FlowTrack: Controlling Edit-Signal Execution in Inversion-Free Flow Video Editing

Zhong JieLun[1], Di Wang[3], Jisheng Dang[1]*, Bimei Wang[1], Jingwen Zhao[1], Wencan Zhang[2], Hong Peng[1], Bin Hu[4]*, Leigang Qu[2]*

[1] School of Information Science and Engineering, Lanzhou University [2] National University of Singapore
[3] Cuiying Honors College, Lanzhou University [4] School of Computer Science and Technology, Beijing Institute of Technology
Zhong JieLun: Computer Science and Technology, Third-year undergraduate student * Corresponding authors: Jisheng Dang, Bin Hu, and Leigang Qu

NEURAL INFORMATION PROCESSING SYSTEMS

1. Edit-signal execution control

1.1 Text-guided video editing

Text-guided editing takes a source video and source and target texts. The edited video should match target semantics while preserving motion, camera trajectories, background, and layout.

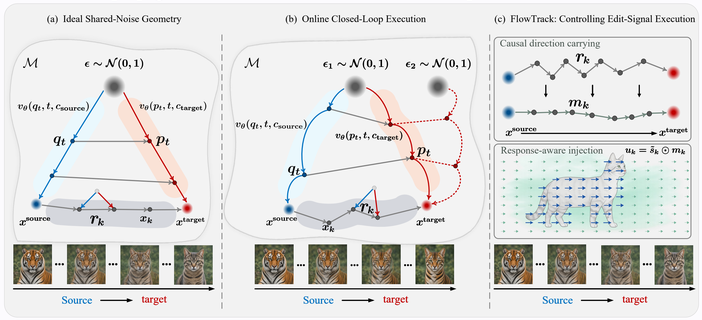
FlowEdit and FlowDirector use pretrained velocity fields without source-specific inversion. FlowTrack controls how the resulting edit signal enters the solver.



Sunglasses insertion

A local semantic change can affect the full video latent. Editing must preserve identity, motion, and layout across frames while limiting changes to unedited content.

Ideal shared-noise coupling gives velocity differences an endpoint-displacement interpretation. Neural trajectories need not be globally straight. During sampling, the model provides an online conditional-mean observation of an unknown edited endpoint. Earlier injections alter later signals, so direct execution can propagate directional fluctuations and weak responses.



1.2 Flow Matching and inversion-free updates

$$z_\tau = (1-\tau)\varepsilon + \tau y, \qquad \frac{\mathrm{d}z_\tau}{\mathrm{d}\tau} = v_\theta(z_\tau, \tau, c)$$

Shared-noise source and target branches

$$x_0 = x^{\mathrm{src}}, \qquad q_k = (1-\tau_k)\varepsilon_k + \tau_k x^{\mathrm{src}}$$
$$\Delta x_k = x_k - x^{\mathrm{src}}, \qquad p_k = q_k + \Delta x_k$$

Conditional velocity difference and state update

$$r_k = v_\theta(p_k, \tau_k, c_{\mathrm{tgt}}) - v_\theta(q_k, \tau_k, c_{\mathrm{src}})$$
$$\text{Direct: } u_k = r_k, \qquad x_{k+1} = x_k + \Delta\tau_k u_k$$

Here y is a clean video latent and ε is Gaussian noise. The velocity target is y − ε. Both editing branches share the same noise at each step and are rebuilt after each update.

1.3 Directional memory and response support

FlowTrack retains the pretrained generator and raw-signal computation. The executed signal follows a strictly convex objective that separates directional memory from response-dependent attenuation.

Directional memory

$$m_k = \frac{r_k + \lambda m_{k-1}}{1+\lambda}$$

Response-aware injection

$$u_k = \widetilde{s}_k \odot m_k, \quad \widetilde{s}_k \in [0,1]$$

Unified local objective

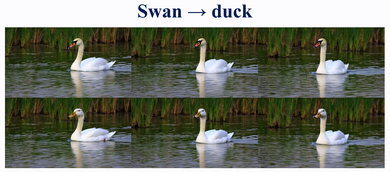
$$\mathcal{L}_k(z) = \frac{1}{2}\|z - r_k\|_F^2 + \frac{\lambda}{2}\|z - m_{k-1}\|_F^2 + \frac{1}{2}\|a_k^{1/2} \odot z\|_F^2, \quad \lambda \geq 0,\ a_k \geq 0$$

Closed-form execution

$$u_k^\star = \arg\min_z \mathcal{L}_k(z) = \frac{r_k + \lambda m_{k-1}}{1+\lambda+a_k} = \underbrace{\frac{1+\lambda}{1+\lambda+a_k}}_{\widetilde{s}_k} \odot \underbrace{\frac{r_k + \lambda m_{k-1}}{1+\lambda}}_{m_k}$$

Here τ is sampling time and c denotes text conditions. Cost a is nonnegative and follows current signal magnitudes. Field multiplication and division are elementwise.

Swan → duck



Carrying preserves memory before attenuation. Low support weakens current injection without recursively attenuating memory.

Cap insertion



Continuous support follows current evidence. Values in [0, 1] attenuate injection without amplifying it or creating missing semantics.

2. Editing quality and source preservation

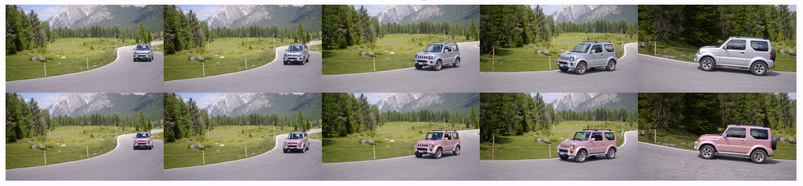
2.1 Quantitative comparison

MFS measures motion fidelity and CLIP measures target alignment. PSNR, LPIPS, MSE, SSIM, and Structure assess visual and structural preservation. Arrows indicate the preferred direction.

