COUNTERFACTUAL DISCOVERY OF MOTION-EFFECT NODE SETS IN VIDEO DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video diffusion transformers make camera and object motion observable, but observation and attribution do not by themselves identify which internal computations support a particular motion effect. Large attribution can reflect denoising shared across motions rather than support specific to the requested direction. We introduce CAP-MED, a counterfactual activation-patching protocol for motion-effect node discovery. It uses target-vs-other attribution to select compact node sets across denoising steps, layers, and submodules, then tests whether exchanging their activations recovers and suppresses the effect between matched motion-instructed and empty-action rollouts. On Wan2.2, specificity-selected sets achieve median normalized recovery of 0.99 and suppression of 0.97, compared with 0.15 and 0.59 for attribution-magnitude selection at the same node budget. Selector controls and independent camera and object readouts support the motion-specific interpretation of these effects. Further head-level interventions identify conditional dependence between selected sets and a value-dominant effect within selected camera-pan heads. LTX-Video experiments test the full workflow on a second architecture and extend it to color temperature. The evidence characterizes selected node sets and their functional roles rather than a complete motion-generation circuit.

1 INTRODUCTION

Text-conditioned video diffusion transformers can pan across a scene, zoom toward an object, or move a subject relative to its surroundings (Yang et al., 2025; Wan et al., 2025). These behaviors use a shared denoising network, yet require different changes across space and time. An internal analysis must distinguish components that support a particular motion from those that contribute broadly to generation. Without this distinction, an intervention may reproduce much of an instructed video while revealing little about the internal support for its motion.

Motion-aware features (Xiao et al., 2024) and spatiotemporal attention patterns (Wen et al., 2025) provide candidates for this analysis. Decoding or visualizing a motion signal, however, does not determine the effect of replacing it (Belinkov, 2022; Zhang & Nanda, 2024). Attribution brings the search closer to intervention by estimating how activation differences affect a motion score. The remaining selection problem is that a component with large target attribution may also contribute strongly to other motions. We test whether comparing attribution across directions yields more effective intervention sets than ranking by target magnitude alone.

The candidate space spans denoising steps, transformer layers, and attention and feed-forward submodules. A node is the full token-by-channel output of one submodule at one layer and denoising step. This resolution separates attention and feed-forward contributions while keeping the search tractable. Even at this resolution, testing every set by activation replacement is too expensive (Conmy et al., 2023; Kramár et al., 2024). We therefore use edge attribution patching with integrated gradients (EAP-IG) to combine differences between paired activations with the sensitivity of the motion score along directed computational edges (Syed et al., 2024; Hanna et al., 2024).

We propose **CAP-MED**, a counterfactual activation-patching protocol for motion-effect node discovery. The protocol first ranks layers by *target-vs-other specificity*: target attribution divided by mean attribution to alternative motion directions. The selected layers are combined with fixed steps and submodules to form a node set (Figure 1). We test each set on paired rollouts that share the scene, seed, and sampler, with a motion instruction present in one and absent in the other. Inserting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **The CAP-MED protocol.** (a) A rollout with a motion instruction (V_1) is compared with one without it (V_0), and the motion effect TE_a is the difference in motion classifier score. (b) EAP-IG scores how much each node drives this effect, summed per layer for the target motion and for the other motions. (c) Layers with a high target score relative to the other motions are kept, giving a set N_a. (d) Inserting V_1 activations into V_0 at N_a tests whether these nodes are enough to create the motion, and inserting V_0 activations into V_1 tests whether they are needed.

instructed activations into the empty-action rollout measures recovery of the motion effect. Reversing the exchange measures suppression in the instructed rollout. The comparison with magnitude ranking uses the same attribution scores and node budget. Additional controls fix the step–submodule grid to test whether layer choice accounts for the difference.

We evaluate the main comparison on Wan2.2 and a model-specific variant on LTX-Video for camera motion and color temperature. Our contributions are:

- A target-vs-other selection method for motion-effect node sets, tested through matched recovery and suppression interventions across denoising steps and transformer layers.
- A fixed-budget comparison showing stronger exchange effects than magnitude ranking, with matched-grid controls and independent motion measurements that test alternative explanations for the advantage.
- Head-level analyses showing conditional dependence between two selected sets and larger value-exchange effects in camera-pan heads. These use separate head sets from the main submodule-level comparison.

2 Related Work

Video diffusion models and motion control. Diffusion transformers were introduced for image generation (Peebles & Xie, 2023). Video models incorporate temporal structure and latent representations (Blattmann et al., 2023), with large transformers supporting text-conditioned generation (Yang et al., 2025; Kong et al., 2024; HaCohen et al., 2024; Wan et al., 2025). MotionCtrl separates camera and object control (Wang et al., 2024), CameraCtrl adds camera-trajectory conditioning (He et al., 2025), and MotionDirector adapts motion from reference videos (Zhao et al., 2024). Trajectory attention and training-free guidance provide further control mechanisms (Xiao et al., 2025; Hou & Chen, 2025). Our experiments address the internal support for motion already produced by a pretrained model, using activation exchange to measure the effect of selected components.

Diffusion representations and motion localization. Cross-attention has been used for prompt-based editing and text-to-region attribution (Hertz et al., 2023; Tang et al., 2023b). ConceptAttention obtains saliency maps by projecting concept embeddings into the attention output space (Helbling et al., 2025). Intermediate diffusion features support recognition and semantic correspondence across layers and denoising timesteps (Yang & Wang, 2023; Luo et al., 2023; Tang et al., 2023a), while Revelio uses sparse autoencoders to examine interpretable features (Kim et al., 2025). For video, MOFT identifies and manipulates motion-aware features without additional training (Xiao et al., 2024). IMAP localizes motion concepts through video-token and attention-head statistics (Jun et al., 2026). We focus on choosing a fixed-budget set from attribution scores and measuring its effect in both exchange directions.

Circuit discovery and activation patching. Transformer-circuit studies analyze communication among heads and feed-forward layers (Elhage et al., 2021). Activation interventions have localized gender-bias effects, factual recall, and indirect object identification (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023). Causal abstraction and path patching formalize tests of internal hypotheses (Geiger et al., 2025; Goldowsky-Dill et al., 2023). ACDC searches for behavior-preserving circuits (Conmy et al., 2023), while EAP and AtP* reduce search cost with gradient-based approximations (Syed et al., 2024; Kramár et al., 2024). EAP-IG averages gradients along an interpolation path (Hanna et al., 2024), using integrated gradients (Sundararajan et al., 2017). We use this estimator to compare selection rules over a denoising-time node space.

Our analysis separates two design choices: how an intervention is evaluated and the resolution at which it is applied. Patching results depend on the behavioral metric and reference condition (Zhang & Nanda, 2024). We specify both and evaluate patched videos with additional motion readouts. At the level of intervention resolution, sparse feature circuits and transcoders resolve computation into interpretable features (Marks et al., 2025; Dunefsky et al., 2024). Our main interventions exchange coarser submodule outputs, each containing many tokens and channels. Head and operand exchanges provide separate functional analyses at finer resolution.

3 CAP-MED: DISCOVERING MOTION-EFFECT NODE SETS

3.1 PAIRED ROLLOUTS AND THE INTERVENTION UNIT

A rollout is a denoising trajectory that generates a video from fixed conditioning and seed. For scene prompt S, seed U, and signed motion direction $a \in \mathcal{A}$, let A_1 append the direction-a instruction and let A_0 leave that instruction slot empty. The source video $V_1 = G_\theta(S, A_1, U)$ and base video $V_0 = G_\theta(S, A_0, U)$ share the scene, seed, sampler, and generation settings. The base video is uninstructed, not necessarily motionless. For classifier logits l_a and $l_{\bar{a}}$, where $\bar{a}$ is the opposite direction on the same axis, define

$$Y_a(V) = l_a(V) - l_{\bar{a}}(V), \qquad \mathrm{TE}_a(S, U) = Y_a(V_1) - Y_a(V_0). \tag{1}$$

TE_a is the score change induced by the instruction under a fixed scene and seed.

A node $n = (t, \ell, m)$ denotes the full token-by-channel activation field at denoising step t, layer ℓ, and submodule $m \in \mathcal{M}_{\mathrm{full}} = \{\mathrm{CA}, \mathrm{SA}, \mathrm{FFN}\}$. These are cross-attention, self-attention, and feed-forward outputs. We write $H_n^{(c)}$ for the activation in condition c. The node budget counts these fields, each of which contains all tokens and channels.

3.2 ATTRIBUTION-BASED CANDIDATE SCORING

Edge attribution patching combines a paired activation difference with a gradient of the behavioral score (Syed et al., 2024). Following EAP-IG (Hanna et al., 2024), the estimator below averages edge-specific gradients along a paired-activation interpolation using the integrated-gradients construction (Sundararajan et al., 2017). Let $\mathcal{T}_{\mathrm{rank}}$ be the scored denoising steps and $\mathcal{E}_t$ the forward-valid upstream-downstream edges at step t. For N_{pair} matched attribution pairs and K_{IG} integration points, the edge score is

$$A_{a,t}(u \to v) = \frac{1}{N_{\mathrm{pair}}} \sum_{p=1}^{N_{\mathrm{pair}}} \left\langle h_{p,t,u}^{(1)} - h_{p,t,u}^{(0)},\ \frac{1}{K_{\mathrm{IG}}} \sum_{k=1}^{K_{\mathrm{IG}}} g_{p,t}^{a,k}(u \to v) \right\rangle. \tag{2}$$

Here $h^{(1)}$ and $h^{(0)}$ are instructed and empty-action upstream activations, and $g_{p,t}^{a,k}(u \to v)$ denotes the edge-specific gradient signal for Y_a at integration point k. The inner product sums over tokens and channels. Scoring includes CA, SA, and FFN.

3.3 TARGET-VS-OTHER SELECTION

To compare a layer's contributions across motions, we first sum its absolute edge attributions over scored steps, upstream submodules, and downstream endpoints:

$$E_a(\ell) = \sum_{t \in \mathcal{T}_{\text{rank}}} \sum_{m \in \mathcal{M}_{\text{full}}} \sum_{v:((t,\ell,m)\to v) \in \mathcal{E}_t} \left| A_{a,t}((t,\ell,m) \to v) \right|. \tag{3}$$

We divide the target sum by the mean sum for the other directions, with a small stabilizer $\epsilon > 0$, and select the top K layers:

$$\text{Spec}_a(\ell) = \frac{\epsilon + E_a(\ell)}{\epsilon + (|\mathcal{A}| - 1)^{-1} \sum_{b \neq a} E_b(\ell)}, \qquad \mathcal{L}_a^{\text{spec}} = \text{TopK}_\ell \, \text{Spec}_a(\ell). \tag{4}$$

A ratio above one means that target attribution exceeds the non-target mean. Layers are selected by rank, so even a ratio below one can enter the set. There is no additional specificity threshold.

