

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

Supplementary Material

This supplementary material provides comprehensive implementation details, in-depth algorithmic descriptions, and extensive experimental analyses to further validate the effectiveness of FlowDirector. We begin by outlining the specific experimental configurations and hyperparameter settings in Section A, followed by the complete inference pseudocode in Section B. Section D details the process of mask generation via cross-attention maps. Subsequently, we present a deeper investigation into our core contributions, including a qualitative ablation analysis of the Direction-Aware Flow Correction in Section E, a parameter study of the Motion-Appearance Decoupling Flow Correction in Section F, and a detailed efficiency analysis of the Differential Averaging Guidance strategy in Section G. Finally, we discuss current limitations in Section H and showcase an extensive gallery of additional qualitative results across diverse editing scenarios in Section I.

A. Detailed Experimental Settings

Our implementation leverages the pre-trained Wan-2.1 1.3B model [6] as the foundational backbone. The editing procedure is executed over a full 50-step denoising trajectory without employing skip sampling strategies. During inference, we disable Classifier-Free Guidance (CFG) for the source video branch, whereas the target video generation utilizes a fixed CFG scale of 10.5. To better align the temporal dynamics, a timestep shift of 12 is applied throughout the sampling process. Regarding the specific hyperparameters of FlowDirector, the Direction-Aware Flow Correction is configured with an amplification factor $\alpha = 0.25$ and a softening coefficient $\lambda = 0.25$. For mask generation, we apply average pooling with a kernel size of 9 to ensure boundary smoothness. The Motion-Appearance Decoupling Correction adapts its regularization strength (ζ, ϕ) according to the editing magnitude: we assign $(\zeta, \phi) = (0.01, 0.3)$ for edits involving significant motion changes, and $(0.007, 0.5)$ for those with subtler dynamics. Furthermore, the Differential Averaging Guidance (DAG) is computed using a guidance weight of $\omega = 2.75$, deriving a robust velocity estimate from $L_{HQ} = 3$ stochastic samples and establishing the baseline from the $K = 2$ candidates with the lowest cosine similarity. To stabilize the final output details, the noise realization remains frozen during the last eight denoising steps.

B. FlowDirector Inference Algorithm

We present the complete inference procedure of FlowDirector in Algorithm 1. Our framework operates in an inversion-

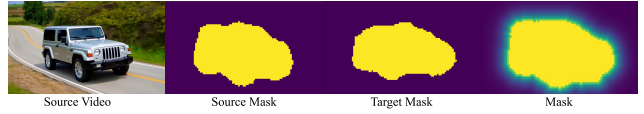


Figure 1. The final mask is obtained by merging the Source Mask and the Target Mask and softening the edges. Lighter colors indicate lower attention values.

free manner, initializing the denoising trajectory directly from the source video X^{src} . At each timestep t , the update process integrates three core strategies: (1) **Direction-Aware Flow Correction (DA-FC)** is applied during the candidate generation phase, where we amplify anti-parallel flow components to facilitate structural changes; (2) **Differential Averaging Guidance (DAG)** aggregates these corrected candidates to estimate a robust editing velocity v_{edit} , utilizing a baseline formed by high-variance samples to reduce trajectory jitter; and (3) **Motion-Appearance Decoupling Flow Correction (MAD-FC)** rectifies the state update. Specifically, we estimate the clean data states \hat{x}_0 and their temporal averages (anchors A_h) to enforce motion consistency while allowing appearance changes via the parameter ϕ . The final state Z_{i-1} is updated by combining the masked editing flow with this decoupling correction term.

Table 1. Same-backbone comparison on Wan2.1-1.3B.

Method	Pick (%) \uparrow	CLIP-T \uparrow	CLIP-F \uparrow	WarpSSIM \uparrow	$Q_{\text{edit}}\uparrow$
FLATTEN (+Wan2.1)	21.35	<u>34.12</u>	95.18	<u>78.01</u>	<u>26.65</u>
TokenFlow (+Wan2.1)	<u>21.43</u>	33.72	95.73	76.85	25.92
VideoDirector (+Wan2.1)	21.28	33.65	<u>96.52</u>	77.35	26.08
FlowDirector	21.82	34.64	97.34	78.49	27.19

Table 2. Quantitative comparison with the v2v frameworks.

	Method	Pick (%) \uparrow	CLIP-T \uparrow	WarpSSIM \uparrow	$Q_{\text{edit}}\uparrow$
Training-free	FRESCO-V2	20.66	31.82	72.96	23.33
	FlowDirector	<u>21.82</u>	34.64	<u>78.49</u>	<u>27.19</u>
Training-based	VACE	22.43	<u>34.35</u>	80.74	27.81

C. Extended Comparisons

To eliminate the influence of different foundation models and enable a fairer comparison, we re-implement the compatible components of FLATTEN [1], TokenFlow [2], and VideoDirector [7] on the same Wan2.1-1.3B backbone [6]. In this way, the performance gap can be attributed more directly to the editing framework itself rather than to differences in the underlying generative model. As shown in Table 1, FlowDirector consistently achieves the best re-

Algorithm 1 FlowDirector Inference Algorithm

Input: Source X^{src} , Prompts $c_{\text{src}}, c_{\text{tar}}$, Steps N , Hyperparams $\alpha, \zeta, \omega, L_{\text{HQ}}, K$.

