dynamics.

3.5. Differential Averaging Guidance

To stabilize the editing trajectory and reduce variance, we introduce *Differential Averaging Guidance* (DAG). The key idea is to form a high-quality velocity estimate by averaging multiple stochastic editing flows, and to use their discrepancy from a conservative baseline as a guidance signal. At each sampling step t, we compute a high-quality estimate V_{HQ} by averaging L_{HQ} velocity fields obtained under different noise samples:

$$V_{\text{HQ}}(Z_t^{\text{edit}}, t) \;=\; \frac{1}{L_{\text{HQ}}} \sum_{\ell=1}^{L_{\text{HQ}}} V_{\text{edit}}^{(\ell)}(Z_t^{\text{tar}}, Z_t^{\text{src}}, t, c^{\text{tar}}, c^{\text{src}}). \tag{12}$$

From the same pool, we construct a baseline estimate V_{BL} by selecting the K candidates with the lowest cosine similarity to V_{HQ} and averaging them:

$$V_{\text{BL}}(Z_t^{\text{edit}}, t) \;=\; \frac{1}{K} \sum_{i \in \mathcal{I}_K} V_{\text{edit}}^{(i)}(\cdot), \tag{13}$$

where $\mathcal{I}_K$ indexes the K least similar velocity fields, $K \leq L_{\text{HQ}}$, and $(\cdot)$ denotes the same arguments as in Eq. 12. The differential signal $\bar{D} \;=\; V_{\text{HQ}} - V_{\text{BL}}$ is then used to guide the final velocity:

$$V_{\text{DAG}} \;=\; V_{\text{HQ}} + w\,\bar{D}, \tag{14}$$

with $w > 0$ controlling the guidance strength.

In flow-matching generative models, large shift values create wide step intervals late in sampling. Different noise draws then perturb the editing trajectory and cause texture flicker. We fix a single noise sample for the late denoising phase. After a chosen cutoff timestep, we stop resampling and reuse this noise for all remaining steps. This stabilizes textures and other fine details.

4. Experiments

4.1. Experimental Setups

We build FlowDirector on the Wan-2.1 1.3B model [42] and perform inversion-free video editing with 50 denoising steps without skip sampling. For Direction-Aware Flow Correction, we fix $\alpha = 0.25$ to amplify anti-parallel components and $\lambda = 0.25$ to soften the attention mask. Motion-Appearance Decoupling Flow Correction uses task-dependent hyperparameters (ζ, ϕ): for edits with strong motion we set $(\zeta, \phi) = (0.01, 0.3)$, while for milder motion we use $(0.007, 0.5)$. Differential Averaging Guidance (DAG) estimates a high-quality velocity from three stochastic samples, forms a baseline by averaging the two candidates with the lowest cosine similarity to this mean, *i.e.*, $L_{HQ} = 3, K = 2$, and applies guidance with strength $w = 2.75$, reusing a fixed noise realization in the last eight denoising steps.

For qualitative and quantitative evaluation, we construct 150 video-text editing pairs from Internet videos and the DAVIS [30] dataset, covering insertion, deletion, and object editing. We evaluate at 41 and 81 frames and report results for both settings. Unless otherwise specified, all experiments for FlowDirector are run on a single NVIDIA H20 141G GPU. We compare against five state-of-the-art video editing methods: FateZero [32], FLATTEN [5], TokenFlow [7], RAVE [15], and VideoDirector [46], using their official implementations and default hyperparameters. All main results are obtained with the 1.3B model. Additional results for the 14B model and further implementation details are provided in the supplementary material.

