TEMPORAL EVIDENCE HEADS: FROM TEMPORAL SENSITIVITY TO ANSWER RELEVANCE IN VIDEO-MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video multimodal large language models (Video-MLLMs) can misjudge event order even when they recognize the video content. We find that heads that respond strongly to temporal reordering are not always the heads that most affect support for the correct answer. We propose Temporal Evidence Heads (TEH), a method that requires no parameter training. TEH first identifies temporally sensitive heads and then ranks them through bidirectional representation swapping between original and temporally reordered videos. The resulting answer-support measurements guide head selection. At inference, TEH adjusts the selected heads along a sample-specific direction computed from the current video's original–counterfactual representation difference. Experiments on TempCompass and VidHalluc show that TEH improves temporal reasoning across multiple Video-MLLMs. Head-selection ablations show that answer-support changes help select more effective heads than temporal sensitivity alone, and a fixed-head construction ablation shows that TEH retains its benefit under an alternative temporal permutation.

1 INTRODUCTION

Video multimodal large language models (Video-MLLMs) can answer questions, describe events, and reason over videos, but they still fail on action stages, event order, and cross-frame relations (Liu et al., 2024; Li et al., 2025; Kong et al., 2025). A model may recognize the people, objects, and actions in a video while reversing the order of two events. Figure 1 illustrates this temporal hallucination: the video presents a clear sequence, yet the answer exposes a gap between content recognition and temporal-relation reasoning.

Existing inference-time methods improve temporal reasoning along two main directions. One line compares an original video with a variant whose temporal cues are weakened or reordered, and uses the output difference to emphasize temporal information, as in MotionCD and TCD (Zhang et al., 2024a; Kong et al., 2025). Another line locates internal modules or attention heads related to temporal information and directly modifies their representations, as in TA-AE and CircuitProbe (Cai et al., 2025; Zhang et al., 2025b). The second line raises a basic question: does a strong response to a temporal perturbation mean that a head matters for the answer? Reordering frames also changes their positional organization, so a large internal response may reflect input rearrangement without being closely related to the current prediction. A second issue concerns intervention direction: a fixed direction estimated from other videos may not suit the current input. These observations motivate two questions—which temporally sensitive heads change answer support, and how should they be adjusted for the current video?

Our analysis reveals that some attention heads respond consistently to changes in event order, yet response magnitude need not induce the same ordering as changes in answer support. This distinction motivates using temporal sensitivity to narrow the search space and answer-support changes to guide head selection. Based on this observation, we introduce Temporal Evidence Heads (TEH), illustrated in Figure 2. TEH first reorders the same sampled frames to find candidate heads that respond to temporal order. It then swaps the corresponding representations between the original and counterfactual videos and measures whether support for the correct answer recovers or drops. These discovery-time measurements determine an ordered candidate family from which validation selects the intervention budget and strength.

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: Temporal hallucinations in Video-MLLMs. Although the videos contain clear event order, the models may reverse key actions in their answers, revealing a gap between content recognition and temporal-relation reasoning.

At inference, the selected head locations are shared across videos, but the adjustment direction is recomputed for each input. TEH derives this direction from the representation difference between the current video and its temporal counterfactual, allowing the intervention to follow the content and temporal structure of each sample. The method requires no parameter training: base-model parameters remain frozen, while head selection uses labeled discovery and validation data.

Our main contributions are:

- We identify a gap between temporal sensitivity and answer relevance: an attention head can respond consistently to changes in event order without producing a correspondingly large change in support for the correct answer.
- We propose TEH, which first finds temporally sensitive heads and then ranks them by how representation swapping changes support for the correct answer.
- We introduce a sample-specific intervention direction computed from each video's own representation difference. Experiments across multiple Video-MLLMs and temporal reasoning settings, together with controls over head selection, direction, and fixed-head counterfactual construction, demonstrate the effectiveness of this design.

2 Related Work

2.1 Multimodal Hallucination Evaluation and Inference-Time Mitigation

Multimodal models may generate objects, attributes, or relations that lack visual evidence. POPE evaluates this problem through object existence (Li et al., 2023), while HallusionBench examines vision–language consistency, language hallucination, and visual illusion (Guan et al., 2024). Training-free mitigation methods typically strengthen visual evidence or suppress language priors during inference. VCD contrasts predictions from an image and a perturbed counterpart (Leng et al., 2024); OPERA reallocates attention during decoding (Huang et al., 2024); and SECOND contrasts visual information at different scales (Park et al., 2025). These methods mainly target static images. Video errors also depend on how events are organized across frames, so stronger single-frame evidence alone does not resolve mistakes in temporal relations.