Output: Edited Video Z_0 .

```
1:  $Z_N \leftarrow X^{\text{src}}$ 
2: for  $i = N$  to 1 do
3:    $t \leftarrow t_i, \Delta t \leftarrow t_{i-1} - t_i$ 
4:    $\mathcal{V} \leftarrow \emptyset$ 
5:   for  $k = 1$  to  $L_{\text{HQ}}$  do
6:     Sample  $\epsilon_k \sim \mathcal{N}(0, I)$  to obtain  $z_t^{\text{src}}, z_t^{\text{tar}}$ 
7:      $v_{\text{src}} \leftarrow v_{\theta}(z_t^{\text{src}}, t, c_{\text{src}}), v_{\text{tar}} \leftarrow v_{\theta}(z_t^{\text{tar}}, t, c_{\text{tar}})$ 
8:      $v_{\text{raw}} \leftarrow v_{\text{tar}} - v_{\text{src}}$ 
9:      $v_{\parallel} \leftarrow \frac{\langle v_{\text{raw}}, v_{\text{src}} \rangle}{\|v_{\text{src}}\|^2} v_{\text{src}}, v_{\perp} \leftarrow v_{\text{raw}} - v_{\parallel}$ 
10:     $M_{\text{opp}} \leftarrow \mathcal{K}(\langle v_{\text{raw}}, v_{\text{src}} \rangle < 0)$  // Element-wise indicator
11:     $\tilde{v}_k \leftarrow v_{\perp} + M_{\text{opp}} \odot (1 + \alpha)v_{\parallel}$ 
12:     $\mathcal{V} \leftarrow \mathcal{V} \cup \{\tilde{v}_k\}$ 
13:   end for
14:    $v_{\text{HQ}} \leftarrow \frac{1}{L_{\text{HQ}}} \sum_{\tilde{v} \in \mathcal{V}} \tilde{v}$ 
15:    $S_k \leftarrow \frac{\langle \tilde{v}_k, v_{\text{HQ}} \rangle}{\|\tilde{v}_k\| \|v_{\text{HQ}}\|} \forall \tilde{v}_k \in \mathcal{V}$ 
16:    $\mathcal{I}_K \leftarrow$  Indices of  $K$  smallest values in  $S$ 
17:    $v_{\text{BL}} \leftarrow \frac{1}{K} \sum_{k \in \mathcal{I}_K} \tilde{v}_k$ 
18:    $v_{\text{edit}} \leftarrow v_{\text{HQ}} + \omega \cdot (v_{\text{HQ}} - v_{\text{BL}})$ 
19:    $\hat{x}_0^{\text{src}} \leftarrow Z_i - t \cdot v_{\theta}(Z_i, t, c_{\text{src}})$  // (C, T, H, W)
20:    $\hat{x}_0^{\text{tar}} \leftarrow Z_i - t \cdot v_{\theta}(Z_i, t, c_{\text{tar}})$  // (C, T, H, W)
21:    $A_h^{\text{src}} \leftarrow \text{Mean}_T(\hat{x}_0^{\text{src}}), A_h^{\text{tar}} \leftarrow \text{Mean}_T(\hat{x}_0^{\text{tar}})$ 
22:    $\hat{v}_{\text{final}} \leftarrow v_{\text{edit}} \odot \text{Mask}(c_{\text{src}}, c_{\text{tar}})$ 
23:    $Z_{i-1} \leftarrow Z_i + \Delta t \cdot \hat{v}_{\text{final}} - \zeta [\hat{x}_0^{\text{tar}} - \hat{x}_0^{\text{src}} - \phi(A_h^{\text{src}} - A_h^{\text{tar}})]$ 
24: end for
25: return  $Z_0$ 
```

sults across all evaluation metrics. These results show that our method maintains clear advantages even under the same backbone setting, demonstrating that the gains of FlowDirector come from the proposed framework and flow correction strategies rather than from a stronger foundation model.

We further compare FlowDirector with two recent video-to-video editing frameworks, FRESCO V2 [8] and VACE [3], in Figure 2 and Table 2. The results show that FlowDirector consistently outperforms FRESCO V2 across all metrics and remains competitive with VACE, despite the latter being supported by extensive training. This comparison suggests that FlowDirector is not only effective against methods adapted to the same backbone, but also highly competitive among broader video-to-video editing frameworks. Taken together, these results further support the effectiveness of our design in improving editing alignment, temporal consistency, and overall edit quality.



Figure 2. Qualitative comparison with v2v frameworks.

D. Mask Generation via Cross-Attention Maps

Although the editing flow V_{edit} effectively drives the semantic transformation, applying it globally can cause unintended modifications in the background. To address this, we construct an explicit spatial mask by leveraging the intrinsic localization capabilities of Diffusion Transformers (DiTs). By visualizing the cross-attention maps across different network depths (see Figure 4 and Figure 5), we empirically identified the **18th block** as the optimal block. The maps extracted from this block exhibit high activation concentrations on object structures, enabling us to cleanly separate editable regions from the background.

Attention Extraction and Aggregation. During the denoising step t , we perform a forward pass with the source prompt c_{src} and extract the cross-attention map $\mathbf{A} \in \mathbb{R}^{B \times N_h \times L_{\text{vis}} \times L_{\text{text}}}$. Here, N_h denotes the number of attention heads, while L_{vis} and L_{text} represent the visual and textual token counts. Given a set of indices S corresponding to the key editing tokens in the prompt (e.g., “jeep”), we aggregate the attention scores to obtain a consolidated spatial map $\mathbf{a} \in \mathbb{R}^{L_{\text{vis}}}$:

$$\mathbf{a}_i = \frac{1}{N_h |S|} \sum_{h=1}^{N_h} \sum_{j \in S} \mathbf{A}_{i,j}^{(h)}. \quad (1)$$

This vector \mathbf{a} is then reshaped into a spatiotemporal patch grid $\mathcal{G} \in \mathbb{R}^{F_p \times H_p \times W_p}$.

Mask Construction and Refinement. To generate a robust binary mask \mathbf{M}_{src} , we process \mathcal{G} through the following steps:

- Spatial Smoothing:** We apply 2D average pooling with a **kernel size of 9** to each frame in \mathcal{G} . This mitigates high-frequency noise in the raw attention maps.
- Upsampling:** The smoothed map is upsampled to the original video resolution via trilinear interpolation and replicated along the channel dimension to form $\mathcal{A} \in \mathbb{R}^{C \times T \times H \times W}$.
- Binarization:** We perform a numerically stable normalization $\hat{\mathcal{A}} = \mathcal{A} / (\max(\mathcal{A}) + \epsilon)$ and derive a binary mask using an adaptive threshold $\tau = \text{mean}(\hat{\mathcal{A}})$:

$$(\mathbf{M}_{\text{src}})_{c,t}(x,y) = \begin{cases} 1, & \text{if } \hat{\mathcal{A}}_{c,t}(x,y) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

An analogous procedure yields the target mask \mathbf{M}_{tar} from the target prompt c_{tar} . The comprehensive editing region is defined as the union $\mathbf{M} := \mathbf{M}_{\text{src}} \cup \mathbf{M}_{\text{tar}}$.

Soft Blending. To ensure a seamless transition between the edited object and the background, we soften the binary mask using a Euclidean distance transform $d_{c,t}(x, y)$ computed on the background region. The final soft mask $\widetilde{\mathbf{M}}$ is formulated as:

$$\widetilde{\mathbf{M}}_{c,t}(x, y) = \mathbf{M}_{c,t}(x, y) + (1 - \mathbf{M}_{c,t}(x, y)) e^{-\lambda d_{c,t}(x, y)}. \quad (3)$$

We set the decay rate $\lambda = 0.25$. This mask directly modulates the editing velocity field, freezing irrelevant regions while preserving the structural integrity of the edited subject:

$$\widetilde{V}_{\text{edit}} = V_{\text{edit}} \odot \widetilde{\mathbf{M}}. \quad (4)$$



Figure 3. **Ablation study of Direction-Aware Flow Correction.** Without DA-FC, it is difficult to achieve an effective balance between editing strength and consistency. In contrast, incorporating DA-FC enables significant semantic modifications of the target object while effectively preserving irrelevant regions and maintaining motion consistency. Our modules are crucial for achieving high-quality video editing.