4.2. Qualitative Results

We conduct a qualitative evaluation of FlowDirector on diverse video content under a variety of editing prompts. Our method handles large-scale object editing, video colorization, local attribute modification, object addition and removal, multi-object editing, and compositions of these operations (Figure 1 and Figure 4). For large-scale object editing, FlowDirector seamlessly converts the primary subject from one category to another while allowing substantial shape deformation (*e.g.*, the first and second rows, "brown bear" $\rightarrow$ "horse", and the fifth and seventh rows, "paper boat" $\rightarrow$ "duck"). For video colorization, it can colorize black-and-white videos while preserving the original structure and content (*e.g.*, the third and fourth rows, "colorful video"). For fine-grained object editing, FlowDirector modifies small objects while keeping the rest of the scene unchanged (*e.g.*, "watermelon" $\rightarrow$ "apple" in the third and fourth rows). It also maintains complex motion, such as the continuous falling motion in the "paper boat" case, and in challenging long-occlusion scenarios it preserves the identity and appearance of the edited object consistently before and after the occlusion (*e.g.*, the last two rows, "bike" $\rightarrow$ "motorcycle" with $\sim$20 occluded frames). Furthermore, FlowDirector adeptly handles attribute insertion on human subjects under strong appearance changes, simultaneously altering identity ("Leonardo" $\rightarrow$ "Donald Trump"), gesture ("raising the glass" $\rightarrow$ "giving a thumbs up"), and adding an extra red hat (fifth to seventh rows). Across these cases, the edits remain faithful to textual prompts, preserve details and motion in unedited regions, and exhibit strong temporal coherence and visual plausibility. We also compare FlowDirector with several state-of-the-art video editing methods (Figure 5), where FlowDirector consistently produces higher-quality edits with fewer artifacts. More comprehensive qualitative results and detailed comparisons are provided in the supplementary material.

Figure 5. **Qualitative comparison.** Our method outperforms previous methods across diverse editing tasks, demonstrating superior visual quality and temporal consistency. Best viewed zoomed-in.

Table 1. **Quantitative comparison.** We report Pick-Score, CLIP-T, CLIP-F, WarpSSIM, and Q_{edit} for 41-frame and 81-frame videos. Each entry is reported as a/b, corresponding to 41-frame and 81-frame results, respectively. ↑ indicates that higher is better. The "-" symbol indicates cases where a method cannot be evaluated at a given length due to memory limits or lack of support. We bold the **best value** for each metric and underline the second-best value, both computed excluding FlowDirector (14B).

Method	*Perceptual Quality*			*Temporal Quality*	
	Pick Score (%) ↑	**CLIP-T** ($\times 10^{-2}$) ↑	**CLIP-F** ($\times 10^{-2}$) ↑	**WarpSSIM** ($\times 10^{-2}$) ↑	$\mathbf{Q_{edit}}$ ($\times 10^{-6}$) ↑
FateZero [32]	20.41 / -	32.01 / -	92.25 / -	78.37 / -	25.09 / -
FLATTEN [5]	20.84 / 20.18	33.56 / 33.12	92.80 / 92.35	77.44 / 78.21	26.01 / 25.90
TokenFlow [7]	20.99 / 20.63	32.69 / 32.27	93.82 / 94.15	74.98 / 75.43	24.51 / 24.34
RAVE [15]	21.01 / 20.76	33.25 / 33.10	94.03 / 93.58	76.32 / 75.91	25.38 / 25.13
VideoDirector [46]	20.61 / -	32.56 / -	95.48 / -	75.89 / -	24.70 / -
FlowDirector (1.3B)	**21.82 / 21.69**	**34.64 / 34.63**	**97.34 / 96.90**	**78.49 / 79.10**	**27.19 / 27.31**
FlowDirector (14B)	22.61 / -	34.95 / -	97.30 / -	79.86 / -	28.67 / -

4.3. Quantitative Results

We conduct a quantitative comparison between FlowDirector and other state-of-the-art methods. Following prior work [5, 15, 32], we evaluate text-content alignment using CLIP-T, the average CLIP [33] embedding similarity between the text prompt and video frames. Temporal coherence is measured by CLIP-F, the average pairwise frame CLIP similarity. Structure preservation during editing is assessed via WarpSSIM, the average SSIM between the final edited video and the source video warped using RAFT [41] optical flow. Consistent with [5, 15], we utilize the composite metric Q_{edit} (WarpSSIM·CLIP-T) for a holistic evaluation of editing performance. Furthermore, we use Pick-Score [20] to evaluate overall perceptual quality and prompt alignment based on human preferences. Our experimental results (Table 1) demonstrate that our proposed method significantly outperforms current state-of-the-art techniques regarding both text alignment and temporal consistency, and achieves strong performance on Pick-Score. Furthermore, our method achieves superior results on the comprehensive Q_{edit} metric. This demonstrates that our method achieves the best overall results across diverse editing tasks.

