Dual-PAG: Bidirectional Perturbed-Attention Guidance for Residue-Free Video Editing

Jeongseon Oh [*1] **Suwoong Yeom** [*1] **ChangHee Yang** [*1] **Suk-Ju Kang** [†1]

Abstract

Text-driven video editing aims to modify desired semantic content while preserving the motion, layout, and temporal coherence of a source video. However, velocity-based editing methods often leave source residue when the target edit requires a large object-level transformation. To address this issue, we propose **Dual-PAG**, a bidirectional Perturbed Attention Guidance framework that suppresses source-side structural traces while reinforcing target-side geometry. Dual-PAG further uses cross-attention-based masking to localize the correction to the edit region. Dual-PAG preserves the single-trajectory nature of inversion-free video editing while more effectively reducing source residue in challenging object replacement scenarios.

1. Introduction

Recent advances in large-scale video generation models (Wan et al., 2025; Liu et al., 2024b; Kong et al., 2024; Blattmann et al., 2023) have enabled increasingly realistic and temporally coherent video synthesis, making text-driven video editing an important research problem for modifying existing videos with natural-language instructions. Given a source video, a source prompt, and a target prompt, the goal is to modify the source video according to the target prompt while preserving the source motion, layout, and temporal coherence. Among recent inversion-free (Kulikov et al., 2025; Liu et al., 2022; Lipman et al., 2022; Ho & Salimans, 2022; Wu et al., 2024; Cong et al., 2024) approaches, FlowDirector (Li et al., 2025) serves as a representative and strong baseline for velocity-based video editing, achieving competitive editing quality through rectified-flow trajectory steering.

Despite its effectiveness, velocity-discrepancy-based editing often struggles when the target edit requires a substantial object-level structural change. Rather than fully replacing the source structure, the model may preserve a source-like silhouette and overlay target appearance on top of it, producing ghost-like residue and unnatural object geometry. Figure 1 illustrates this failure mode. FlowDirector changes the object appearance, but residual structures remain near the original contour, while the velocity-level editing update is weak or diffuse around the edited region. Such residual source cues can persist across frames, reducing the reliability of the intended transformation.

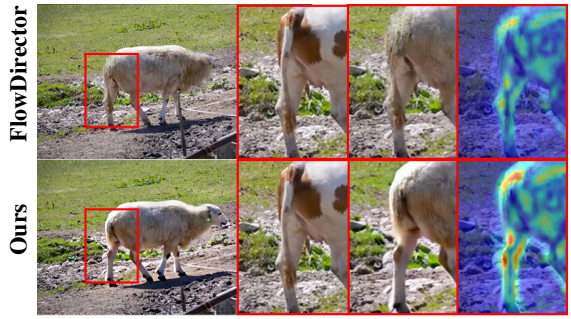

Figure 1. **Velocity-level analysis of source-residue artifacts.** Each row compares FlowDirector and Dual-PAG using the edited frame, source-region crop, edited crop, and final velocity-update magnitude map. The heatmaps visualize where the editing update is concentrated.

These observations indicate a limitation of the base editing direction. The source-to-target velocity discrepancy can provide an effective global transformation signal, but it does not explicitly distinguish source-preserving components from target-forming components. A straightforward alternative is to tune the source or target CFG scales (Ho & Salimans, 2022). However, scale adjustment primarily changes the overall strength of prompt-conditioned predictions rather than isolating structure-sensitive components. This distinction is important because prior diffusion editing studies have shown that self-attention is closely tied to structure preservation and content consistency. Self-attention features help preserve source layout in image-to-image translation (Tumanyan et al., 2023), mutual self-attention enables consistent non-rigid editing by querying source contents (Cao et al., 2023), and recent attention analysis shows that self-attention is crucial for retaining geometric and shape details

[*]Equal contribution. [†]Corresponding author. [1] Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea. Correspondence to: Suk-Ju Kang <sjkang@sogang.ac.kr>.

Preprint. June 8, 2026.