E. Qualitative Ablation Analysis Of Direction-Aware Flow Correction

We conduct a qualitative analysis to investigate the impact of Direction-Aware Flow Correction (DA-FC) on editing fidelity and structural integrity. As illustrated in Figure 3, relying solely on the basic FlowDirector (Direct ODE) presents a dilemma. When configured without skip steps, the direct integration accumulates errors along the path, leading to severe deviations in both appearance and motion consistency compared to the source video. Conversely, employing skip steps to mitigate this drift imposes excessive constraints on the generative trajectory, which heavily restricts the editing magnitude and fails to produce significant structural changes.

Consequently, without DA-FC, the editing process struggles to strike an effective balance between semantic transformation and content preservation. Our Direction-Aware Flow Correction resolves this by intervening at the velocity level: it amplifies the anti-parallel components essential

for structural editing while suppressing the parallel components that contribute to drift. This enables FlowDirector to achieve robust semantic transformations without the consistency degradation seen in the full ODE or the conservative limitations of skip sampling.

F. Ablation Study of Motion-Appearance Decoupling Flow Correction

The Motion-Appearance Decoupling Flow Correction (MAD-FC) module serves a critical role in balancing source motion fidelity with target appearance transformation. This balance is governed by two key hyperparameters: the appearance anchor coefficient ϕ and the overall correction strength ζ . In this section, we analyze their individual impacts on editing quality based on our empirical observations.

We first examine the influence of the appearance anchor coefficient ϕ . This parameter regulates the adherence to the source video’s visual attributes. As illustrated in Figure 10, setting ϕ to an excessively high value (*e.g.*, $\phi = 2$) imposes rigid constraints derived from the source appearance anchors. Consequently, the generated output remains visually nearly identical to the original bear, effectively suppressing the desired semantic transformation. Even at $\phi = 1$, the result retains significant bear-like morphological features, such as the head shape and fur texture. Conversely, lowering ϕ relaxes these constraints, allowing the target semantics to manifest. We observe that $\phi = 0.3$ strikes an optimal balance, facilitating a successful morphological transformation into a dinosaur—characterized by changes in skin texture and body structure—while preserving the underlying walking motion of the original video.

Next, we investigate the correction strength ζ , which determines the intensity of the motion consistency enforcement. As shown in Figure 11, a low correction strength (*e.g.*, $\zeta = 0.003$) provides insufficient guidance to counteract the stochastic variance of the diffusion process. This results in slight pose misalignments and temporal instability, where the edited character fails to strictly follow the source motion. Increasing ζ to 0.01 significantly improves alignment, ensuring the edited subject’s posture and trajectory align precisely with the source video. Based on these findings, we utilize a configuration of ($\zeta = 0.01, \phi = 0.3$) for edits involving complex structural changes, and adjust to (0.007, 0.5) for milder motion scenarios to prioritize stability.

G. Detailed Analysis of Differential Averaging Guidance

A conventional averaging strategy performs multiple rounds of iterative inference to obtain several editing flows, each corresponding to a different editing direction. These flows

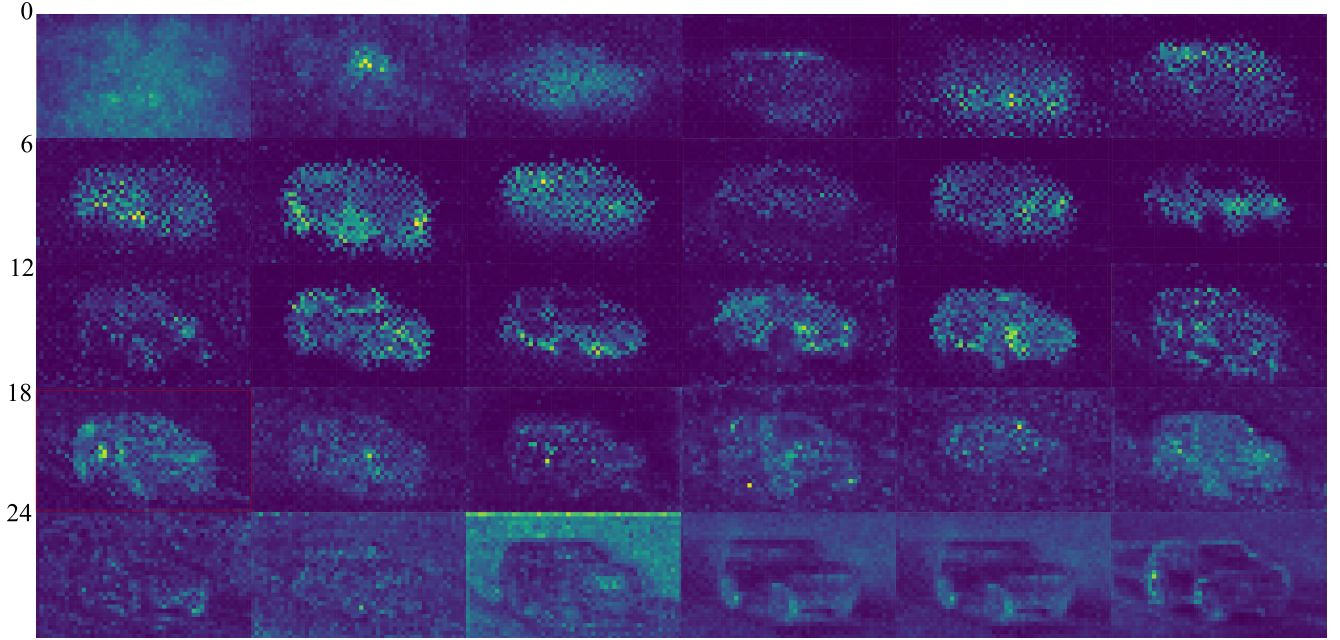


Figure 4. Visualization of the cross-attention maps of the keyword “jeep” in the source prompt across different DiT blocks. The attention map of the 18th block clearly outlines the shape of the jeep.