4.4. Ablation Study

Direction-Aware Flow Correction. As shown in Table 2 and Table 3, removing the parallel component aligned with the source direction effectively improves the consistency of the editing results. As the scaling factor α increases,

Table 2. **Ablation results for Direction-Aware Flow Correction (DA-FC), Motion-Appearance Decoupling Flow Correction (MAD-FC) and Differential Averaging Guidance (DAG).** We highlight the best values for each metric.

Method	CLIP-T ↑	CLIP-F ↑	WarpSSIM ↑	Q_{edit} ↑
Director ODE (no skip)	34.59	94.66	62.71	21.69
Director ODE (skip)	32.23	96.84	78.90	25.43
w/o DA-FC	32.25	95.70	78.83	25.42
w/o MAD-FC	34.71	97.10	69.26	24.04
w/o DAG	34.62	97.19	78.32	27.11
FlowDirector	34.64	97.34	78.49	27.19

Table 3. **Ablation results for Direction-Aware Flow Correction (DA-FC).** Para denotes the parallel component aligned with the source semantic direction. Therefore, w/o Para indicates removal of only this component, while a '-' value for α signifies no amplification of the opposite component. We highlight the best values for each metric.

Method	α	CLIP-T ↑	CLIP-F ↑	WarpSSIM ↑	Q_{edit} ↑
w/o DA-FC	–	32.25	95.70	78.83	25.42
+ w/o Para	–	33.02	96.51	78.85	26.04
+ w/o Para	0.05	33.14	96.49	78.80	26.11
+ w/o Para	0.15	34.57	97.30	78.48	27.13
+ w/o Para	**0.25**	34.64	97.34	78.49	27.19

the semantic change of the edited object becomes progressively stronger. However, the optical-flow-based metric keeps decreasing. We attribute this to the fact that stronger edits introduce pronounced shape deformations, so warping the edited frame using optical flow estimated from the source video leads to severe misalignment, lowering the WarpSSIM score. In addition, we performed a qualitative ablation analysis of DA-FC and present an ablation study on mask generation in the supplementary materials.

Motion-Appearance Decoupling Flow Correction. Figure 6 illustrates that FlowDirector (Direct ODE) suffers from motion distortion and drift in complex action editing due to accumulated inter-frame errors (red boxes, w/o corr row: the ball incorrectly appears in front), whereas MAD-FC preserves coherent motion even under fast movement with occlusion (red boxes, ours row: the crossover dribble is correctly maintained). Quantitative results in Table 2 show a clear improvement in WarpSSIM computed from the source-video optical flow, confirming that MAD-FC better preserves motion consistency before and after editing. In the supplementary material, we further provide ablations on the MAD-FC parameters ζ (overall guidance strength) and ϕ (appearance alignment strength), offering a detailed analysis of their impact on the final editing quality.

Differential Averaging Guidance. As shown in Figure 7, without DAG the high variance of sampling noise makes the editing trajectory jitter across timesteps, which introduces visible artifacts and texture flicker. With DAG, these issues are largely suppressed, and the textures remain temporally

Figure 6. **Ablation study of Motion-Appearance Decoupling Flow Correction.** Without MAD-FC, the edited video exhibited severe distortion. After using MAD-FC correction, the motion of the edited video remained largely consistent with the original video.

Figure 7. **Ablation study of Differential Averaging Guidance.** In the "Zoom" column: Top row is with DAG, bottom row is without DAG.

stable and consistent with the underlying motion. As reported in Table 2, enabling DAG leads to slight improvements on all metrics. The numerical gains are relatively modest because many perceptual degradations, such as artifacts, texture flicker, and other local inconsistencies, are not fully reflected by these metrics. For example, optical-flow-based WarpSSIM and CLIP-T are often insensitive even when such issues are clearly noticeable to human observers. In the supplementary materials, we provide a detailed analysis of runtime and resource consumption, and propose several optimization and acceleration strategies for DAG.

5. Conclusion

In this paper, we propose FlowDirector, a novel training-free video editing framework that operates without inversion, instead modeling video transformation as a continuous evolution in data space. We propose three training-free flow correction strategies to optimize the editing path. Direction-aware flow correction amplifies components opposite to the source semantics to promote thorough editing. Motion-appearance decoupling flow correction models the motion consistency before and after editing as an optimizable energy term and continuously minimizes it to preserve temporal structure. Differential averaging guidance achieves variance reduction close to that of multiple averaging with fewer averaging operations and improves trajectory stability. Extensive experiments show that FlowDirector achieves state-of-the-art performance in text-driven video editing.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 6250070674) and the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007).