For fixed intervention steps $\mathcal{T}_{\text{sel}}$ and patch submodules $\mathcal{M}_{\text{patch}}$, the selected set is

$$\mathcal{N}_a^{\text{spec}} = \mathcal{T}_{\text{sel}} \times \mathcal{L}_a^{\text{spec}} \times \mathcal{M}_{\text{patch}}. \tag{5}$$

This construction selects layers while fixing steps and patch submodules, $\mathcal{M}_{\text{patch}} \subseteq \mathcal{M}_{\text{full}}$. It is the selection rule used in the main Wan comparison. The LTX extension ranks individual nodes with a magnitude-balanced variant (Section H.2).

3.4 BIDIRECTIONAL ACTIVATION EXCHANGE

For a proposed set $\mathcal{N}$, let $\mathcal{H}_{\mathcal{N}}^{(c)} = \{H_n^{(c)} : n \in \mathcal{N}\}$ be the donor fields from condition c. Sufficiency inserts instructed donor fields into the base rollout, while necessity supplies base donor fields to the instructed rollout:

$$\begin{aligned} V_{0 \leftarrow 1}^{\mathcal{N}} &= G_\theta(S, A_0, U \mid \mathcal{H}_{\mathcal{N}}^{(0)} \leftarrow \mathcal{H}_{\mathcal{N}}^{(1)}), \\ V_{1 \leftarrow 0}^{\mathcal{N}} &= G_\theta(S, A_1, U \mid \mathcal{H}_{\mathcal{N}}^{(1)} \leftarrow \mathcal{H}_{\mathcal{N}}^{(0)}). \end{aligned} \tag{6}$$

The recipient continues denoising after each overwrite. Recovery and suppression are measured by

$$\text{Eff}_{\mathcal{N},a}^{\text{suff}} = Y_a(V_{0 \leftarrow 1}^{\mathcal{N}}) - Y_a(V_0), \qquad \text{Eff}_{\mathcal{N},a}^{\text{nec}} = Y_a(V_1) - Y_a(V_{1 \leftarrow 0}^{\mathcal{N}}). \tag{7}$$

For either direction $d \in \{\text{suff}, \text{nec}\}$, we report the sign-normalized ratio and sign-correctness:

$$\rho_{\mathcal{N},a}^d = \frac{\text{Eff}_{\mathcal{N},a}^d \, \text{sign}(\text{TE}_a)}{|\text{TE}_a| + \epsilon}, \qquad \text{SC}_{\mathcal{N},a}^d = \mathbf{1}\Big[\text{sign}(\text{Eff}_{\mathcal{N},a}^d) = \text{sign}(\text{TE}_a)\Big]. \tag{8}$$

We report median per-pair ρ and the fraction of sign-correct trials. These summarize effect magnitude relative to the instruction-induced change and agreement with its sign. A ratio near one refers to the measured score difference, not physical displacement. Sufficiency and necessity denote these exchange tests.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND SELECTED LAYERS

The main experiments use Wan2.2 T2V-A14B (Wan et al., 2025), with 40-layer experts, 81-frame videos at 480×832, and ten FlowUniPC steps based on UniPC (Zhao et al., 2023) (Section A.1). We evaluate ten signed directions: camera pan, tilt, and zoom, and object lateral and depth motion. Our VideoMAE classifiers (Tong et al., 2022) supply the logits in Equation (1). Their held-out accuracies are 98.99% on 2,878 camera videos and 95.42% on 1,114 object videos (Section B).

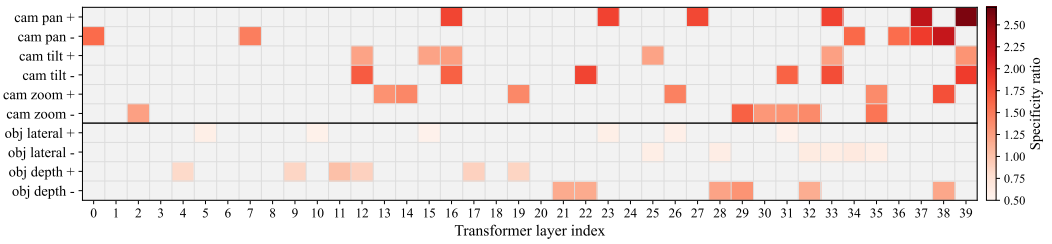


Figure 2: Layers selected by target-vs-other attribution. Rows are signed motion directions and columns are transformer layers. Colored cells mark the six selected layers per direction and give their raw specificity ratio. Object-lateral layers are selected by rank even though their ratios are below one.

Table 1: Per-axis CAP-MED validation under a matched 36-node budget. Succ. is sign-correctness, shown as rounded percentage and count. ρ is the median normalized intervention ratio. The aggregate row reports 95% bootstrap confidence intervals with 10,000 resamples over 400 trials.

	Sufficiency				Necessity			
Axis	Axis succ.	EAP-IG succ.	Axis ρ	EAP-IG ρ	Axis succ.	EAP-IG succ.	Axis ρ	EAP-IG ρ
cam pan +	100% (40/40)	83% (33/40)	+1.00	+0.83	100% (40/40)	80% (32/40)	+0.99	+0.91
cam pan −	98% (39/40)	75% (30/40)	+1.00	+0.47	100% (40/40)	95% (38/40)	+1.01	+0.78
cam tilt +	93% (37/40)	65% (26/40)	+0.88	+0.09	98% (39/40)	83% (33/40)	+0.97	+0.57
cam tilt −	95% (38/40)	75% (30/40)	+0.98	+0.13	100% (40/40)	90% (36/40)	+0.97	+0.81
cam zoom +	95% (38/40)	73% (29/40)	+1.01	+0.17	98% (39/40)	75% (30/40)	+0.95	+0.33
cam zoom −	100% (40/40)	68% (27/40)	+1.00	+0.19	100% (40/40)	83% (33/40)	+1.01	+0.18
obj lateral +	78% (31/40)	33% (13/40)	+0.34	−0.32	75% (30/40)	68% (27/40)	+0.43	+0.57
obj lateral −	88% (35/40)	50% (20/40)	+0.98	−0.00	95% (38/40)	85% (34/40)	+1.01	+0.73
obj depth +	80% (32/40)	20% (8/40)	+0.65	−0.45	83% (33/40)	80% (32/40)	+0.69	+0.61
obj depth −	90% (36/40)	68% (27/40)	+0.99	+0.26	95% (38/40)	60% (24/40)	+0.95	+0.04
Aggregate	91.5% [88.8, 94.0]	60.8% [56.2, 65.0]	+0.99 [+.96, +1.00]	+0.15 [+.06, +.26]	94.2% [92.0, 96.2]	79.8% [76.0, 83.5]	+0.97 [+.95, +.99]	+0.59 [+.42, +.78]

Attribution uses ten prompt pairs, ten sampled steps, and five integration points. The *Axis-specific* selector combines six specificity-ranked layers with steps $\{0, 2, 4\}$ and $\{\mathrm{SA}, \mathrm{FFN}\}$, giving 36 nodes. CA is included in scoring but has low attribution mass (Section A.2). The *EAP-IG top-36* baseline ranks individual nodes by raw target-attribution magnitude from the same graph. Each selector is evaluated on 40 pairs per direction and arm, giving 400 trials per selector and arm. The node count is matched, but allocation across steps, layers, and submodules can differ. Section 4.3 fixes the step–submodule allocation. Figure 2 shows middle-to-late camera selections and weaker, more heterogeneous object specificity. Both lateral selections have ratios below one.

4.2 Recovery and Suppression of Motion Effects

At 36 nodes, Axis-specific selection produces stronger aggregate exchange effects than EAP-IG top-36 (Table 1). Sufficiency is sign-correct on 366/400 versus 243/400 trials, with median normalized recovery of 0.99 versus 0.15. Necessity is sign-correct on 377/400 versus 319/400 trials, with median suppression of 0.97 versus 0.59. Paired-bootstrap differences exclude zero for both aggregate metrics and both arms. Median absolute effects are also larger, by factors of 8.3 in sufficiency and 3.8 in necessity. These factors use absolute score changes, not median ρ.

Axis-specific selection has larger median ρ in all ten sufficiency directions and nine necessity directions. Object lateral+ is the necessity exception, at 0.43 versus 0.57, and also has a relatively weak sufficiency effect. We next test the role of layer selection and whether a larger budget improves this weaker cell.

4.3 Layer Selection, Intervention Location, and Budget

The main comparison changes node allocation as well as ranking. To examine layer choice, we compare five selectors at 36 nodes on camera pan±, tilt±, and object lateral± (Table 2). Two

controls retain the Axis-specific step–submodule grid: random layers and layers with high attribution but low specificity. The latter selects the least specific layers from a high-attribution pool. Global EAP-IG and a linear probe also select 36 nodes, with potentially different allocations (Section C). The high-attribution/low-specificity control gives median recovery/suppression of 0.07/0.35, compared with 0.95/1.00 for Axis-specific selection. Thus, placing a high-attribution layer set on the same grid does not reproduce the effect.

Table 2: Selector comparison across six held-out signed directions at 36 nodes. SC is sign correctness, and $\widetilde{\rho}$ is median normalized recovery (sufficiency) or suppression (necessity).

	Sufficiency		Necessity	
Selector	SC (%)	$\widetilde{\rho}$	SC (%)	$\widetilde{\rho}$
Axis-specific	91	+0.95	92	+1.00
EAP-IG top-36	65	+0.10	82	+0.49
Random layers, same grid	71	+0.16	78	+0.29
High-attribution, low-specificity	61	+0.07	83	+0.35
Linear probe	54	+0.01	70	+0.34

The six directions are camera pan±, tilt±, and object lateral±. Random layers and high-attribution/low-specificity layers use the same step–submodule grid as Axis-specific selection. The global and linear-probe selectors match the 36-node budget. Color denotes magnitude, not statistical significance.