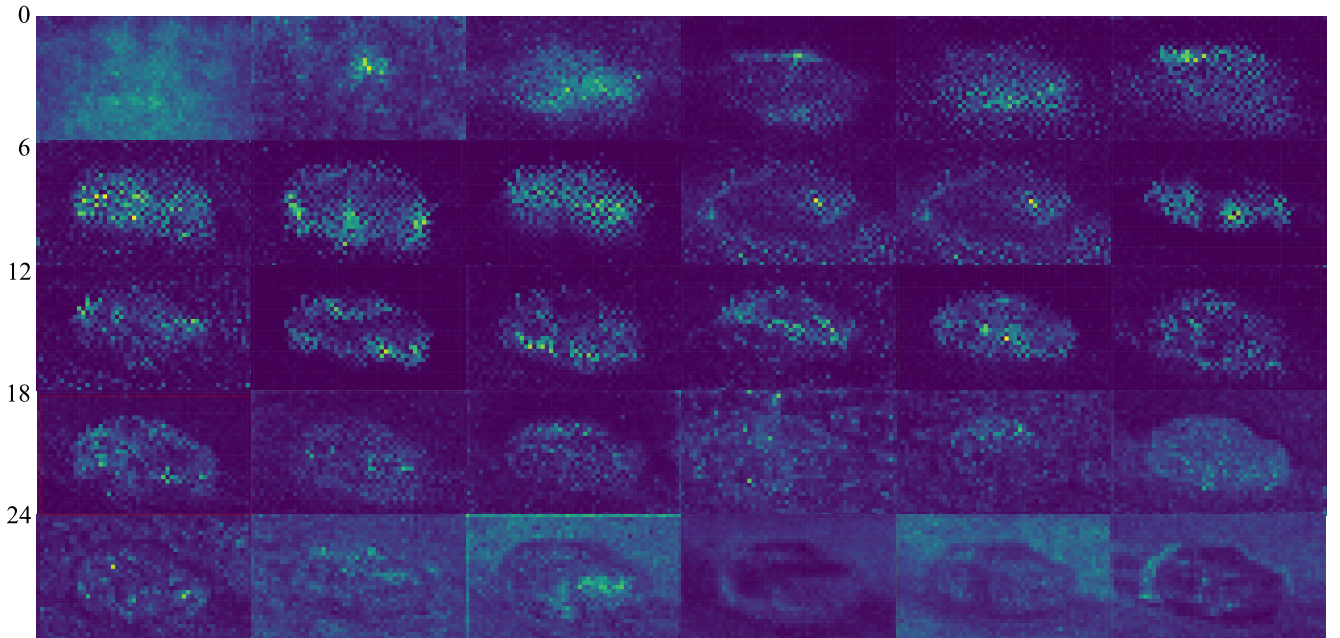


Figure 5. Visualization of the cross-attention maps of the keyword “Porsche car” in the source prompt across different DiT blocks. The attention map of the 18th block clearly outlines the car.

are then averaged to produce a more robust and consolidated direction, which is subsequently used to update the video features. In this section, we compare our Differential Averaging Guidance (DAG) with this conventional averaging strategy to demonstrate both the effectiveness and efficiency

of DAG. As illustrated in Figure 6, when editing the source video (*e.g.*, a bear) into the target video (*e.g.*, a dinosaur), the regions undergoing substantial semantic changes are mainly concentrated on the bear’s back. Without any dedicated guidance strategy, noticeable artifacts and inter-frame

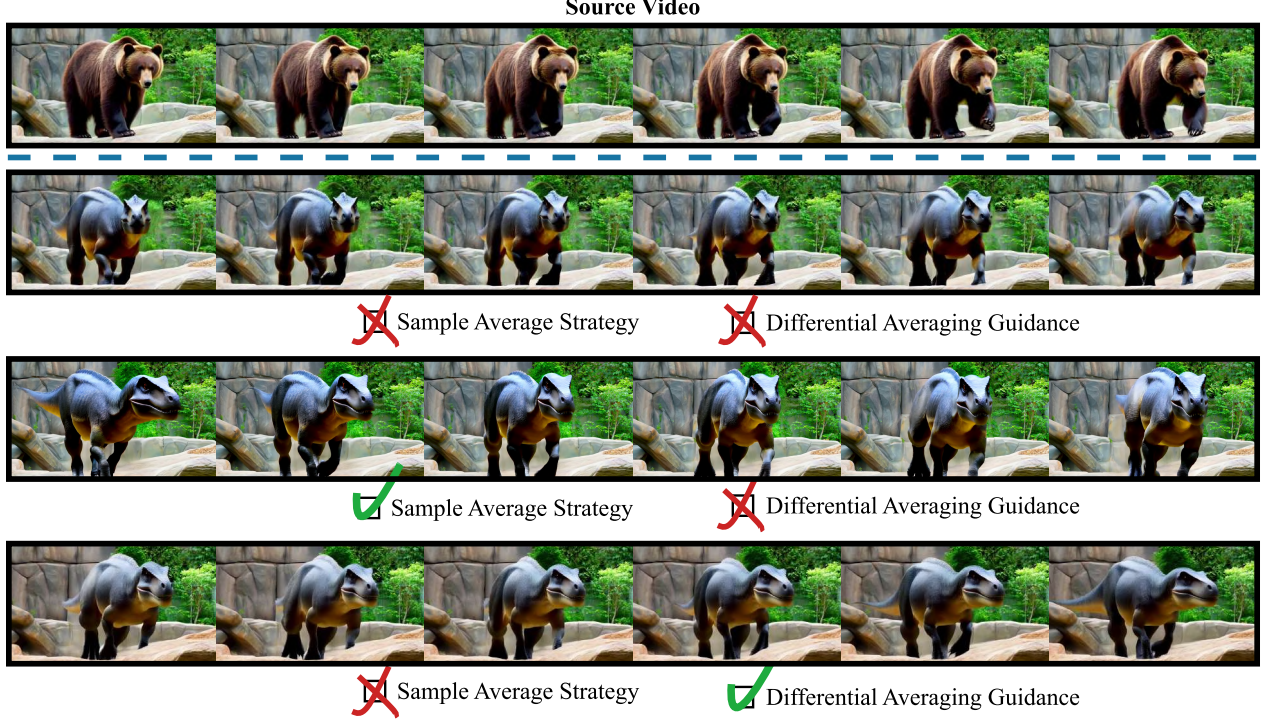


Figure 6. **Qualitative comparison between the editing results of a multi-round inference averaging strategy and using a DAG.** The Sample Average strategy is set to use a regular averaging strategy for 20 rounds of iterative inference at every denoising step to obtain the editing flow. The DAG setting uses 3 rounds of iterative inference to obtain a high-quality estimate and perform reinforcement-guided generation of the editing flow. Best viewed zoomed in.

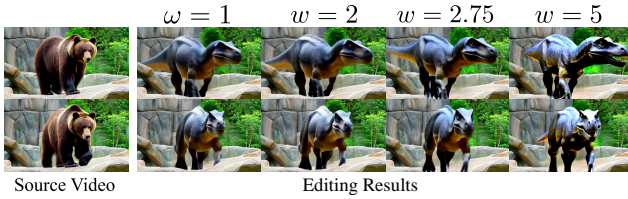


Figure 7. **Ablation study of guidance strength ω .** The ω controls the guidance strength of the differential signal. By enhancing the differential signal, artifacts can be effectively eliminated and the editing results can be optimized. We use $\omega = 2.75$ as the default value. Best viewed zoomed in.