References

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. *Cvpr*, 2023. 3

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *Cvpr*, 2023. 3

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Cvpr*, 2021. 2

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation, 2023. 3

[5] 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. https://arxiv.org/abs/2310.05922v3, 2023. 2, 3, 6, 7

[6] 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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. https://arxiv.org/abs/2403.03206v1, 2024. 3

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. https://arxiv.org/abs/2307.10373v3, 2023. 2, 3, 6, 7

[8] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. *Computer Vision and Pattern Recognition*, 2023. 3

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[11] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, 2020. 2, 3

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *Neural Information Processing Systems*, 2022. 3

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[14] Hyeonho Jeong and Jong Chul Ye. Ground-A-Video: Zero-shot Grounded Video Editing using Text-to-image Diffusion Models. https://arxiv.org/abs/2310.01107v2, 2023. 2

[15] 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. https://arxiv.org/abs/2312.04524v1, 2023. 2, 3, 6, 7

[16] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6007–6017. IEEE, 2023. 3

[17] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. https://arxiv.org/abs/2303.13439v1, 2023. 2

[18] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 2416–2425. IEEE, 2022. 3

[19] Jeongsol Kim, Yeobin Hong, Jonghyun Park, and Jong Chul Ye. Flowalign: Trajectory-regularized, inversion-free flow-based image editing, 2025. 3

[20] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Neural Information Processing Systems*, 2023. 7

[21] 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. 3

[22] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free

text-based editing using pre-trained flow models. *arXiv preprint arXiv: 2412.08629*, 2024. 2, 3, 4

[23] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, Deva Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 2014. 2

[24] Y. Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 2022. 3

[25] Chang Liu, Rui Li, Kaidong Zhang, Yunwei Lan, and Dong Liu. StableV2V: Stablizing Shape Consistency in Video-to-Video Editing. https://arxiv.org/abs/2411.11045v1, 2024. 2

[26] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-P2P: Video Editing with Cross-attention Control. https://arxiv.org/abs/2303.04761v1, 2023. 2, 3

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3, 4

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations*, 2021. 3

[29] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6038–6047. IEEE, 2023. 2, 3

[30] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 6

[31] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *International Conference on Learning Representations*, 2023. 2, 3

[32] 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. https://arxiv.org/abs/2303.09535v3, 2023. 2, 3, 6, 7

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 2021. 7

[34] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *Computer Vision and Pattern Recognition*, 2021. 2, 3

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, 2022. 3

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[37] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sanggil Lee, and Sungroh Yoon. Edit-A-Video: Single Video Editing with Object-Aware Consistency. https://arxiv.org/abs/2303.07945v4, 2023. 2

[38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[40] 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. 2, 3

[41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *European Conference on Computer Vision*, 2020. 7

[42] 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. 3, 6

[43] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope Text-to-Video Technical Report, 2023.

[44] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation, 2024. 3

[45] Yuanzhi Wang, Yong Li, Mengyi Liu, Xiaoya Zhang, Xin Liu, Zhen Cui, and Antoni B. Chan. Re-Attentional Controllable Video Diffusion Editing. https://arxiv.org/abs/2412.11710v1, 2024. 2

[46] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. VideoDirector: Precise Video Editing via Text-to-Video Models. https://arxiv.org/abs/2411.17592v2, 2024. 2, 3, 6, 7

[47] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. https://arxiv.org/abs/2212.11565v2, 2022. 2, 3

[48] Zongze Wu, Nicholas Kolkin, Jonathan Brandt, Richard Zhang, and Eli Shechtman. Turboedit: Instant text-based image editing. *European Conference on Computer Vision*, 2024. 3

[49] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv: 2312.04965*, 2023. 2, 3

[50] Shuheng Zhang, Yuqi Liu, Hongbo Zhou, Jun Peng, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. AdaFlow: Efficient Long Video Editing via Adaptive Attention Slimming And Keyframe Selection. https://arxiv.org/abs/2502.05433v1, 2025. 2