The allocation also matters (Section F). At 12 nodes, moving pan+ exchange from step 0 to step 2 raises median recovery from 0.03 to 0.99. A three-layer comparison gives 95% sign-correctness and $\rho = 1.00$ for late layers, versus 40% and -0.02 for middle layers. Increasing the object lateral+ budget to twelve layers raises recovery to 1.06 and suppression to 1.00, closing much of the recovery gap. Adding 18 aligned CA nodes to the default pan+ set changes median paired recovery by 0.00 percentage points, with a 95% interval of $[-0.265, +0.221]$. This tests CA's added effect alongside the selected SA/FFN fields.

4.4 Target and cross-axis responses

To check whether stronger target recovery accompanies broad score changes, we re-score each patched video on all ten signed directions. Let $\Delta_b = Y_b(V_T) - Y_b(V_0)$ in sufficiency and $\Delta_b = Y_b(V_1) - Y_b(V_T)$ in necessity, where V_T is the corresponding patched video. The signed target-off-axis ratio is defined as:

$$\mathrm{STOR}(a) = \frac{\Delta_a}{\epsilon + (|\mathcal{A}| - 1)^{-1} \sum_{b \neq a} |\Delta_b|}. \tag{9}$$

Positive STOR indicates a change in the expected target direction. Its denominator includes the algebraically dependent opposite score, making STOR a logit-space comparison rather than a ratio over independent physical axes. Axis-specific selection has larger target responses and STOR in most cells (Table 3). In object sufficiency, EAP-IG gives negative STOR in three of four directions, while Axis-specific selection remains positive. Object lateral+ necessity again favors EAP-IG.

4.5 Motion readouts beyond the scoring classifier

We evaluate the same patched videos with motion readouts unused in node selection. VBench CameraPredict (Huang et al., 2024a;b; Karaev et al., 2024b) treats target-label presence as success in sufficiency and absence as success in necessity, with empty predictions mapped to static. Axis-specific selection has higher success in all twelve camera comparisons (Table 4).

For object motion, camera-compensated CoTracker3 readouts (Karaev et al., 2024a) provide an independent comparison (Section D). Compared with EAP-IG top-36, Axis-specific selection has higher geometric sign-correctness in the two selection-independent negative-direction sufficiency cells: 91.2% versus 63.8% ($N = 80$ per selector), and across all four necessity directions: 82.5% versus 68.1% ($N = 160$). The same ordering appears in dense optical flow across all five motion families: sufficiency sign-correctness is 80.7% [77.0, 84.5] versus 62.7% [58.0, 67.5] ($N = 400$ per selector). For camera motion, the target-to-off-axis magnitude ratio likewise favors Axis-specific

Table 3: Per-axis target and cross-axis responses under activation exchange ($N = 40$ per axis). Target Δ is the median selected-axis response. STOR is the median ratio of signed target response to mean absolute off-axis response plus ϵ, including the opposite direction. Positive STOR indicates the expected target direction.

	Sufficiency				Necessity			
	Axis-specific		EAP-IG top-36		Axis-specific		EAP-IG top-36	
Axis	Target Δ	STOR	Target Δ	STOR	Target Δ	STOR	Target Δ	STOR
cam pan +	+13.43	+2.50	+9.12	+1.63	+13.42	+2.57	+10.14	+2.02
cam pan −	+12.89	+2.40	+4.30	+0.98	+13.62	+2.31	+8.72	+2.21
cam tilt +	+11.76	+2.03	+1.00	+0.28	+13.68	+2.22	+6.77	+1.77
cam tilt −	+13.18	+2.07	+1.24	+0.28	+13.02	+2.26	+10.87	+1.75
cam zoom +	+10.35	+2.39	+1.00	+0.45	+9.80	+2.03	+3.17	+1.31
cam zoom −	+14.88	+2.97	+2.62	+0.60	+15.01	+2.94	+1.29	+0.82
obj lateral +	+0.61	+0.26	−1.48	−0.53	+1.09	+0.40	+1.21	+0.61
obj lateral −	+5.76	+1.57	−0.18	−0.12	+8.58	+1.64	+2.64	+0.97
obj depth +	+0.99	+0.60	−0.50	−0.20	+1.99	+1.12	+0.78	+0.52
obj depth −	+10.96	+1.83	+1.67	+0.63	+8.93	+2.04	+0.50	+0.19

Table 4: External camera-motion sanity with VBench CameraPredict. Success requires target-label presence in sufficiency and absence in necessity. Entries are percentage (count), with $N = 40$ per axis. Empty predictions map to static. The aggregate pools six camera directions.

	Sufficiency		Necessity	
Axis	Axis-spec succ.	EAP-IG succ.	Axis-spec succ.	EAP-IG succ.
cam pan +	92.5% (37/40)	75.0% (30/40)	87.5% (35/40)	37.5% (15/40)
cam pan −	72.5% (29/40)	45.0% (18/40)	87.5% (35/40)	57.5% (23/40)
cam tilt +	80.0% (32/40)	52.5% (21/40)	47.5% (19/40)	35.0% (14/40)
cam tilt −	90.0% (36/40)	52.5% (21/40)	55.0% (22/40)	37.5% (15/40)
cam zoom +	67.5% (27/40)	35.0% (14/40)	82.5% (33/40)	55.0% (22/40)
cam zoom −	85.0% (34/40)	45.0% (18/40)	80.0% (32/40)	45.0% (18/40)
Aggregate	81.2% (195/240)	50.8% (122/240)	73.3% (176/240)	44.6% (107/240)

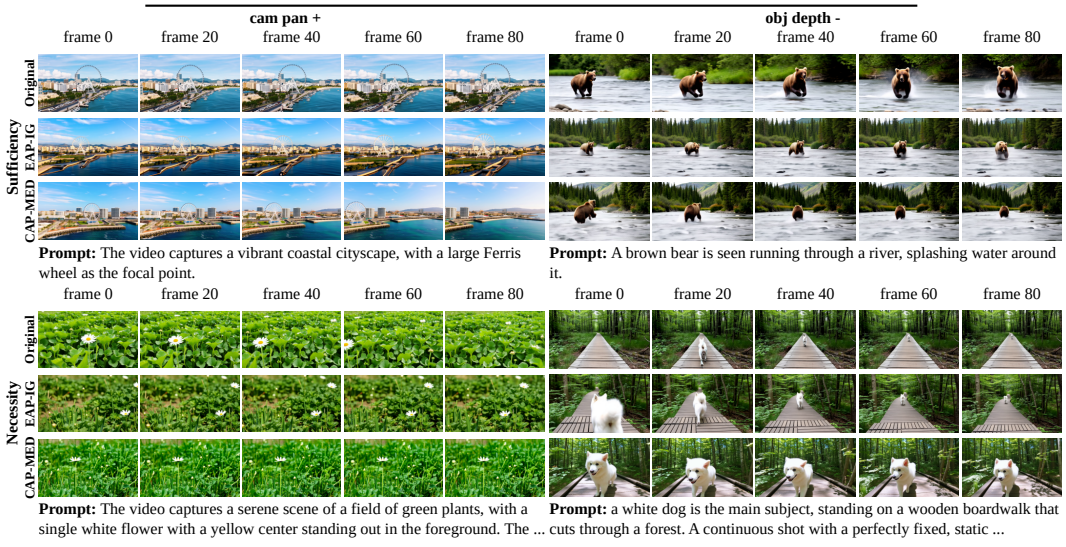


Figure 3: Camera and object motion under activation exchange. Each panel shows frames 0, 20, 40, 60, and 80. "Original" is the recipient, the empty-action video in sufficiency (top) and the instructed video in necessity (bottom). EAP-IG and CAP-MED denote EAP-IG top-36 and the Axis-specific selector at 36 nodes. The recipient and two patched videos are shown, without a donor row. Scene descriptions are listed in Section I.

selection, at 4.61 versus 1.16. Although the readouts agree on the selector ordering, object readouts show lower directional agreement on unpatched videos than camera readouts.

We also compare patched flow time courses with their own instructed source and with other instructed sources from the same motion cell. Median own-source correlation is larger in all five families. For pan, the own-source and other-source medians are 0.955 and 0.020, and the own-source correlation is larger on 98.8% of trials. Patched motion follows the matched source more closely than the same-cell reference profile. Figure 3 shows camera and object examples, with further sequences in Section I. Because full-field exchange can also transfer appearance, Section E evaluates video quality relative to the instructed and empty-action references.

4.6 Head-level dependence and operand exchange

To examine contributions within attention, we perform head-step interventions on separate selected sets. A matched $K = 100$ comparison favors a combined step-balanced, target-vs-other selector over global attribution ranking in camera-pan sufficiency and object-lateral sufficiency and necessity (Section G). Paired confidence intervals exclude zero in all three cells, which test step balancing and specificity jointly. Two further experiments use different head sets.

Dependence between two sets. We apply two disjoint sets in sequence. After inserting instructed activations at the first set in an empty-action rollout, restoring the second set to its empty-action state largely blocks recovery. In the reverse arm, restoring the second set to its instructed state reverses suppression induced at the first set (Table 5). Across four motion families, all eight changes follow this pattern, with median reductions of 96.5 and 95.7 percentage points in the two arms. The effect induced at the first set depends on the state supplied at the second under these exchanges.

Table 5: Two-set composed activation exchange. All entries report 100ρ.

	Sufficiency		Necessity	
Axis	First-set injection	After second-set empty-action restoration	First-set removal	After second-set instructed restoration
Camera pan	98.2	1.1	100.6	2.0
Camera tilt	100.5	4.6	101.2	-0.8
Camera zoom	93.4	11.5	87.7	-2.4
Object lateral	98.1	-1.7	101.6	8.8

Routing and value operands. For a fixed EAP-IG-ranked Pan+ top-300 head-step set, we write the pre-output head state as

$$Z_h = \mathrm{softmax}(Q_h K_h^\top / \sqrt{d})\, V_h. \tag{10}$$

We exchange normalized QK operands or V before output projection on ten new matched scenes. Value exchange gives classifier median recovery/suppression of 0.974/0.956, compared with 0.001/0.010 for QK exchange. Independent flow gives the same magnitude ordering (Table 6). Value-only exchange still uses recipient routing. The larger value-exchange effect therefore does not imply that routing is unnecessary.

Table 6: Operand exchange within a selected Pan+ top-300 head-step set on ten matched scenes.

	Classifier		Independent flow	
Exchanged operand	Suff. ρ, SC	Nec. ρ, SC	Suff. ρ, SC	Nec. ρ, SC
QK routing	0.001, 60%	0.010, 60%	0.052, 80%	0.205, 70%
V value	0.974, 90%	0.956, 100%	0.711, 80%	0.900, 100%
Full QKV	0.948, 80%	0.942, 100%	0.770, 90%	1.019, 100%

Each cell reports median normalized effect and sign-correctness. QK denotes the normalized query/key operands described by the intervention, not a separately reported post-softmax attention-matrix patch.