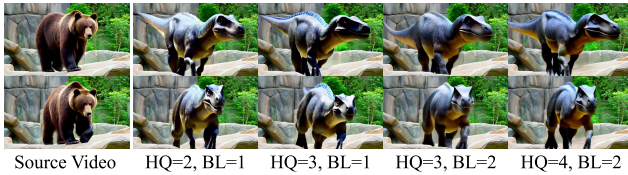


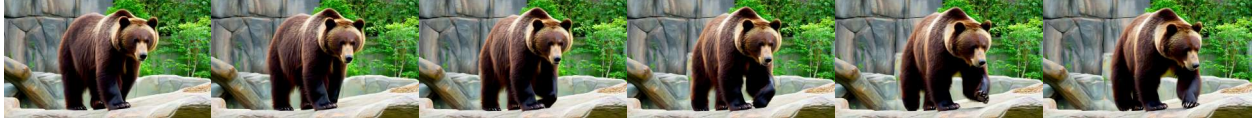
Figure 8. By taking different averages to construct high-quality estimates and baseline estimates, different guidance enhancement effects are produced. Best viewed zoomed in.

texture flickering appear in this region (second row of Fig-

Table 3. **Efficiency Comparison.** We report the inference time and peak GPU memory usage for editing a 41-frame video on a single NVIDIA H800 80G GPU. The upper section compares existing SOTA methods, while the lower section analyzes the efficiency of different strategies within our framework. The symbol “-” indicates cases where the method exceeded the single-GPU memory limit and required specific optimization strategies to execute; consequently, these metrics are omitted to ensure a fair comparison of native performance.

Method	Editing Time	GPU Memory
FateZero [5]	-	-
FLATTEN [1]	6 min 13s	44.7GB
TokenFlow [2]	1 min 15s	29.7GB
RAVE [4]	5min 46s	26.4GB
VideoDirector [7]	-	-
w/o DAG	57s	18 GB
Conventional Averaging	19 min 3s	≈18 GB
w/ DAG	2 min 54s	≈18 GB

ure 6). Although these defects are visually salient, they are not adequately reflected by standard quantitative metrics, which are often insensitive to localized artifacts and tem-



A large brown **bear** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery. The camera remains fixed, capturing the **bear**'s deliberate movements.



A large **dinosaur** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery. The camera remains fixed, capturing the **dinosaur**'s deliberate movements.



A large **dinosaur** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery. The camera remains fixed, capturing the **bear**'s deliberate movements.

Figure 9. **An example of editing failure due to incomplete target text replacement.** When attempting to edit a “bear” into a “dinosaur,” if the target prompt erroneously retains descriptions of the “bear” (e.g., “...capturing the bear’s deliberate movements” instead of a full replacement with dinosaur-related descriptions), the edited video exhibits significant residual features of the original “bear.”

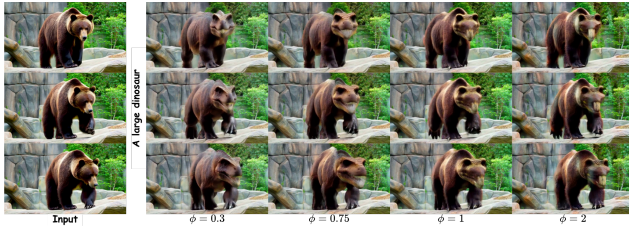


Figure 10. **Ablation study on the appearance anchor coefficient ϕ .** We fix $\zeta = 0.003$ and vary ϕ in the “bear \rightarrow dinosaur” task. A very high ϕ (2.0) imposes excessive source constraints, causing the result to revert to the original bear appearance. As ϕ decreases, the constraints relax, allowing the dinosaur features to emerge. $\phi = 0.3$ successfully achieves the semantic transformation while maintaining the original motion pattern.

poral texture instability. To quantify this effect, we compute the CLIP similarity between the edited videos and two defect-related prompts, namely “artifact” and “distortion”. As shown in Table 4, enabling DAG consistently reduces the similarity to both prompts, decreasing the score from 0.2404 to 0.2012 for “artifact” and from 0.2363 to 0.2115 for “distortion.” These results provide quantitative support for the qualitative observations and confirm that DAG effectively suppresses perceptually disturbing artifacts.

We further compare DAG with the conventional averaging strategy. When conventional averaging is applied using the results from twenty rounds of iterative inference

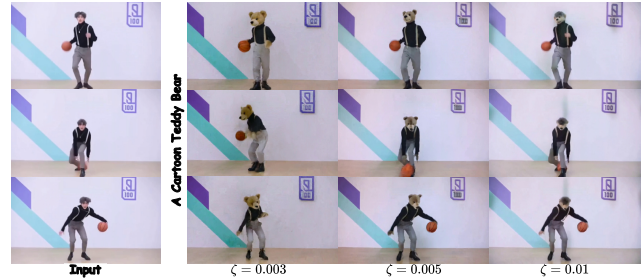


Figure 11. **Ablation study on the correction strength ζ .** We fix $\phi = 0.3$ and vary ζ . Increasing ζ from 0.003 to 0.01 significantly improves motion alignment with the source input, ensuring that the edited character strictly follows the original movements.

Table 4. CLIP similarity to defect-related prompts (lower is better).

Setting	Artifact↓	Distortion↓
w/o DAG	0.2404	0.2363
w/ DAG	0.2012	0.2115

(third column of Figure 6), the artifacts are largely alleviated. However, this improvement comes at a substantial computational cost. As reported in Table 3, editing a 41-frame video requires about 19 minutes on a single NVIDIA H800 80G GPU.

In contrast, our DAG approach requires only three rounds of iterative inference to obtain a high-quality estimation, which is then used to construct a differential signal that guides and reinforces the editing flow. As shown in Fig-

ure 6 (fourth row), DAG not only removes artifacts more effectively than the twenty-round averaging strategy, but also better resolves incomplete edits, yielding a dinosaur with a morphology more clearly separated from the original bear. Meanwhile, it edits the same 41-frame video in only 2 minutes and 54 seconds on a single NVIDIA H800 80G GPU, with comparable memory usage. These results demonstrate that DAG is both more effective and substantially more efficient than conventional averaging.