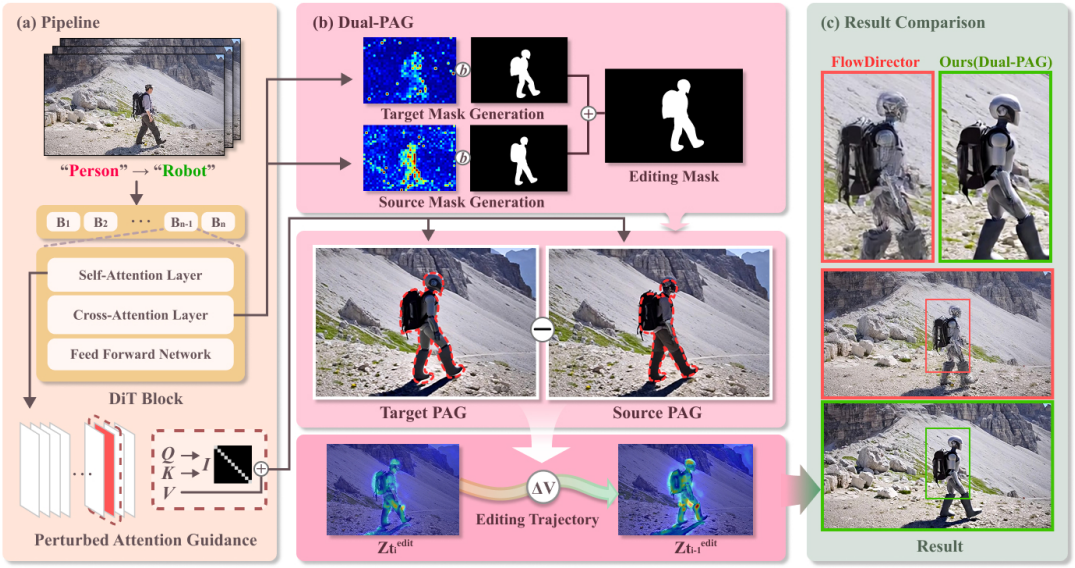

Figure 2. **Overview of Dual-PAG.** (a) Given a source video and a source-to-target prompt, PAG perturbs selected self-attention layers in the DiT-based editing pipeline. (b) Dual-PAG adds Target-PAG for target formation and subtracts Source-PAG for residue suppression, while a cross-attention mask localizes the correction to the edited region. (c) The resulting trajectory reduces source residue and improves target formation compared with FlowDirector.

during text-guided editing (Liu et al., 2024a).

We therefore revisit Perturbed Attention Guidance (PAG) (Ahn et al., 2024), which perturbs self-attention to expose structure-sensitive prediction components. Since the base source-to-target velocity edits along a single direction without explicitly separating source removal from target construction, applying PAG to the source- and target-conditioned branches can reveal complementary structural cues associated with source persistence and target formation. Based on this observation, we propose **Dual-PAG**. Our contributions are threefold: (i) we identify source residue as an entanglement failure in velocity-based editing; (ii) we introduce Source-PAG and Target-PAG to guide source suppression and target formation; and (iii) we demonstrate through velocity-level analyses that bidirectional guidance produces more localized source-to-target editing updates in challenging object replacement scenarios.

2. Preliminaries

2.1. Rectified Flow-based Editing.

Rectified Flow models (Kulikov et al., 2025; Liu et al., 2022; Lipman et al., 2022; Ho & Salimans, 2022) formulate generation as a continuous transport process over latent states. Let Z_t be the latent video state at time $t \in [0, 1]$, and let C be the text condition. The text-conditioned trajectory is defined by

$$\frac{dZ_t}{dt} = V_\theta(Z_t, t, C), \tag{1}$$

where V_θ denotes the learned velocity field that predicts the instantaneous update direction of Z_t under condition C.