4.7 Cross-Architecture and Non-Motion Validation

Cross-architecture validation. Instead of the Wan layer-grid rule, it ranks individual nodes by a magnitude-balanced specificity score, which combines target-attribution magnitude with target-vs-other specificity so that selected nodes are both strongly attributed to the target direction and relatively

specific to it (Section H.2). Attribution uses fifty discovery pairs per signed camera direction over 840 step-layer-submodule nodes. Both selectors exchange 36 nodes on forty discovery-disjoint pairs per direction, using signed optical flow for evaluation. All attribution scores and selected nodes are recomputed on LTX. An earlier L2-ranked head experiment evaluates the exchange stage alone (Section H.1).

Magnitude-balanced specificity gives higher pooled recovery and suppression (Table 7), but raw target-attribution ranking gives larger median effects for both pan directions in both arms. The pooled values are point estimates without uncertainty intervals.

Table 7: LTX-native full-workflow comparison at $K = 36$ nodes.

	Sufficiency		Necessity	
Selector	SC (%)	Recovery (%)	SC (%)	Recovery (%)
Magnitude-balanced specificity	88.75	27.96	96.25	83.73
Raw EAP-IG	85.42	22.55	95.83	82.21

Forty evaluation pairs per signed direction give 240 trials per selector and arm. Recovery is the reported median normalized effect in percentage units, 100ρ. The pooled results are point estimates. Uncertainty intervals are not reported for this comparison.

Color-temperature validation. To test a non-motion target, we use warmer/cooler prompt pairs matched in scene and seed, with frame-mean CIELAB b^* as an appearance readout. All selectors rank the same 840 nodes and exchange 36 on discovery-disjoint pairs. Specificity gives greater recovery and suppression than EAP-IG and uniform random selection (Table 8). Paired specificity-minus-EAP gains have positive 95% intervals clustered over forty shared scenes, and a secondary RGB readout gives the same ordering (Section H.3). All selectors are sign-correct on every trial, so effect magnitude separates them. The remaining appearance directions are attribution contrasts, not intervention targets.

Table 8: Color-temperature intervention on matched LTX pairs at $K = 36$ nodes.

	Sufficiency			Necessity		
Selector	SC	Median ρ	Induced Δb^*	SC	Median ρ	Removed Δb^*
Uniform random	100%	0.228	6.21	100%	0.249	6.64
EAP-IG	100%	0.583	14.71	100%	0.699	19.36
Specificity	100%	0.802	19.45	100%	0.819	21.58

5 Discussion and conclusion

The Wan2.2 experiments show that motion-effect node sets are better identified by target-vs-other specificity than by target attribution magnitude alone. At the same node budget, specificity-selected sets produce substantially stronger recovery and suppression, and matched-grid controls show that this advantage is not explained simply by using the same steps and submodules. Independent motion readouts reproduce the same ordering, supporting the interpretation that the selected sets carry motion-specific intervention effects. The main exception is object lateral+, whose weaker effect at 36 nodes largely recovers as the layer budget increases.

Overall, the results support a view in which motion effects are concentrated in compact but direction-dependent sets of internal activations rather than in the largest-attribution components alone. Head-level interventions further show conditional dependence between selected sets and a value-dominant contribution within camera-pan heads. The LTX experiments reproduce the full workflow on a second architecture, while also showing that the selector advantage is not uniform across directions. Because full-field exchange can transfer non-motion information, these results do not establish content-preserving editing or a minimal motion circuit. A finer decomposition of these fields, together with tests of compositional, articulated, and multi-object motion, remains future work.

AI USE STATEMENT

The experiments analyze videos generated by Wan2.2 and LTX-Video. Qwen2.5-VL-7B provides an auxiliary depth-motion readout, as described in Section D. Large language models assisted with implement methods, manuscript drafting and language editing, organization, interpretation and scope of previously reported results, and LaTeX formatting. The authors are responsible for the scientific claims and reported results.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Rc7y9HFC34.

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. URL https://openreview.net/forum?id=_CDixzkzeyb.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KI1zldOFz9.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024b.

Youngjun Jun, Seil Kang, Woojung Han, and Seong Jae Hwang. I'm a map! interpretable motion-attentive maps: Spatio-temporally localizing concepts in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024a.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *European Conference on Computer Vision*, 2024b.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4659–4669, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. Atp*: An efficient and scalable method for localizing llm behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023a.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023b.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, and Ashwinee Panda. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37:76115–76138, 2024.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2z1HT5lw5M.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18938–18949, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023.

A EXPERIMENTAL PROTOCOL

A.1 GENERATION AND EVALUATION PAIRS

The main Wan2.2 T2V-A14B experiments use 81-frame videos at 480×832 with ten FlowUniPC steps and 40 transformer layers. Each comparison uses source and base rollouts matched in scene, seed, schedule, and generation settings. The source adds the target motion instruction and the base leaves its action slot empty. There are 40 evaluated pairs per signed direction and intervention arm in the ten-direction comparison, giving 400 trials per selector and arm and 1,600 selector-specific intervention trials in total. The six camera directions are pan right/left, tilt up/down, and zoom in/out. The four object directions are lateral left-to-right/right-to-left and depth toward/away from the camera.

A.2 ATTRIBUTION AND MODULE CHOICE

The Wan2.2 attribution configuration uses ten prompt pairs, ten sampled denoising steps, and five integrated-gradient points. Both selectors use the same attribution graph, with CA, SA, and FFN included in scoring and layer aggregation. The default Axis-specific intervention grid is $\{0, 2, 4\} \times \mathcal{L}_a^{\text{spec}} \times \{\text{SA}, \text{FFN}\}$ with six selected layers. The baseline ranks individual nodes by raw target-axis EAP-IG magnitude and selects 36.

Across ten signed directions, CA accounts for 6.1% of total attribution mass, compared with 61.6% for SA and 32.3% for FFN. CA mass ranges from 5.0% to 7.2% across individual directions. Unconstrained top-36 selection includes no CA node in nine directions and one CA node in the remaining camera direction. Thus, CA contributes one of 360 selected nodes, with mean counts 0.1 CA, 24.4 SA, and 11.5 FFN per direction. These attribution measurements motivate the default patch grid, but do not establish that CA is functionally absent. The direct CA-addition intervention is reported in Section F.3.

Table 9: Attribution mass and unconstrained selected-node composition across ten signed directions.

Statistic	CA	SA	FFN
Mean share of total attribution mass (%)	6.1	61.6	32.3
Mean nodes in unconstrained top-36	0.1	24.4	11.5

A.3 SELECTED LAYERS

Figure 2 visualizes and Table 10 lists the six layers selected by Equation (4) for each signed direction ($K = 6$), with the target-vs-other specificity ratio in parentheses (reference value 1.0). Camera selections mostly lie at middle or later depths and have ratios above one; lateral-motion ratios are below one. These are attribution diagnostics. Intervention evidence is reported separately.

A.4 AGGREGATION AND UNCERTAINTY

The reported normalized ratio is the median of the per-pair ρ values defined in Equation (8), not a ratio formed after separately aggregating numerator and denominator. Sign-correctness averages the per-pair indicators. The aggregate table reports 95% bootstrap intervals with 10,000 resamples over 400 trials. Its marginal intervals appear in Table 1. The aggregate paired-bootstrap comparison gives nonzero differences for both sign-correctness and normalized ratio in both intervention directions. A median within-pair difference need not equal the difference between two marginal medians. Specific paired estimates and confidence intervals are reported separately for the head comparison, CA-addition experiment, and color-temperature test. The color-temperature intervals use scene-clustered resampling over 40 shared scene identities. These aggregation and resampling units should not be conflated.

Table 10: Selected layers and their motion-specificity scores.

Axis direction	Layer index (specificity ratio)					
	1	2	3	4	5	6
Camera motion						
cam pan+	39 (2.60)	37 (2.23)	23 (1.83)	33 (1.83)	16 (1.81)	27 (1.79)
cam pan−	38 (2.19)	37 (1.87)	34 (1.60)	0 (1.59)	36 (1.58)	7 (1.46)
cam tilt+	39 (1.31)	16 (1.24)	33 (1.23)	25 (1.21)	12 (1.21)	15 (1.21)
cam tilt−	39 (1.87)	22 (1.82)	33 (1.77)	12 (1.68)	16 (1.65)	31 (1.64)
cam zoom+	38 (1.75)	26 (1.44)	14 (1.40)	19 (1.39)	35 (1.37)	13 (1.32)
cam zoom−	29 (1.65)	35 (1.52)	32 (1.38)	31 (1.31)	30 (1.29)	2 (1.23)
Object motion						
obj lateral+	23 (0.58)	5 (0.58)	26 (0.57)	10 (0.56)	15 (0.55)	31 (0.55)
obj lateral−	34 (0.66)	32 (0.63)	33 (0.61)	28 (0.60)	25 (0.59)	35 (0.59)
obj depth+	11 (1.01)	9 (0.85)	12 (0.88)	17 (0.88)	19 (0.86)	4 (0.82)
obj depth−	29 (1.31)	28 (1.21)	38 (1.18)	21 (1.16)	22 (1.16)	32 (1.14)

Each entry gives the layer index followed by its target-versus-other specificity ratio in parentheses. Six layers are selected per direction. The reference ratio is 1.0. The six positions preserve the recorded order and do not imply a sorted score sequence.

A.5 Compute

On one NVIDIA H200, a representative 20-trial intervention cell required 3.6 GPU-hours for sufficiency or 4.4 GPU-hours for necessity, including pair screening. Generation took about 2.4 minutes per video. Across ten signed directions, each attribution run used ten prompt pairs, ten sampled steps, and five integration points and took 19.15–19.64 wall-clock hours, with median 19.30 hours and total 193.2 H200 GPU-hours. The 40-scene, three-axis LTX activation-exchange rerun took 2.2 GPU-hours in total. These figures describe the reported workloads rather than a measured total for the entire research project.

A.6 Motion instructions

Table 11: Motion-instruction text used in the reported examples.