2.2 Temporal Understanding and Hallucination in Video-MLLMs

Video benchmarks expose temporal limitations from complementary perspectives. MVBench and Video-MME cover broad video-understanding and temporal tasks (Li et al., 2024; Fu et al., 2025), while TempCompass focuses on event order and fine-grained temporal perception (Liu et al., 2024). MHBench, VidHalluc, and VidHal characterize video hallucinations involving motion, event relations, and several temporal attributes (Kong et al., 2025; Li et al., 2025; Choong et al., 2024). V-STaR further shows that recognizing video content does not imply reliable spatiotemporal reasoning (Cheng et al., 2026). Mitigation methods either construct a contrastive branch by modifying temporal structure, as in MotionCD and TCD (Kong et al., 2025; Zhang et al., 2024a), or directly strengthen and redistribute

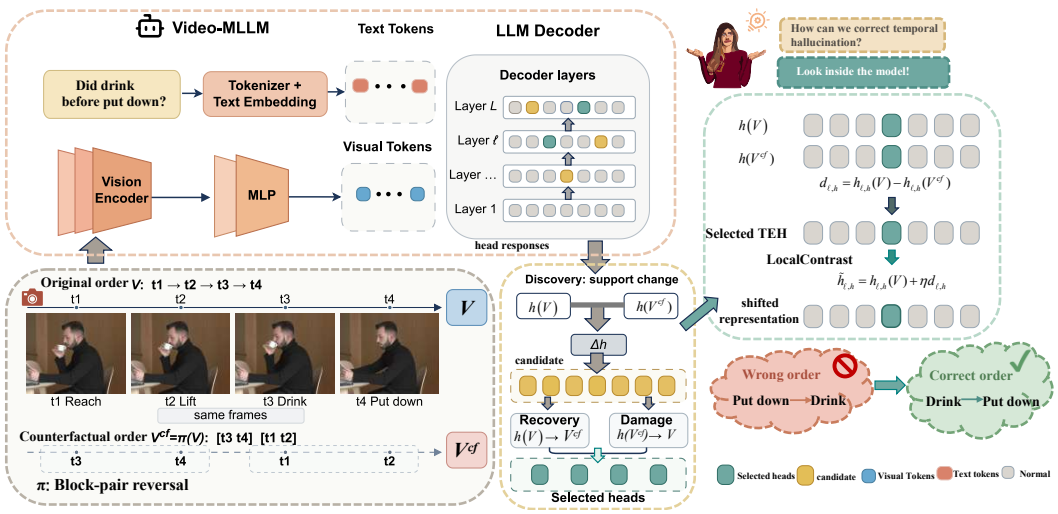

Figure 2: Overview of TEH. TEH first identifies heads that respond to temporal reordering and then tests which of them affect support for the correct answer. Validation selects the head budget and intervention strength. At inference, the selected head locations are fixed, while LocalContrast derives a sample-specific adjustment from the current video and its temporal counterfactual.

visual-temporal evidence, as in DINO-HEAL, SEASON, VERHallu, STEAR, and Unstitching the Chimera (Li et al., 2025; Wu et al., 2026; Zhang et al., 2026; Fan et al., 2026; Yang et al., 2026). These studies indicate that temporal information can improve video reasoning. TEH asks a further question: does intervening on temporally responsive internal locations change support for the correct answer?

2.3 Locating and Intervening on Internal Representations

Local editing and representation replacement provide direct tools for studying model internals. ROME locates and edits parameters associated with factual knowledge, showing that a small set of internal locations can alter a specific prediction (Meng et al., 2022). Activation patching instead replaces representations between inputs and tracks the resulting prediction change. A related line modifies internal representations directly at inference. Activation engineering and representation engineering show that directions in intermediate representations can steer model outputs without retraining (Turner et al., 2023; Zou et al., 2023). Prior work also cautions that counterfactual construction and evaluation metrics shape how such interventions should be interpreted (Zhang & Nanda, 2024; Makelov et al., 2024). In multimodal image models, CausalLens selects multi-head locations by visual sensitivity and intervenes on them to reduce hallucination (Ji et al., 2026).

Head-level intervention in video models is most closely related to our setting. TA-AE selects temporally related locations from head activations under original and interval-sampled videos, then injects a cross-sample mean offset (Cai et al., 2025). CircuitProbe finds and amplifies attention heads related to temporal concepts (Zhang et al., 2025b). These methods show that temporally related internal locations can be found and used. TEH further tests whether swapping a located head's representation changes answer support, and computes the adjustment direction separately for each video instead of relying on a fixed cross-sample offset.