Building on inversion-free rectified-flow editing, the editing direction is defined as the discrepancy between target- and source-prompt velocities, and the editing state is updated by

$$\Delta V_t = V_\theta(Z_t^{tar}, t, P^*) - V_\theta(Z_t^{src}, t, P) \tag{2}$$

$$Z_{t_{i-1}}^{edit} = Z_{t_i}^{edit} + (t_{i-1} - t_i)\Delta V_{t_i} \tag{3}$$

Here, P and P^* denote the source and target prompts, respectively. In practice, we follow the original framework and use CFG-guided velocity predictions. Our method further augments this rectified-flow editing velocity with an auxiliary semantic correction signal.

2.2. Perturbed Attention Guidance.

Perturbed Attention Guidance (PAG) (Ahn et al., 2024) constructs an auxiliary degraded prediction by perturbing the self-attention (SA) (Vaswani et al., 2017) operation. Given query Q, key K, and value $\mathcal{V}$, the standard self-attention operation computes token-to-token interactions through an attention map A and aggregates the value features as

$$\mathrm{SA}(Q, K, \mathcal{V}) = A\mathcal{V}, \quad A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{4}$$

where d denotes the feature dimension. PAG replaces the attention map A with the identity matrix I in order to remove off-diagonal token-to-token interactions while preserving

Prompt: A (woman→man) walking in the grass near a tree

Figure 3. **Temporal comparison on the Lucia sequence.** Rows show the source video, FlowDirector, and our Dual-PAG results; columns show three timestamps. Our method maintains better temporal consistency and identity preservation.

each token's own value feature:

$$\mathrm{PSA}(Q, K, \mathcal{V}) = I\mathcal{V} = \mathcal{V}. \tag{5}$$

Since self-attention maps encode structural relationships among visual tokens, this identity substitution selectively attenuates self-attention-mediated structural aggregation without directly corrupting the value features. The resulting perturbed prediction serves as a structurally degraded counterpart to the original prediction. This property motivates our use of the PAG-induced discrepancy as an additional guidance signal for rectified-flow video editing.

3. Method

3.1. Overview

Figure 2 summarizes the proposed Dual-PAG framework for reducing source residue and reinforcing target-specific structure. Given a source video and source/target prompts, we query the Wan2.1 (Wan et al., 2025)-based DiT (Peebles & Xie, 2023) under normal and attention-perturbed conditions to obtain the base editing velocity and PAG correction signals. Source-PAG suppresses source-preserving cues, while Target-PAG reinforces target-specific structure. Cross-attention responses localize the editing region. The following subsections describe the PAG corrections, mask construction, and localized velocity composition.

3.2. Source-Target PAG

The base editing velocity ΔV_t gives the main source-to-target direction, but it does not explicitly separate source removal from target formation. To address this, we decompose PAG into two complementary branches: Source PAG, which suppresses source-preserving structure, and Target PAG, which enhances target-specific structure.

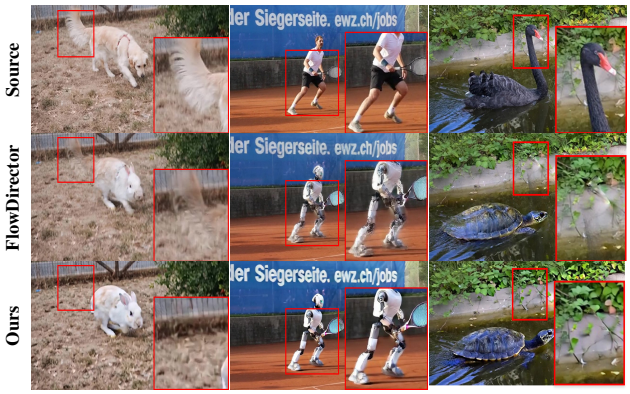

Figure 4. **Cropped comparisons across object-editing examples.** Rows show the source video, FlowDirector, and our Dual-PAG results. Red boxes highlight regions where source-residue artifacts or local structural remnants remain.