Direction	Template
cam pan +	The subjects in the scene remain in their initial states. The perspective executes a smooth, continuous panoramic sweeping motion, rotating horizontally to the right. This motion progressively reveals new, unseen parts of the environment as they continuously enter the frame from the right side. The viewpoint shifts dynamically.
cam pan −	The subjects in the scene remain in their initial states. The perspective executes a smooth, continuous panoramic sweeping motion, rotating horizontally to the left. This motion progressively reveals new, unseen parts of the environment as they continuously enter the frame from the left side. The viewpoint shifts dynamically.
cam tilt +	The subjects in the scene remain in their initial states. The field of view performs a slow, sweeping tilt up motion, shifting vertically. This vertical sweep progressively reveals the upper portions of the environment as they are introduced into the frame from the top edge, while the lower area gradually disappears below.
cam tilt −	The subjects in the scene remain in their initial states. The field of view performs a slow, sweeping tilt down motion, shifting vertically. This vertical sweep progressively reveals the lower portions of the environment as they are introduced into the frame from the bottom edge, while the upper area gradually disappears above.

Motion-instruction text (continued)

Direction	Template
cam zoom +	The scene is relatively static. A continuous, cinematic optical zoom in begins slowly. The frame progressively magnifies the central subject. The depth of field compresses, and the background becomes larger and more blurred as the specific details of the central object fill the entire view.
cam zoom −	The scene is relatively static. A continuous, cinematic optical zoom out begins slowly from a close-up. The frame progressively widens to reveal the entirety of the central subject and the expansive surrounding environment.
obj lateral +	A continuous shot with a perfectly fixed, static viewpoint. The main subject begins moving steadily, traveling horizontally across the frame from the far left side to the far right edge. The background remains perfectly still.
obj lateral −	A continuous shot with a perfectly fixed, static viewpoint. The main subject begins moving steadily, traveling horizontally across the frame from the far right side to the far left edge. The background remains perfectly still.
obj depth +	A continuous shot with a perfectly fixed, static viewpoint. The main subject physically walks or travels directly forward through the 3D environment along the z-axis, changing its spatial relationship with the stationary background as it approaches the foreground. The background itself does not change scale, but the subject progressively occupies more of the frame due to its physical traversal towards the lens.
obj depth −	A continuous shot with a perfectly fixed, static viewpoint. The main subject physically walks or travels directly away from the lens into the deep background along the z-axis. As it physically traverses the distance, its figure occupies progressively less of the frame, becoming smaller and eventually merging with the far background. The overall field of view remains wide and constant; only the subject moves through the depth.
empty action	(blank, with the scene sentence alone)

This table separates the motion-instruction suffixes from the scene sentences given in Section I. The empty-action condition contains the scene without the motion suffix.

B Motion classifiers and annotation

B.1 Camera-motion classifier

The camera readout is a VideoMAE classifier with seven classes: six signed camera directions and static. It is trained on 5,756 annotated Wan2.2-generated videos using a stratified 50/50 split, with 2,878 videos for training and 2,878 for validation. The backbone is `MCG-NJU/videomae-base-finetuned-kinetics`, fine-tuned with a seven-way head. Each input contains 16 uniformly sampled RGB frames resized to 224×224. Training first updates the classifier head and top encoder layers, then unfreezes additional upper encoder layers.

Validation accuracy is 98.99%, with macro F1 0.990 and 29 off-diagonal errors out of 2,878 videos. Static recall is 100%, but the static labels are derived from VBench static predictions. This limits the independence of that class-specific comparison. Tables 12 and 13 retain the camera-classifier and VBench comparisons on the same videos. VBench uses lenient matching, which counts a prediction as correct when the expected label appears in the predicted label set.

Table 12: Camera-classifier and VBench accuracy on the same validation videos. Motion-only excludes static.

Evaluation set	Classifier	VBench
Seven-class accuracy	0.9899	0.9350
Motion-only accuracy	0.9879	0.9221

Table 13: Joint correctness of the camera classifier and VBench on 2,878 validation videos.

	VBench correct	VBench wrong	Total
Classifier correct	2673	176	2849
Classifier wrong	18	11	29
Total	2691	187	2878

Among VBench-correct videos, the classifier is also correct on 99.33%. Among VBench-wrong videos, it is correct on 94.12%. VBench accuracies for pan left, zoom in, and zoom out are 88.96%, 87.31%, and 88.18%, respectively. The classifier improves these categories by roughly 9–11 percentage points. These validation results support the in-domain readout but do not replace independent evaluation of intervention effects.

B.2 Object-motion classifier

The object readout uses a Kinetics-400-pretrained VideoMAE backbone and a four-class head for lateral and depth directions. It is fine-tuned on 1,114 videos and evaluated on a separate stratified split of 1,114 videos from a 2,228-video human-annotated pool of Wan2.2 generations. The pool comes from a 10,000-video annotation pass in which outputs are confirmed, relabeled, or discarded according to realized rather than merely requested motion. Reviewers of the generated videos see the instruction during annotation. Held-out accuracy is 95.42%, with per-class F1 scores between 0.936 and 0.965. Seventeen validation errors cross the lateral/depth families, representing 1.5% of the held-out split. Twenty errors reverse lateral direction and fourteen reverse depth direction. The independent tracking and depth checks in Section D complement this in-domain evaluation.

B.3 Prompt realization

The annotation pass relabels 18.1% of the 4,000 object-prompted videos and 3.2% of the 6,000 camera-prompted videos. These rates measure mismatch between the requested and realized labels in the annotated pool. They do not measure the success rate of an activation intervention. The main empirical scope is the four signed object directions and their fixed instruction templates, rather than arbitrary object behavior.

Table 14: Relabeling rates in the generated-video annotation pass.

Prompt family	Pool size	Relabeling rate
Object	4,000	18.1%
Camera	6,000	3.2%

C Selector controls

Table 2 compares five selectors on six held-out signed directions: pan$\pm$, tilt$\pm$, and object lateral$\pm$. Every selector exchanges 36 nodes. The Axis-specific selector uses the six specificity-ranked layers on the fixed grid $\{0, 2, 4\} \times \{\mathrm{SA}, \mathrm{FFN}\}$. The random-layer control samples six non-Axis-specific layers without replacement and instantiates them on that same grid. The high-attribution/low-specificity control excludes the Axis-specific layers, considers the top twelve layers by target-axis EAP-IG mass, and takes the six least specific layers from that pool. It also uses the same grid. The raw EAP-IG and linear-probe selectors each select 36 nodes.

The random control tests layer choice at a fixed node budget. The high-attribution/low-specificity control retains a high-attribution candidate pool while selecting against specificity. Its median ratios are 0.07 in sufficiency and 0.35 in necessity, compared with 0.95 and 1.00 for Axis-specific selection. The linear-probe selector reaches 0.01 and 0.34. Thus, neither high attribution alone nor the tested probe ranking matches the bidirectional effects of Axis-specific selection. The result does not establish an isolated causal effect of every selector component, because the global node and probe selectors need not have the same allocation as the layer-grid selectors.

D INDEPENDENT MOTION READOUTS

D.1 CAMERA-COMPENSATED OBJECT TRACKING AND DEPTH JUDGMENTS

All 640 archived object-axis intervention trials are re-scored with camera-compensated CoTracker3 readouts, keeping the submitted scoring and selection fixed. The geometric measure uses foreground tracking for lateral motion and its depth component for depth motion. Across 160 distinct source/base pairs per object-motion family, represented by 320 selector-duplicated trial rows, total-effect signs agree with the classifier on 74.4% of lateral and 70.0% of depth pairs. The selection-independent sufficiency check covers the two negative-direction cells, with 80 trials per selector. Necessity covers all four signed directions, with 160 trials per selector. Both selectors use matched scenes, seeds, and 36-node budgets.

Table 15: Independent object-motion comparisons on matched intervention videos.

Readout and scope	Arm	Axis-specific	EAP-IG top-36
CoTracker3, two negative directions ($N = 80$)	Suff.	91.2%	63.8%
CoTracker3, all four directions ($N = 160$)	Nec.	82.5%	68.1%
Qwen2.5-VL depth, two signed directions	Suff.	53.8%	26.3%
Qwen2.5-VL depth, two signed directions	Nec.	60.0%	36.3%

CoTracker3 reports geometric sign-correctness. Qwen2.5-VL reports directional success. N is per selector. Different readouts need not share the same event definition.

The paired discordant CoTracker3 outcomes favor Axis-specific selection by 24 versus 2 in the selection-independent sufficiency comparison and by 36 versus 13 in all-direction necessity. The geometric criterion and total-effect sign share a video reference, so the matched selector contrast is the relevant comparison rather than a universal chance baseline.

A zero-shot Qwen2.5-VL-7B judge has 75.0% directional accuracy on 320 instructed depth judgments. It fails directional pre-validation for lateral motion, with approximately 0.47 agreement on instructed videos, and is therefore used only for depth in this evaluation. The depth success rates in Table 15 retain the selector ordering under a semantic readout. The judge's imperfect validation accuracy limits the interpretation of its absolute success rates.

D.2 DENSE OPTICAL FLOW

All 1,600 archived trials are re-evaluated without the learned classifier as the output readout. Dense optical flow is decomposed into global camera pan, tilt, and zoom plus a camera-compensated object residual. On unpatched videos, directional agreement is 90.0% for pan, 96.3% for tilt, 92.5% for zoom, 62.5% for object depth, and 67.5% for object lateral motion. These values show that object direction remains harder to measure independently.

Table 16: Independent motion evidence on the original matched videos.

Metric and scope	Axis-specific	EAP-IG top-36
VBench CameraPredict sufficiency, six camera cells	81.2%	50.8%
Dense-flow sufficiency sign-correctness, all five families	80.7%	62.7%
95% confidence interval	[77.0, 84.5]	[58.0, 67.5]
Dense-flow camera target/off-axis magnitude ratio	4.61	1.16

VBench and camera-only flow use 240 trials per selector. All-family flow uses 400 trials per selector and intervention arm. The sign-correctness intervals shown are the reported sufficiency intervals.

Across all sixteen object-axis dense-flow cells, with 40 trials per cell, normalized object-residual magnitude ranges from 0.321 to 1.404, compared with 0.009–0.049 for fitted camera motion. Per-cell median object-to-camera magnitude ratios range from 8.8 to 71.6, with a pooled median of approximately 34. This decomposition supports an object-dominated measured signal rather than a fitted camera-motion explanation in those cells.

D.3 SOURCE-SPECIFIC TEMPORAL PROFILES

A per-trial sufficiency analysis uses 80 trials per motion family. The five-frame-smoothed target-axis course of each patched video is correlated with its own instructed source and with the median course of ten other instructed rollouts from the same cell. Table 17 reports the median own-source and other-source correlations and the fraction with a larger own-source correlation. Axis-specific selection has higher own-source correlations and win rates on all five motion families. Pan amplitude also transfers nearly one-to-one, with slope 1.01 and correlation 0.99. The results argue against a shared temporal-onset template as a sufficient explanation, but do not imply that no other donor information is transferred.