Effect of the Guidance Strength ω . To determine the optimal guidance strength ω , we conducted a comprehensive ablation study. The primary objective of this analysis was to investigate the impact of varying ω values on the quality of the video editing results. As illustrated in Figure 7, our experiments reveal a clear correlation between the guidance strength and the final output. Specifically, as the value of ω increases, a noticeable reduction in artifacts from the source video is observed. Concurrently, the semantic deformations in the edited output become more pronounced and accurate. For instance, the morphological structure of the dinosaur in our test case undergoes a more significant and semantically appropriate transformation with a higher ω . However, our study also indicates that an excessively high guidance strength can be detrimental. When ω surpasses a certain threshold, the model begins to introduce unnatural color shifts and motion inconsistencies, which degrade the overall quality of the edited video.

Through a systematic process of experimental analysis and evaluation, we identified $\omega = 2.75$ as the optimal value. This specific setting strikes a balance between artifact suppression and meaningful semantic deformation. At this guidance strength, the model effectively eliminates visual artifacts while producing edits that are semantically coherent and visually compelling, thereby yielding superior editing outcomes.

Effect of high-quality estimates and baseline estimates. We construct high-quality and baseline estimations by averaging the results of multiple inference runs, a strategy that effectively refines the output. As illustrated in Figure 8, we evaluated configurations with varying ensemble sizes, specifically $(L_{HQ}, K) \in \{(2, 1), (3, 1), (3, 2), (4, 2)\}$. Our analysis reveals that as the number of averaging iterations for both estimations increases, there are discernible improvements in the final output, particularly in aspects such as color fidelity and overall appearance. Based on these observations, we have standardized our experimental protocol to the (3, 2) setting. Consequently, the high-quality estimation is generated by averaging the results of three separate inference runs, while the baseline estimation is derived from the average of two of these runs.

Inference Acceleration Strategies. To further enhance the computational efficiency of FlowDirector, we explored several optimization strategies regarding the guidance mechanism. First, we investigated the necessity of applying Classifier-Free Guidance (CFG) to the source generation branch. Our empirical analysis indicates that employing CFG on the source video yields negligible perceptual differences in the final editing results compared to using standard text conditioning alone. Consequently, we adopted an asymmetric guidance strategy: we disable CFG for the source branch (utilizing only text-conditional generation) while retaining it exclusively for the target branch. This reduction effectively halves the computational load for the source velocity estimation. Furthermore, we observe that FlowDirector is compatible with orthogonal acceleration techniques designed for diffusion models. For instance, caching the CFG residual (the difference between conditional and unconditional noise predictions) for the target branch can be seamlessly integrated into our framework, providing further reductions in inference time without compromising editing performance.

H. Limitation

Our method aims to construct a direct editing path from the source video to the target video, bypassing the inversion process, which is prone to structural loss. Since the primary driving force for this direct editing path stems from the discrepancy between the source and target texts, varying degrees of textual difference can lead to markedly different editing outcomes. This results in incomplete text replacement, which causes substantial remnants of the original video content (Figure 9). For example, modifying the source prompt c_{src} (*i.e.*, “A large brown **bear** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery. The camera remains fixed, capturing the **bear**’s deliberate movements.”) to an incompletely substituted target prompt c_{tar} (*i.e.*, “A large **dinosaur** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery. The camera remains fixed, capturing the **bear**’s deliberate movements.”) leads to significant residual “bear” information in the edited video. Furthermore, we observe that the quality of the source text c_{src} also substantially affects the editing results; more comprehensive source texts tend to yield better editing outcomes compared to simpler prompts.

Similarly, our method excels in structure preservation, which is evident in tasks such as significant object editing, texture replacement, object addition/deletion, or compositional tasks. However, its performance in video style transfer is relatively limited. We attribute this to a combination of its tendency towards result preservation and being less driven by textual differences.

I. More Qualitative Results

In this section, we present additional qualitative results to further demonstrate the effectiveness and high quality of our video editing method. Figures 12 to 16 provide further examples of our method performing precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. These results consistently exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Figure 17 also shows some editing results using Wan 2.1 14B [6], achieving higher editing quality and better consistency compared to the 1.3B model.

References

- [1] 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. 1, 5
- [2] 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. 1, 5
- [3] 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*, pages 17191–17202, 2025. 2
- [4] 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. 5
- [5] 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. 5
- [6] 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, Jingen Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghai 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, Wenten 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. 1, 8
- [7] 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. 1, 5
- [8] Shuai Yang, Junxin Lin, Yifan Zhou, Ziwei Liu, and

Chen Change Loy. Zero-shot video translation and editing with frame spatial-temporal correspondence, 2025. 2

A large brown bear is walking slowly across a rocky terrain in a zoo enclosure.



A large tiger is walking slowly across a rocky terrain in a zoo enclosure.



A large dinosaur is walking slowly across a rocky terrain in a zoo enclosure.



A horse walking slowly across a rocky terrain in a zoo enclosure.



A cute and adorable fluffy puppy wearing a witch hat in a halloween autumn evening forest.



A cute and adorable fluffy chinchilla wearing a witch hat in a halloween autumn evening



A cute and adorable fluffy cat wearing a witch hat in a halloween autumn evening forest.



A cute and adorable fluffy koala wearing a witch hat in a halloween autumn evening forest.



Figure 12. **More Qualitative Results.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.