3.2.1. Source PAG: Residue Suppression

A common failure in text-driven video editing is source residue, where the outline or structural cues of the original object remain after editing. We attribute this to source-conditioned velocity components that preserve source-specific semantics and structure. To estimate these components, we apply PAG to the source-conditioned branch and compute the discrepancy between the normal and self-attention-perturbed source velocities:

$$\delta_t^{src} = V_\theta(Z_t^{src}, t, P) - \widetilde{V}_\theta(Z_t^{src}, t, P), \tag{6}$$

where P is the source prompt and $\widetilde{V}_\theta$ denotes the velocity predicted with perturbed self-attention. Because this discrepancy is computed within the source-conditioned branch, it isolates attention-dependent components that support the source prompt. We therefore subtract δ_t^{src} from the editing direction to weaken source-preserving structure in the editable trajectory.

3.2.2. Target PAG: Structure Enhancement

Another failure case in text-driven video editing is weak target formation, where the edited region changes in texture but does not fully develop the distinctive structure of the target object. To enhance such components, we apply PAG to the target-conditioned branch and compute the discrepancy between the normal and self-attention-perturbed target velocities:

$$\delta_t^{tar} = V_\theta(Z_t^{tar}, t, P^*) - \widetilde{V}_\theta(Z_t^{tar}, t, P^*), \tag{7}$$

where P^* is the target prompt and $\widetilde{V}_\theta$ denotes the velocity predicted with perturbed self-attention. Since the perturbed target prediction weakens attention-mediated structure formation, δ_t^{tar} emphasizes target-conditioned components that are lost under perturbation. We therefore add this target-side correction to the editing direction to reinforce target-specific structure rather than merely changing local appearance.

Table 1. Quantitative comparison between FlowDirector (Li et al., 2025) and Dual-PAG on DAVIS-Edit (Pont-Tuset et al., 2017) using manually written editing prompts. We report both editing-oriented metrics and VBench (Huang et al., 2024) video-quality metrics. ↑ denotes higher is better.

Method	Editing-oriented Metrics					VBench Metrics				
	CLIP_T ↑	CLIP_F ↑	Warp-SSIM ↑	Q_{EDIT} ↑	PickScore ↑	Subject Cons. ↑	Aesthetic Quality ↑	Imaging Quality ↑	Temporal Style ↑	Overall Cons. ↑
FlowDirector (Li et al., 2025)	0.2842	0.9566	**0.4104**	**0.1168**	20.4382	0.8876	0.5169	65.5998	0.1250	0.1250
Ours	**0.2935**	**0.9604**	0.3861	0.1129	**20.6200**	**0.8978**	**0.5262**	**66.1564**	**0.1292**	**0.1292**

Table 2. Ablation study of Dual-PAG on DAVIS-Edit (Pont-Tuset et al., 2017). ↑ denotes higher is better.

Method	CLIP_T ↑	CLIP_F ↑	PickScore ↑
w/o Source PAG	0.2930	0.9600	20.6153
w/o Target PAG	0.2851	0.9566	20.4612
Ours	**0.2935**	**0.9604**	**20.6200**

3.3. PAG-Guided Edit Masking

To prevent Dual-PAG from perturbing background regions, we localize the correction using cross-attention responses, following the attention-guided localization strategy of FlowDirector (Li et al., 2025). At each rectified-flow timestep t, we aggregate the corresponding cross-attention maps over selected DiT layers and heads, resize them to the latent resolution, and obtain binary spatial masks:

$$M_t^{src} = \mathcal{B}(A_t^{src}), \quad M_t^{tar} = \mathcal{B}(A_t^{tar}), \tag{8}$$

where A_t^{src} and A_t^{tar} denote the aggregated source- and target-word attention maps, respectively. M_t^{src} and M_t^{tar} are the resulting source and target binary masks in the latent space, and $\mathcal{B}(\cdot)$ denotes normalization followed by thresholding.

The final edit mask is defined as the union of the source and target masks:

$$M_t^{edit} = \text{clip}\left(M_t^{src} + M_t^{tar}, 0, 1\right), \tag{9}$$

where M_t^{edit} specifies the latent spatial region where the Dual-PAG correction is applied. The clipping operation limits overlapping mask values to one.