Table 17: Temporal-profile agreement with matched and other instructed sources.

	Axis-specific			EAP-IG top-36		
Family	r_{own}	r_{other}	Own $>$ other	r_{own}	r_{other}	Own $>$ other
Camera pan	0.955	0.020	98.8%	0.406	−0.026	68.8%
Camera tilt	0.741	0.119	82.5%	0.348	0.041	66.3%
Camera zoom	0.814	0.030	88.8%	0.236	0.042	65.0%
Object lateral	0.679	0.010	83.8%	0.149	−0.004	58.8%
Object depth	0.337	0.033	66.3%	0.096	−0.020	60.0%

E VIDEO QUALITY AND REFERENCE COMPARISONS

The archived 40-trial-per-cell runs are evaluated with DINOv2, LPIPS, and FVD. Table 18 preserves the reported reference comparisons. The necessity margins compare preservation relative to the empty-action condition. The FVD evaluations use MD5-deduplicated reference halves shared across selectors, with the same qualitative ordering reported separately for camera and object trials. These measurements complement motion scores but are not a motion-invariant definition of content preservation.

Table 18: Reference-based quality comparisons on archived intervention videos.

Metric	Scope	Axis-specific	EAP-IG	Preferred
DINOv2 empty-action preservation margin	Camera necessity	+0.344	+0.079	Higher
LPIPS empty-action preservation margin	Pooled necessity	+0.329	+0.082	Higher
LPIPS distance from patched to empty-action video	Pooled sufficiency	0.512	0.523	Lower
FVD to empty-action reference	Pooled necessity	684	958	Lower
FVD to instructed reference	Pooled sufficiency	328	443	Lower

Six reference-free VBench dimensions are also evaluated on all 1,600 trials. Each table entry is a mean per-trial delta. For sufficiency, the three values are Axis-specific patched minus empty action, EAP patched minus empty action, and instructed minus empty action. For necessity, the two values are Axis-specific patched minus instructed and EAP patched minus instructed. The associated classifier median ratios are approximately 0.99/0.15 in sufficiency and 0.97/0.59 in necessity.

Table 19: Reference-free VBench changes relative to the recipient condition.

Dimension	Sufficiency: Axis / EAP / instructed	Necessity: Axis / EAP
Subject consistency	−0.0379/ − 0.0172/ − 0.0432	+0.0399/ + 0.0300
Background consistency	−0.0250/ − 0.0119/ − 0.0268	+0.0234/ + 0.0202
Temporal flickering	−0.0167/ − 0.0162/ − 0.0180	+0.0172/ + 0.0117
Motion smoothness	−0.0068/ − 0.0069/ − 0.0065	+0.0061/ + 0.0038
Imaging quality	−3.6356/ − 0.7703/ − 4.89	+4.2561/ + 3.9156
Aesthetic quality	−0.0369/ − 0.0178/ − 0.0478	+0.0410/ + 0.0404

Axis-specific sufficiency follows the measured quality changes of the instructed reference while recovering substantially more motion than EAP-IG. The smaller EAP-IG deltas occur alongside

substantially weaker induced motion. Necessity shifts every reported dimension toward the empty-action condition. Quality changes should therefore be interpreted jointly with the amount and direction of induced motion. The results do not establish that all perceptual properties are preserved, nor do the tables estimate a separate quality-per-unit-motion statistic.

F TIMESTEP, LAYER, BUDGET, AND SUBMODULE SENSITIVITY

F.1 TIMESTEP AND FIXED-LAYER COMPARISONS

The pan+ sufficiency comparisons use 40 original evaluation pairs. The one-step settings use six selected layers and two submodules, giving twelve nodes. The $\{0, 2, 4\}$ setting uses 36 nodes. Each three-layer comparison uses three steps and two submodules, giving eighteen nodes. Table 20 preserves these different budgets rather than treating every row as a fixed-budget comparison.

Table 20: Pan+ sufficiency sensitivity on 40 pairs.

Ablation	Setting	Nodes	Sign-correctness	Median ρ
Timestep	$\{0\}$	12	57.5%	0.03
Timestep	$\{2\}$	12	97.5%	0.99
Timestep	$\{0, 2, 4\}$	36	100%	1.00
Layer set	$[37, 38, 39]$	18	95.0%	1.00
Layer set	$[15, 16, 17]$	18	40.0%	-0.02
Layer set	$[5, 22, 35]$	18	37.5%	-0.06

The same single-step versus $\{0, 2, 4\}$ contrast is reported for pan$-$, tilt+, and tilt$-$ on smaller samples of 8–10 pairs. Step $\{0\}$ gives 75–80% sign-correctness and median ratios 0.04–0.06, while $\{0, 2, 4\}$ gives 100% and 0.83–1.02. These smaller comparisons support timestep dependence, with their own denominators rather than the 40-pair denominator of Table 20.

F.2 LAYER-BUDGET SWEEP

The layer budget $K \in \{3, 6, 9, 12\}$ gives 18, 36, 54, and 72 nodes at fixed steps $\{0, 2, 4\}$ and submodules $\{\mathrm{SA}, \mathrm{FFN}\}$. Original 40-pair samples are used except for tilt+ necessity, which uses twenty held-out pairs. Each cell reports sign-correctness and median ρ.

Table 21: Layer-budget sensitivity at fixed step and submodule allocation.

Direction / arm and sample	$K = 3$	$K = 6$	$K = 9$	$K = 12$
Pan+ sufficiency, original $N = 40$	100%, 1.00	100%, 1.00	100%, 1.00	92.5%, 1.00
Pan+ necessity, original $N = 40$	100%, 1.03	100%, 0.99	100%, 1.01	100%, 1.01
Tilt+ sufficiency, original $N = 40$	85.0%, 0.78	92.5%, 0.88	95.0%, 0.97	97.5%, 0.98
Tilt+ necessity, held-out $N = 20$	80.0%, 0.80	95.0%, 0.97	90.0%, 1.00	90.0%, 1.02
Lateral+ sufficiency, original $N = 40$	57.5%, 0.06	77.5%, 0.34	90.0%, 0.95	95.0%, 1.06
Lateral+ necessity, original $N = 40$	67.5%, 0.12	75.0%, 0.43	80.0%, 0.85	97.5%, 1.00

Pan+ retains near-full effects across this budget range. Tilt+ sufficiency approaches saturation as the budget increases, while object lateral+ shows the strongest budget sensitivity. The latter reaches near-full median recovery and suppression at $K = 12$ without changing submodules. The result characterizes this fixed-budget limitation and does not establish minimality of the selected sets.

F.3 DIRECT CA-ADDITION INTERVENTION

One frozen camera-pan+ sufficiency cell tests the consequence of adding all eighteen aligned CA nodes at the same six layers and three denoising steps as the default grid. The default SA/FFN set contains 36 nodes and the augmented set contains 54. The evaluation uses forty within-pair comparisons.

Table 22: Direct CA-addition test on camera-pan+ sufficiency.

Grid	Nodes	Median recovered effect (100ρ)	Sign-correctness
SA, FFN	36	100.11	100%
SA, FFN, CA	54	100.28	97.5%

The median of the forty within-pair changes is 0.00 percentage points, with a 95% paired-bootstrap interval of $[-0.265, +0.221]$ from 10,000 resamples. The 0.17-point difference between marginal medians is not this paired estimator. Adding CA increases intervention size by 50% without changing the paired median effect and flips the sign of one trial. This directly supports the submitted SA/FFN prioritization in the tested cell. It is not a general claim that CA has no functional role in motion generation.

G Head-level interventions and operand analysis

G.1 Head-granularity validation and a matched-budget comparison

Attribution-ranked head sets retain activation-exchange support across five motion families. The reported median ratios range from 0.94 to 1.00 in sufficiency and from 0.92 to 1.01 in necessity, with screened samples of 28–46 pairs and approximate saturation at head budgets of 300–5,000. These head-granularity tests refine the intervention unit without establishing channel- or edge-level decomposition.

A separate matched $K = 100$ comparison evaluates a combined step-balanced, target-vs-other selector against global EAP-IG head ranking. Table 23 retains the three reported comparisons on forty original evaluation pairs per cell. The median paired gains and confidence intervals are not differences of the marginal medians. Both sufficiency intervals and the object-lateral necessity interval are above zero. The object-lateral necessity comparison also reports McNemar $p = 0.0018$ for paired sign-correctness. Because the selector combines step balancing and target-vs-other selection, these results do not identify the contribution of either component in isolation.

Table 23: Head-level comparison at a matched budget of $K = 100$, with $N = 40$ per cell.

Axis / intervention	Selector SC, $\widetilde{\rho}$	EAP SC, $\widetilde{\rho}$	Paired $\Delta\rho$ [95% CI]	Win rate
Camera pan / suff.	72.5%, 0.436	50.0%, -0.004	$+0.194$ $[+0.003, +0.438]$	65.0%
Object lateral / suff.	92.5%, 1.016	77.5%, 0.660	$+0.200$ $[+0.112, +0.439]$	72.5%
Object lateral / nec.	90.0%, 0.986	60.0%, 0.533	$+0.332$ $[+0.153, +0.675]$	82.5%

SC is sign-correctness. The selector combines step-balanced and target-vs-other selection. Paired differences and win rates compare it with global EAP-IG ranking.

G.2 Composed activation exchange

The composed intervention applies two disjoint selected head sets in sequence. In the first arm, instructed activations are inserted at the first set during an empty-action rollout, then the second set is restored to its empty-action state. In the reverse arm, empty-action activations are inserted at the first set during an instructed rollout, then the second set is restored to its instructed state. Table 5 reports the resulting normalized effects in percentage units.

Restoring the second set to its empty-action state reduces recovery by 81.9–99.8 percentage points, with a median reduction of 96.5. Restoring it to its instructed state reverses first-set suppression by 90.1–102.0 points, with a median reduction of 95.7. All eight changes are in the predicted direction. The effect introduced or removed through the first set therefore depends on the state supplied at the second set under this composed intervention. This is evidence of two-stage conditional functional dependence, not recovery of the algorithm within either set or a complete motion mechanism.