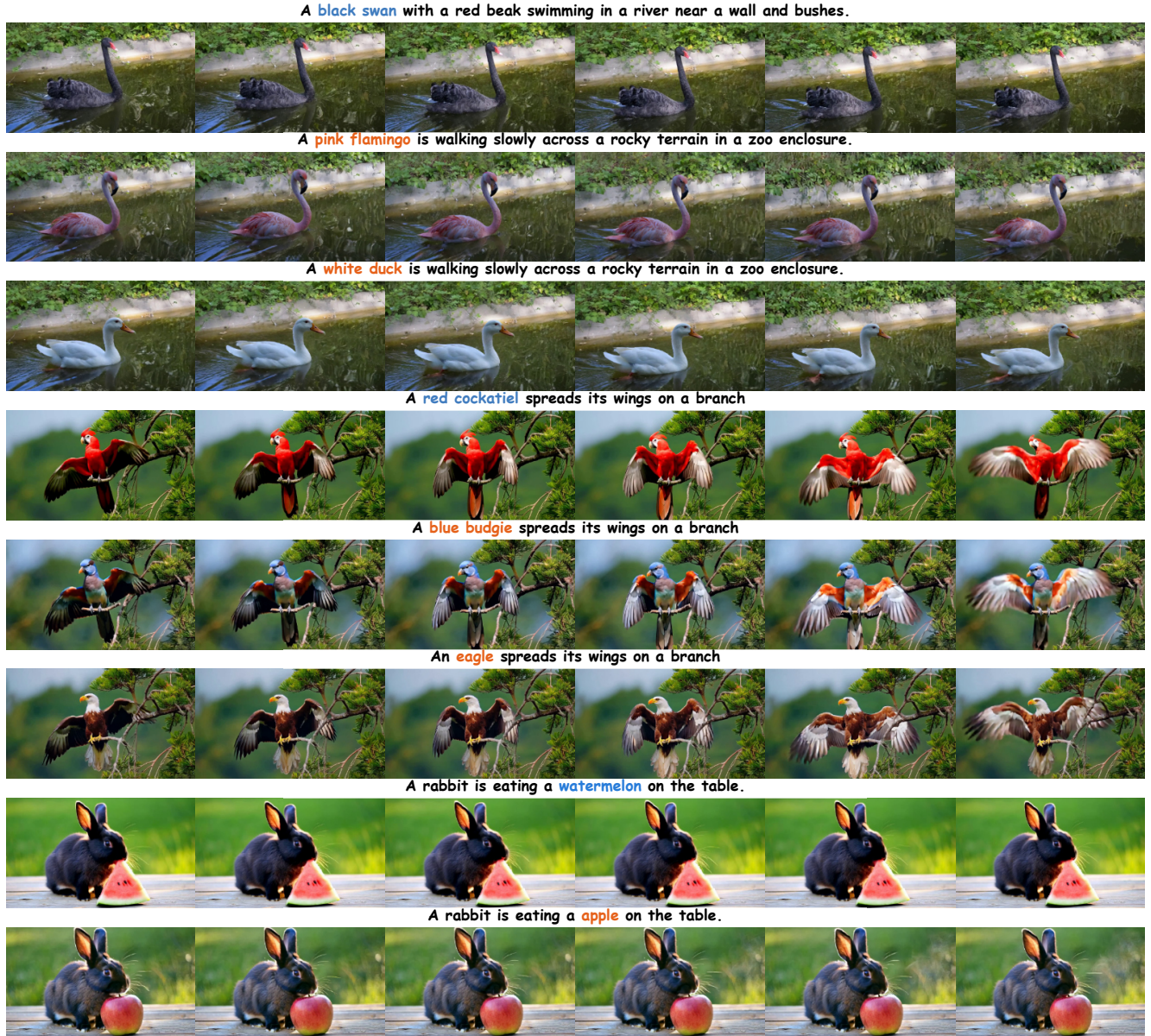


Figure 13. **More Qualitative Results.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.

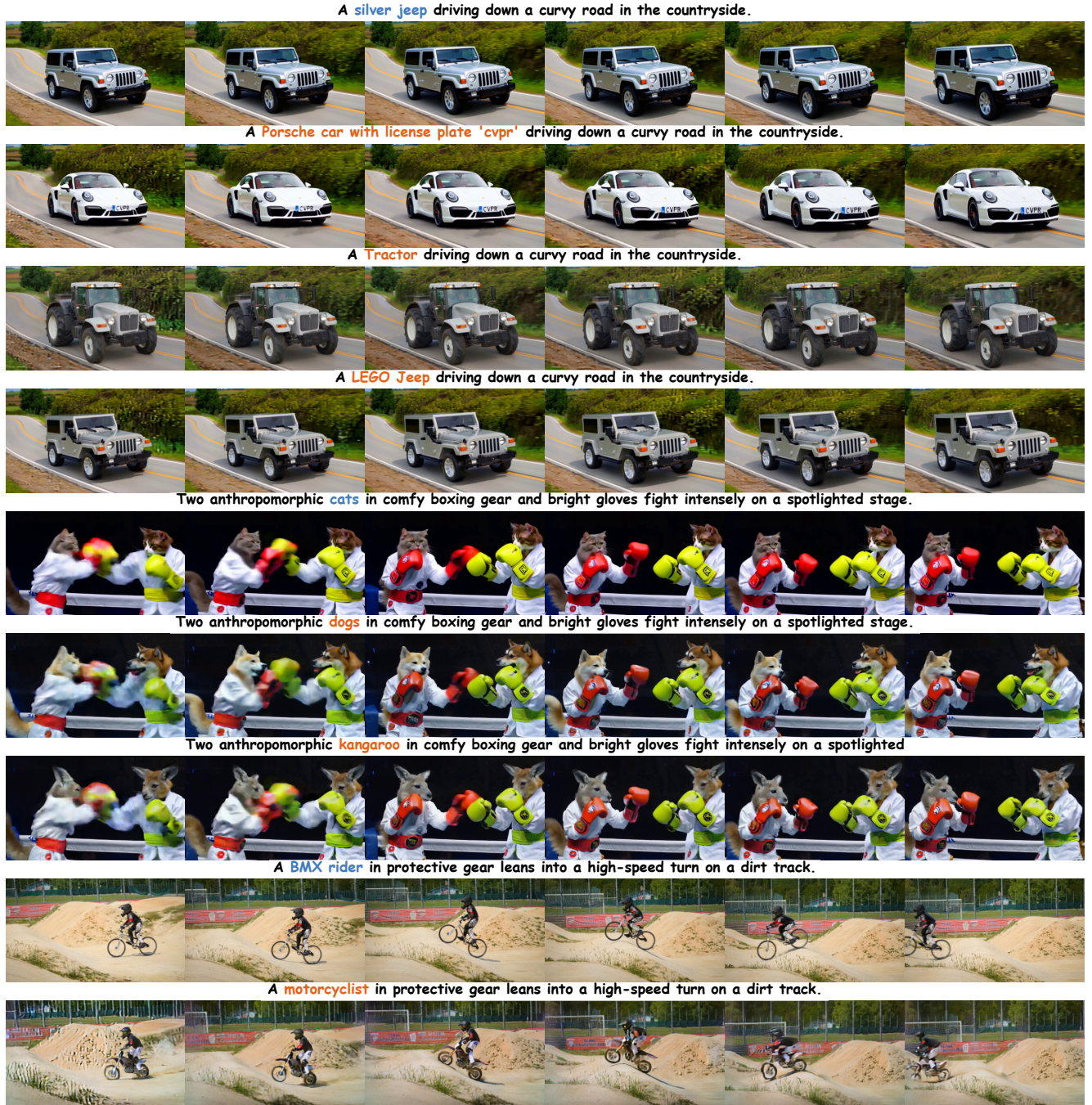


Figure 14. **More Qualitative Results.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.