3.4. Optimization

At each rectified-flow timestep t, we compute the base editing velocity ΔV_t and the source/target PAG discrepancies in the latent space. The final localized editing direction is

$$\Delta V_t^{ours} = \Delta V_t + M_t^{edit} \odot \left(\lambda_{\text{tar}}\delta_t^{\text{tar}} - \lambda_{\text{src}}\delta_t^{\text{src}}\right), \tag{10}$$

where $\odot$ denotes element-wise multiplication over the latent velocity tensor. The mask M_t^{edit} spatially gates the Dual-PAG correction at each timestep, applying it only to prompt-relevant latent regions.

4. Experiments

We evaluate on DAVIS-Edit (Pont-Tuset et al., 2017) using our manually written editing prompts, with FlowDirector (Li et al., 2025) as the main baseline. Due to limited computational resources, we use a reduced inference setting with fewer sampling steps and a smaller averaging parameter n_{avg} than the default configuration. For a fair comparison, the same inference setting is applied to both FlowDirector and Dual-PAG.

4.1. Qualitative Analysis

As shown in Figure 3, Dual-PAG produces a clearer target identity across timestamps while preserving the source motion and background layout. FlowDirector follows the input motion, but the edited subject remains visually ambiguous and retains source-like appearance. In Figure 4, FlowDirector also leaves local source-residue artifacts in the cropped regions, including residual silhouettes and incomplete structural replacement. Dual-PAG suppresses these remnants and forms more target-consistent structures, supporting the effectiveness of explicitly separating source suppression and target formation.

4.2. Quantitative Analysis

As shown in Table 1, Dual-PAG improves the main semantic and preference-based metrics over FlowDirector, including CLIP_T, CLIP_F, and PickScore. It also improves all VBench metrics, suggesting better target alignment and overall video quality. Warp-SSIM and Q_{edit} are lower than FlowDirector, which should be interpreted together with Figure 1. FlowDirector can obtain higher structure-similarity scores by preserving the original source silhouette, even when visible source-residue artifacts remain. In contrast, Dual-PAG suppresses such residual structures and forms new target geometry, which can reduce preservation-oriented scores. This suggests that Warp-SSIM and Q_{edit} are limited as standalone criteria for residue-free object replacement.

4.3. Ablation Study

Table 2 shows the contribution of each PAG branch. Removing Target PAG clearly decreases CLIP_T, CLIP_F, and PickScore, showing that target-side guidance is important for target formation. Removing Source PAG gives scores close to the full model, but remains consistently lower across all metrics. This suggests that Source PAG provides a complementary effect by suppressing source-residue cues that are not fully captured by prompt-alignment metrics. The full Dual-PAG model achieves the best scores, supporting the need to separate source suppression and target formation.

5. Conclusion

We presented Dual-PAG, a bidirectional perturbed-attention guidance framework for residue-free text-driven video editing. Dual-PAG separates source-residue suppression from target-structure formation by subtracting source-side PAG residuals and adding target-side PAG residuals within rectified-flow editing. This branch-specific correction reduces source-preserving structural traces while reinforcing target-consistent geometry. These results show that branch-specific PAG correction is effective for residue-free object editing in rectified-flow video models.

6. Acknowledgement

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026(Project Name: Multimodal Storyverse: A Multi-Agent Collaborative Platform for Expansive Narrative Creation, Project Number: RS-2026-25527029, Contribution Rate: 50%) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00414230: 50%).

References

Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H. H., Jin, K. H., and Kim, S. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.-M., Rosenhahn, B., Xiang, T., and He, S. Flatten: optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations*, volume 2024, pp. 52086–52106, 2024.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., and Michaeli, T. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Li, G., Yang, Y., Song, C., and Zhang, C. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv: 2506.05046*, 2025.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, B., Wang, C., Cao, T., Jia, K., and Huang, J. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7817–7826, June 2024a.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4199–4209, 2023.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., and Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wu, B., Chuang, C.-Y., Wang, X., Jia, Y., Krishnakumar, K., Xiao, T., Liang, F., Yu, L., and Vajda, P. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8261–8270, 2024.