G.3 Operand exchange within selected Pan heads

A separate experiment uses the pre-output head factorization in Equation (10). It independently exchanges the normalized QK routing operands and the V value operands before output projection at a fixed EAP-IG-ranked Pan+ top-300 head-step set. Evaluation uses ten new matched scenes and the same bidirectional activation-exchange definitions. In sufficiency, the source operand is inserted into the base rollout. In necessity, the operand is restored to its base state within the source rollout. Classifier and independent-flow readouts evaluate the same intervention.

Source-value exchange recovers and suppresses most of the classifier-measured Pan effect, whereas query/key exchange produces near-zero median classifier effects. Independent flow preserves the ordering, although its query/key effects are nonzero and value sufficiency does not equal full-state sufficiency. This supports a value-dominant account of the measured effect within the selected Pan head set. It does not make routing unnecessary, because the recipient routing computation remains active when only V is exchanged. It also does not establish the same operand balance for other motion directions or architectures. The head-budget, composed, and operand tests concern distinct selected sets, rather than a single complete decomposition of the main 36-node hypothesis.

H LTX-Video and non-motion validation

H.1 Activation-exchange validation with a frozen head ranking

LTX-Video 2B is a 28-layer rectified-flow architecture distinct from the Wan2.2 model used in the main experiments. The first LTX test uses an early-window L2 activation-difference ranking and evaluates bidirectional head-level activation exchange on camera pan, tilt, and zoom with model-free optical flow. This experiment tests the exchange-validation stage, not target-vs-other attribution selection.

The eligible early-window pool contains 10,752 SA/CA head-step nodes, within a full hooked space of 35,840 across twenty steps. A budget of 300 exchanges 0.84% of the full space or 2.79% of the eligible pool. A budget of 1,000 exchanges 2.79% of the full space or 9.30% of the eligible pool. The ranking is frozen after the initial twelve scenes. Twenty-eight further scenes with source-base effect above 1.0 are added without refitting, and all forty are reported. All twelve intervention cells in Table 24 have 100% sign-correctness. This is conditional validation on the evaluated scenes, not an unfiltered generation success rate.

Table 24: Frozen L2 head ranking on LTX-Video. Each camera family uses forty evaluated scenes.

Axis	Head budget K	Sufficiency median ρ	Necessity median ρ
Pan	300	0.935	0.958
Pan	1,000	0.970	0.987
Tilt	300	0.923	0.999
Tilt	1,000	0.963	1.000
Zoom	300	0.834	0.969
Zoom	1,000	0.941	0.998

Every sufficiency and necessity cell is 100% sign-correct. Head budgets are not interchangeable with the 36-submodule-node budget of the full-workflow test.

H.2 All three workflow stages with LTX-native attribution

The full-workflow experiment executes attribution proposal, target-vs-other selection, and activation-exchange validation directly on Pan+, Pan−, Tilt+, Tilt−, Zoom+, and Zoom−. For each signed direction, EAP-IG is computed from fifty LTX discovery pairs using frozen LTX camera-axis flow-probe directions. Both selectors rank the same predeclared universe of ten sampled denoising steps, all 28 layers, and SA, CA, and FFN, giving 840 individual nodes. No smaller layer or module grid is imposed, and both selectors use $K = 36$.

The target-vs-other score combines magnitude and specificity through their geometric mean. If $M_d(j)$ is the target-direction attribution magnitude at node j and $D_d(j)$ is the stabilized mean same-node

magnitude of the other five directions, the score is

$$S_d^{\mathrm{MB}}(j) = \sqrt{M_d(j)\frac{M_d(j)}{D_d(j)}} = \frac{M_d(j)}{\sqrt{D_d(j)}}. \tag{11}$$

There is no fitted mixing weight. Raw EAP-IG instead ranks $M_d(j)$. This is a native per-node, magnitude-balanced instantiation, not the same layer-ratio/grid construction used in the main Wan2.2 experiment.

Selected sets are evaluated on a fixed, discovery-disjoint registry of forty pairs per signed direction under both exchange arms. Prompts, node budget, intervention operator, and signed Farneback camera-axis readout are matched between selectors. No Wan-derived score or selected node identity is transferred. Tables 7 and 25 report the aggregate and direction-resolved results of this full-workflow comparison.

Table 25: Direction-resolved LTX full-workflow results. Each cell gives sign-correctness and median recovery.

	Sufficiency		Necessity	
Direction	Specificity	Raw EAP-IG	Specificity	Raw EAP-IG
Pan+	100.0% / 33.81%	100.0% / 34.96%	90.0% / 71.12%	92.5% / 72.25%
Pan−	87.5% / 26.08%	82.5% / 26.81%	92.5% / 63.04%	90.0% / 66.40%
Tilt+	97.5% / 43.40%	95.0% / 20.15%	100.0% / 97.57%	97.5% / 93.90%
Tilt−	90.0% / 24.64%	82.5% / 19.62%	100.0% / 93.59%	97.5% / 91.71%
Zoom+	80.0% / 18.82%	82.5% / 8.85%	95.0% / 78.20%	97.5% / 71.65%
Zoom−	77.5% / 24.40%	70.0% / 15.00%	100.0% / 75.62%	100.0% / 73.10%

The specificity-selected sets have positive median recovery in every signed direction and both arms. They have higher pooled sign-correctness and median recovery point estimates than raw attribution. The direction-level table is not uniformly favorable: raw attribution has larger median effects in both pan directions and both arms, and sign-correctness orderings vary in some cells. The experiment establishes a native execution of the full workflow on a second architecture with the stated selector instantiation. It does not establish statistically significant dominance across architectures, seeds, or every motion direction.

H.3 Color temperature as a non-motion target

The color-temperature test evaluates warmer and cooler instructions on LTX-Video 2B. Source and base prompts share the scene and seed and explicitly request unchanged composition, viewpoint, and motion. Only the temperature instruction changes. These prompt constraints specify the intended contrast and do not certify that every other output property is invariant.

The primary readout is the deterministic, training-free frame-mean CIELAB b^* coordinate, with positive change indicating a warmer shift and negative change a cooler shift. The secondary readout is the RGB red-minus-blue opponent statistic $(R - B)/(R + B + 10^{-6})$, averaged over pixels and frames of the same videos. The secondary readout is not used for node selection, pair eligibility, or threshold tuning.

The search universe contains 840 individual step-layer-submodule nodes, with ten sampled steps, 28 layers, and SA, CA, and FFN. For each signed query, attribution is estimated from fifty discovery pairs. Raw EAP-IG ranks target-effect magnitude. Specificity contrasts the target with a frozen same-family pool of five other appearance directions. Uniform random selection samples 36 nodes from the same universe. All three selectors therefore share the candidate space and node budget without an appearance-specific layer or timestep subgrid.

Evaluation uses the same frozen pair IDs, prompts, seeds, source references, and base references for all selectors. The evaluation pairs are disjoint from the discovery pairs. Each polarity has forty pairs, giving eighty signed-direction trials per selector and arm from forty shared scene identities. The random set is one predeclared uniform draw per polarity, without redraw based on intervention outcomes. All 480 matched intervention trials complete with no intention-to-treat failures. Table 8 reports the CIELAB results.

The median sign-aligned source-base effect is $\Delta b^* = 27.02$. The paired median specificity-minus-EAP recovery difference is $+0.092$ with 95% interval $[+0.078, +0.124]$ in sufficiency and $+0.078$ with interval $[+0.069, +0.097]$ in necessity. These intervals use 10,000 bootstrap samples clustered by the forty shared scene identities. Uniform random selection has substantially smaller recovered effects despite the identical node count. Since every selector is sign-correct on every trial, recovered magnitude, rather than sign alone, separates the selectors in this task.

Table 26: Secondary RGB readout on the same color-temperature videos.

Selector	Sufficiency median ρ	Necessity median ρ
Uniform random	0.203	0.298
EAP-IG	0.541	0.709
Specificity	0.756	0.835

Every secondary-readout cell is also 100% sign-correct. The paired specificity-minus-EAP difference is $+0.122$ with 95% interval $[+0.110, +0.141]$ in sufficiency and $+0.068$ with interval $[+0.055, +0.118]$ in necessity. Thus, the ordering does not depend only on the primary CIELAB readout. The evidence extends the directly tested scope from motion to one non-motion appearance attribute. The other appearance directions define the contrast pool and are not presented here as additional validated intervention tasks.

I ADDITIONAL QUALITATIVE EXAMPLES

Figures 4 to 8 show additional camera and object directions using the row conventions of Figure 3. Positive/negative directions appear left/right, and sufficiency/necessity above/below. These selected illustrations show the recipient and patched videos. A separate donor row is not shown. The scene sentences below complete the prompts with the action templates in Table 11: sufficiency recipients use only the scene, while necessity recipients append the corresponding template. TL/TR/BL/BR denote panel positions.

Scene sentences for Figure 3 (*cam pan* $+$ / *obj depth* $-$).

- *TL (cam pan +, suff.):* "The video captures a vibrant coastal cityscape, with a large Ferris wheel as the focal point."
- *TR (obj depth −, suff.):* "A brown bear is seen running through a river, splashing water around it."
- *BL (cam pan +, nec.):* "The video captures a serene scene of a field of green plants, with a single white flower with a yellow center standing out in the foreground."
- *BR (obj depth −, nec.):* "a white dog is the main subject, standing on a wooden boardwalk that cuts through a forest."

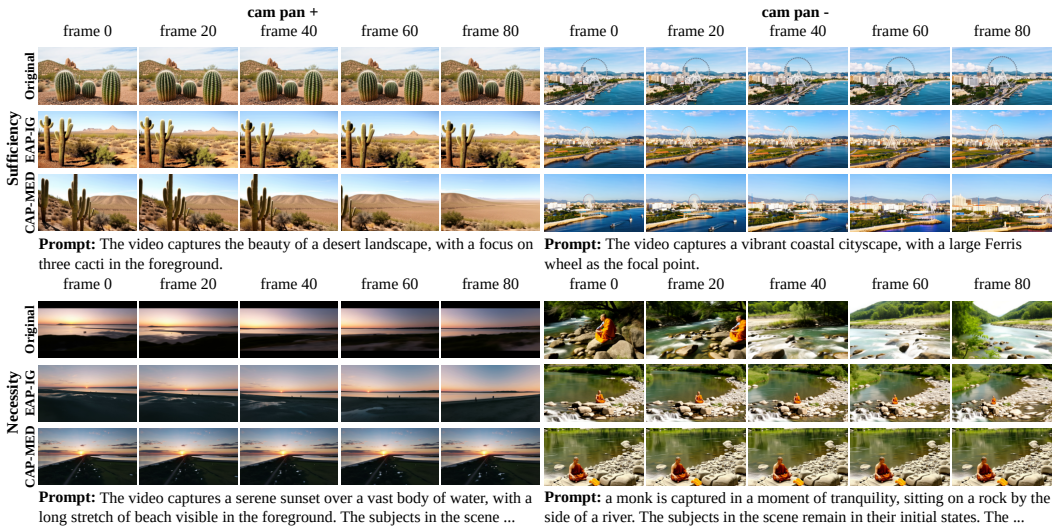


Figure 4: Qualitative examples for camera pan. Left column: *cam pan* + (rightward sweep). Right column: *cam pan* − (leftward sweep).