Method	MFS ↑	CLIP ↑	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	Structure ↓
FateZero	73.7298	26.0856	18.6122	236.8776	155.9120	69.1964	55.3834
FLATTEN	73.9243	26.9928	16.3716	269.9812	255.4401	62.9804	71.3739
FlowDirector	85.2422	**27.6559**	27.4104	52.4975	22.9882	87.7195	8.9845
RAVE	82.3514	27.3857	20.8132	137.9647	96.9810	78.1291	25.2783
TokenFlow	81.1081	26.4377	23.0263	137.5256	57.7608	75.4634	26.6105
Wan-Edit	83.1296	26.1248	25.5700	71.1971	39.8569	83.4118	12.2960
DFVEdit	82.8349	27.1560	25.1480	72.4622	37.9288	82.9375	11.9711
FlowTrack	**86.0860**	27.4871	**29.2961**	**39.5408**	**15.3677**	**90.2092**	**5.9808**

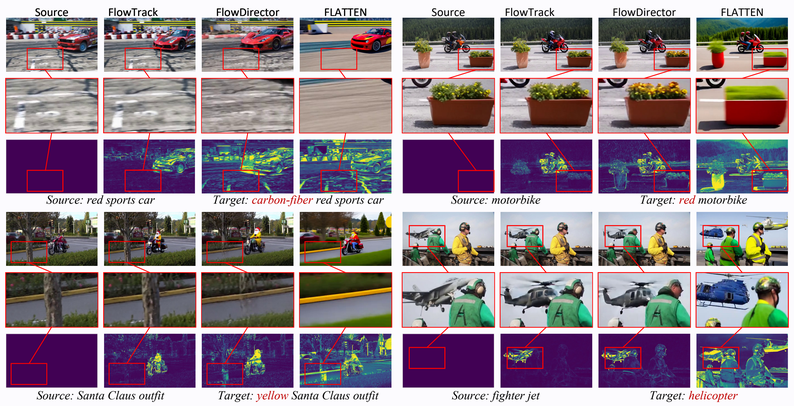
FlowTrack achieves the best results on six of seven metrics. Relative to FlowDirector, MFS increases by 0.8438 and LPIPS decreases by 24.7%, while CLIP decreases by 0.1688.

Silver SUV → pink SUV



2.2 Qualitative comparison and edit-response maps

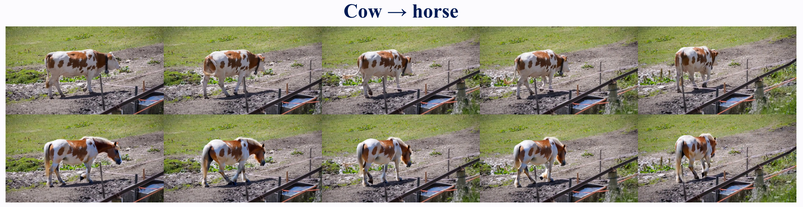
Output crops and edit-response maps show source preservation and the spatial distribution of editing responses. Warmer colors indicate stronger responses.



Silver car → silver jeep



Cow → horse



3. Component analysis and editing examples

3.1 Component ablation and temporal stability

The ablation study uses the same pretrained model and raw-signal computation, varying only carrying and support.

Method	MFS ↑	CLIP ↑	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	Struct. ↓
Direct	83.8333	**27.8676**	25.0004	75.0427	38.8298	83.0756	12.3624
Carry-only	83.9296	27.8392	25.1247	73.4815	37.6697	83.3543	12.0138
Support-only	85.9782	27.5157	29.1260	40.2292	15.9692	90.1019	6.2062
FlowTrack	**86.0860**	27.4871	**29.2961**	**39.5408**	**15.3677**	**90.2092**	**5.9808**

FlowTrack vs. Support-only: temporal error reductions (paired 95% CI)

Aligned edge flicker ↓	**0.84% [0.66%, 1.04%]**
Aligned SSIM dissimilarity ↓	**1.04% [0.79%, 1.35%]**
Relative flow EPE ↓	**3.13% [2.46%, 3.82%]**

Support accounts for most preservation gains. Carrying further reduces temporal errors (paired 95% CIs above zero). Its CLIP difference is statistically inconclusive.

3.2 Sensitivity to the carrying coefficient

λ	MFS ↑	CLIP ↑	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	Struct. ↓
0.00	89.0420	27.6360	27.9142	47.8503	20.5650	88.0821	6.7080
0.05	89.0756	27.5780	27.9593	47.5330	20.3940	88.1195	6.6210
0.10	89.1514	27.5512	27.9998	47.2680	20.1720	88.1477	6.5350
0.20	**89.1663**	**27.5472**	**28.0725**	**46.9238**	**19.8820**	**88.2065**	**6.4310**
0.40	89.1671	27.5205	28.2119	46.3439	19.3430	88.2917	6.2070

On the fixed subset, λ = 0.2 yields 99.4% of the MFS gain from λ = 0 to 0.4, with higher CLIP.

3.3 Appearance and semantic transformations

Bicycle → purple bicycle



Hawk → blue bird



Snowboarder → snowboarder in red pants



Woman in dress → man in suit



Hockey player → Batman



Boat → kayak