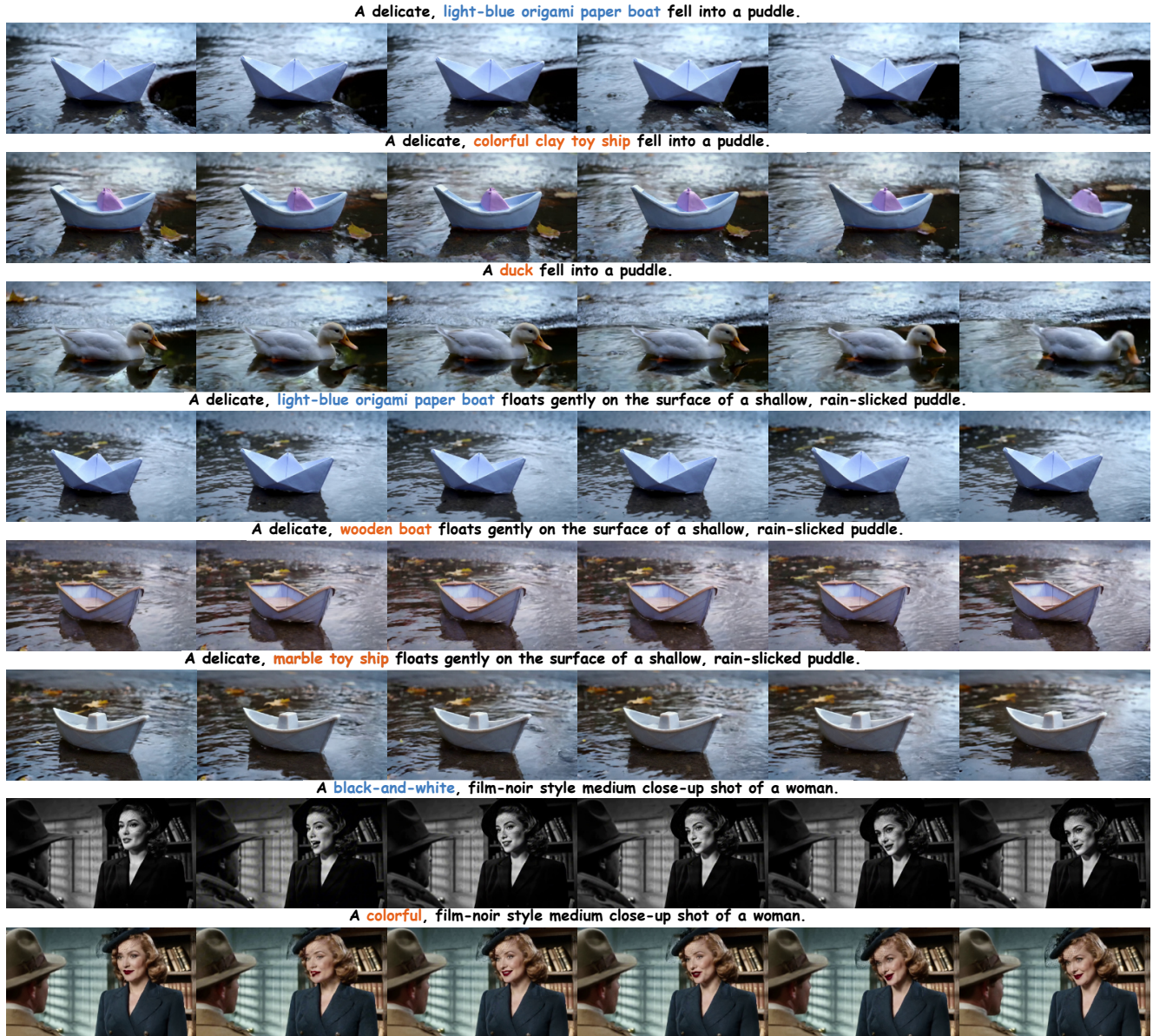


Figure 15. **More Qualitative Results.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.



Figure 16. **More Qualitative Results.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.

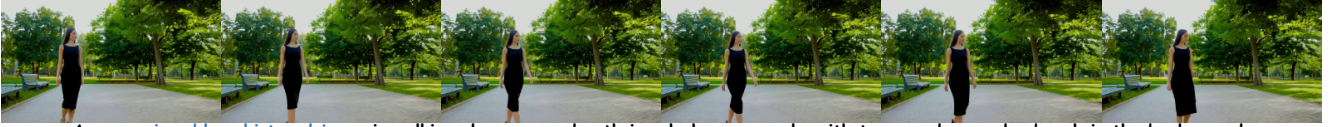
A delicate, **light-blue origami paper boat** floats gently on the surface of a shallow, rain-slicked puddle.



A delicate, **LEGO toy ship** floats gently on the surface of a shallow, rain-slicked puddle.



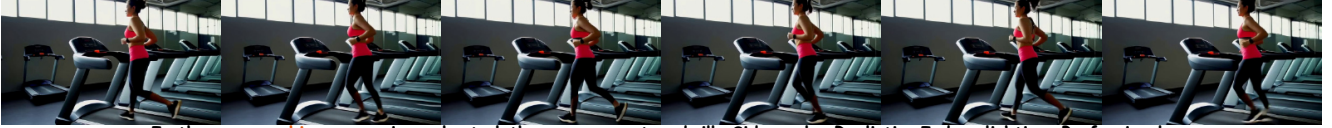
A woman in a **black dress** is walking along a paved path in a lush green park, with trees and a wooden bench in the background.



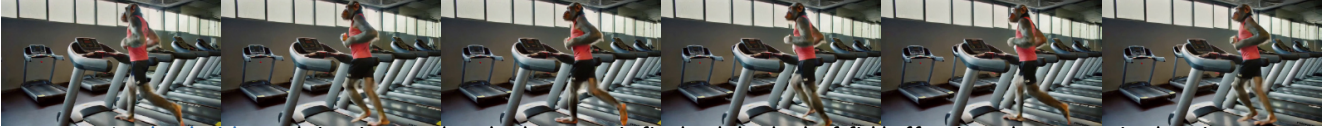
A woman in a **blue shirt and jeans** is walking along a paved path in a lush green park, with trees and a wooden bench in the background.



In the gym, a **woman** in workout clothes runs on a treadmill. Side angle. Realistic, Indoor lighting, Professional.



In the gym, a **chimpanzee** in workout clothes runs on a treadmill. Side angle. Realistic, Indoor lighting, Professional.



A **red cockatiel** spreads its wings on a branch, the camera is fixed and the depth of field effect is used to capture its dynamic



An **eagle** spreads its wings on a branch, the camera is fixed and the depth of field effect is used to capture its dynamic movement.



A **large brown bear** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery.



A **large dinosaur** is walking slowly across a rocky terrain in a zoo enclosure, surrounded by stone walls and scattered greenery.



Figure 17. **Qualitative Results for Wan 14B.** Our method performs precise and semantically faithful edits while preserving the spatial content and motion dynamics of unedited regions. The results exhibit strong alignment with the editing instructions, high visual fidelity, and consistent temporal coherence across frames. Best viewed zoomed-in.