TL: "The video captures the beauty of a desert landscape, with a focus on three cacti in the foreground." *TR:* "The video captures a vibrant coastal cityscape, with a large Ferris wheel as the focal point." *BL:* "The video captures a serene sunset over a vast body of water, with a long stretch of beach visible in the foreground." *BR:* "a monk is captured in a moment of tranquility, sitting on a rock by the side of a river."

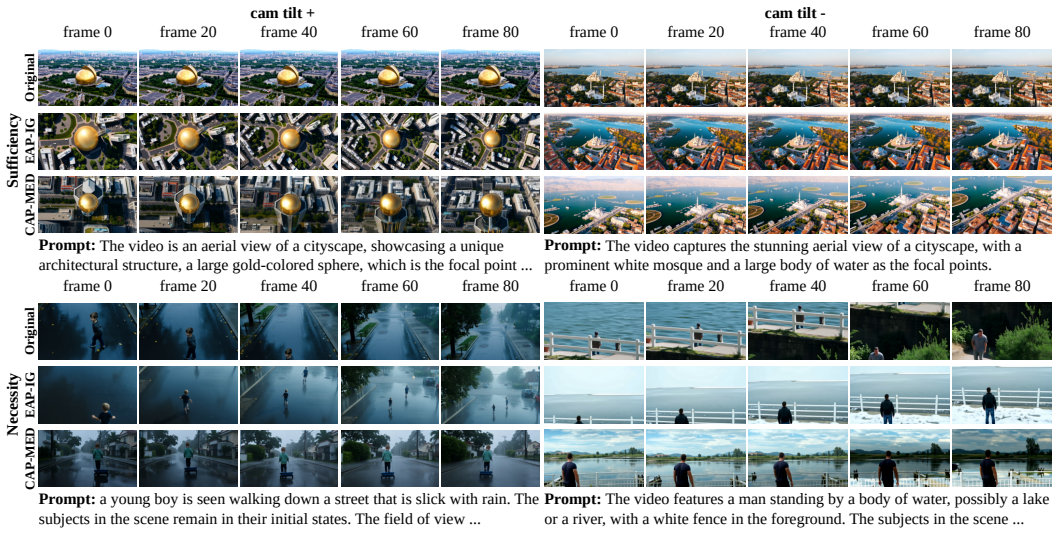


Figure 5: Qualitative examples for camera tilt. Left column: *cam tilt* + (upward tilt). Right column: *cam tilt* − (downward tilt).

TL: "The video is an aerial view of a cityscape, showcasing a unique architectural structure, a large gold-colored sphere, which is the focal point of the scene." *TR:* "The video captures the stunning aerial view of a cityscape, with a prominent white mosque and a large body of water as the focal points." *BL:* "a young boy is seen walking down a street that is slick with rain." *BR:* "The video features a man standing by a body of water, possibly a lake or a river, with a white fence in the foreground."

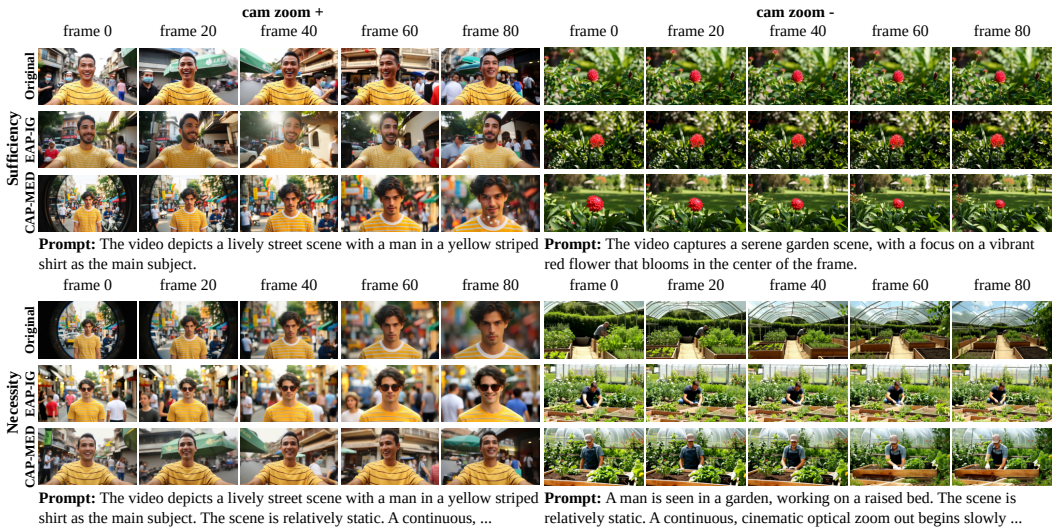


Figure 6: Qualitative examples for camera zoom. Left column: *cam zoom* + (zoom in). Right column: *cam zoom* − (zoom out).
TL: "The video depicts a lively street scene with a man in a yellow striped shirt as the main subject." *TR:* "The video captures a serene garden scene, with a focus on a vibrant red flower that blooms in the center of the frame." *BL:* "The video depicts a lively street scene with a man in a yellow striped shirt as the main subject." *BR:* "A man is seen in a garden, working on a raised bed."

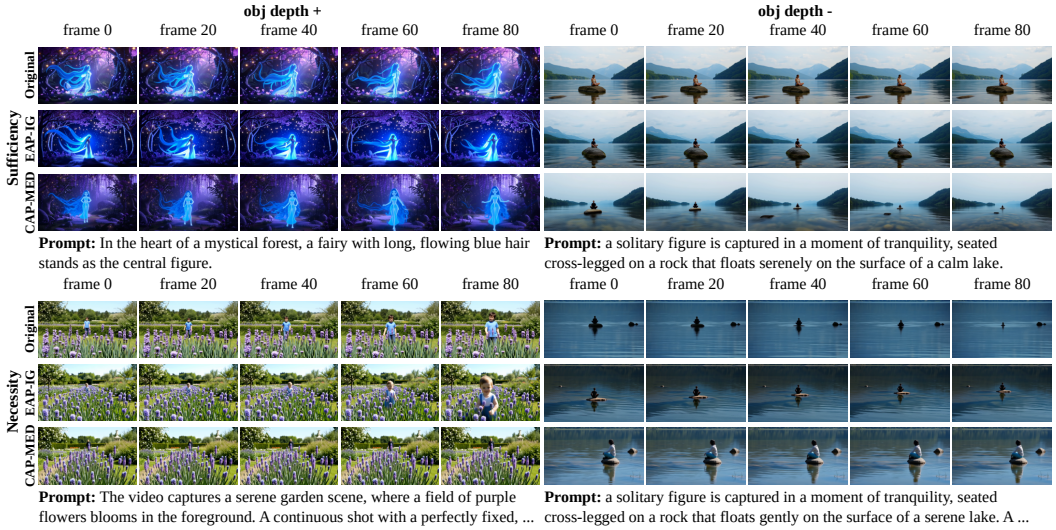


Figure 7: Qualitative examples for object depth motion. Left column: *obj depth* + (subject moving toward the camera along z). Right column: *obj depth* − (subject moving away from the camera along z).
TL: "In the heart of a mystical forest, a fairy with long, flowing blue hair stands as the central figure." *TR:* "a solitary figure is captured in a moment of tranquility, seated cross-legged on a rock that floats serenely on the surface of a calm lake." *BL:* "The video captures a serene garden scene, where a field of purple flowers blooms in the foreground." *BR:* "a solitary figure is captured in a moment of tranquility, seated cross-legged on a rock that floats gently on the surface of a serene lake."

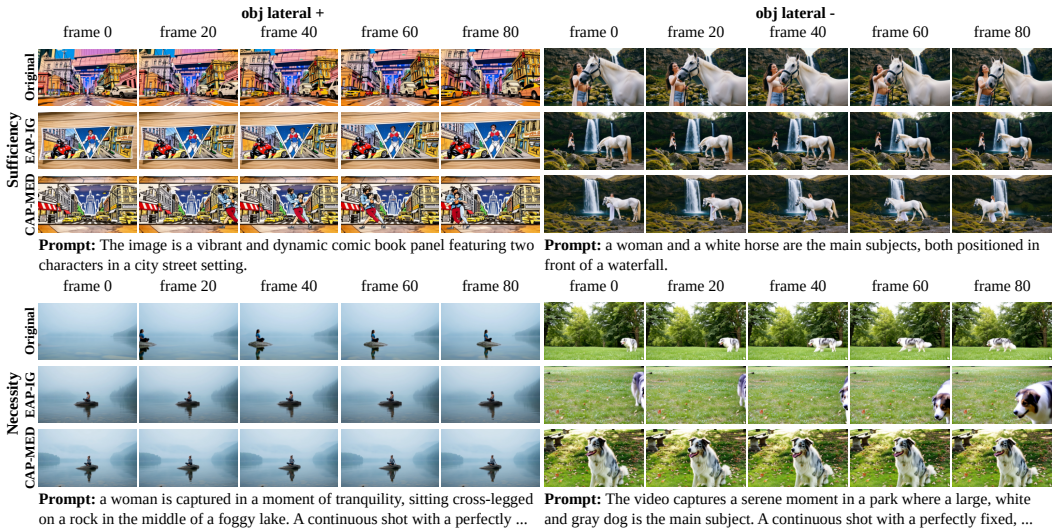


Figure 8: Qualitative examples for object lateral motion. Left column: *obj lateral* + (left-to-right traversal). Right column: *obj lateral* − (right-to-left traversal).

TL: "The image is a vibrant and dynamic comic book panel featuring two characters in a city street setting." *TR:* "a woman and a white horse are the main subjects, both positioned in front of a waterfall." *BL:* "a woman is captured in a moment of tranquility, sitting cross-legged on a rock in the middle of a foggy lake." *BR:* "The video captures a serene moment in a park where a large, white and gray dog is the main subject."