3 Method

3.1 What Does TEH Address?

Given a video V, a question q, and a finite candidate set $\mathcal{Y}$, a Video-MLLM assigns each candidate $y \in \mathcal{Y}$ a score $s_y(V, q)$ and predicts the highest-scoring answer. We normalize these scores within

the candidate set to express model support on a common scale:

$$\widehat{y} = \arg\max_{y \in \mathcal{Y}} s_y(V, q),$$
$$p(y \mid V, q) = \frac{\exp s_y(V, q)}{\sum_{y' \in \mathcal{Y}} \exp s_{y'}(V, q)}. \tag{1}$$

We use $p(y \mid V, q)$ below as the model's support for candidate y.

TEH applies head-level temporal intervention in three stages. First, it compares internal responses to an original video and a temporal counterfactual to find candidate heads that respond to event order. Second, it swaps the candidate representations in both directions and measures how support for the correct answer changes. Third, it adjusts a small set of selected heads along the original–counterfactual representation difference of the current video.

The method requires no parameter training. The candidate ranking is estimated from labeled discovery data. Validation data select the number of heads K and the intervention strength η. After selection, the ranking, K, and η remain fixed for all subsequent evaluation, and the base Video-MLLM parameters stay frozen. Candidate scoring follows the backbone-specific protocol described in Supplement A.

3.2 Which Heads Respond to Temporal Order?

We first locate heads whose representations change with event order. The original and counterfactual videos contain the same sampled frames and differ only in temporal organization. Given a sampled sequence $V = (v_1, \ldots, v_T)$, we group adjacent frames into local temporal blocks $(v_1, v_2), (v_3, v_4), \ldots, (v_{T-1}, v_T)$. We preserve the order inside each block and reverse the order of the blocks. Denoting this block-pair reversal by π, the temporal counterfactual is

$$V^{\mathrm{cf}} = \pi(V) = \big((v_{T-1}, v_T), \ldots, (v_1, v_2)\big). \tag{2}$$

V and V^{cf} use the same sampled frames, question, spatial preprocessing, and input template. The sampled frame content is preserved, while its temporal organization is altered.

For each candidate answer y, we first measure the output response to the frame reordering:

$$\Delta s_y^{(i)} = s_y(V_i, q_i) - s_y(V_i^{\mathrm{cf}}, q_i), \qquad y \in \mathcal{Y}. \tag{3}$$

$\Delta s_y^{(i)}$ is the answer-level response induced by the temporal counterfactual. We then inspect per-head representations at the candidate-scoring position to locate where this response appears internally. Let $h_{\ell,h}(V, q) \in \mathbb{R}^{d_h}$ denote the representation of head h in layer ℓ, where d_h is its dimensionality. For discovery sample i, the paired representation difference is

$$\Delta h_{\ell,h}^{(i)} = h_{\ell,h}(V_i, q_i) - h_{\ell,h}(V_i^{\mathrm{cf}}, q_i). \tag{4}$$

Both branches use the same question and answer candidates, so $\Delta h_{\ell,h}^{(i)}$ captures the head's representational response to the change in frame order.

We measure this response with the root mean square of the per-sample difference and average it over N discovery samples:

$$a_{\ell,h} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{d_h} \left\| \Delta h_{\ell,h}^{(i)} \right\|_2^2}. \tag{5}$$

A larger $a_{\ell,h}$ indicates a stronger representational response to event-order changes. This response provides a candidate-discovery signal for the subsequent answer-support analysis. We rank all heads by $a_{\ell,h}$ and retain the top M as the candidate pool:

$$\mathcal{C}_a = \mathrm{TopM}(\{a_{\ell,h}\}_{\ell,h}). \tag{6}$$

To assess repeatability, we divide the discovery data into two disjoint subsets by video identity and compute a sensitivity ranking on each half. Spearman rank correlation and top-M overlap summarize their agreement. The resulting candidates then enter representation swapping between the original and counterfactual branches.

3.3 Does Swapping Representations Change Answer Support?

We next ask a more direct question for each candidate head: does replacing its representation change support for the correct answer? We use activation patching (Zhang & Nanda, 2024) to swap candidate representations between the original and counterfactual videos. Let y_i be the correct candidate for discovery sample i. Recovery inserts the original-video representation into the counterfactual branch and measures whether correct-answer support rises. Damage inserts the counterfactual representation into the original branch and measures whether that support falls. Both signs are defined relative to the correct answer for the original video; positive values indicate recovery in the counterfactual branch or damage in the original branch:

$$\begin{aligned} e_{\ell,h}^{\mathrm{rec},i} &= p_i^{\mathrm{cf}\leftarrow V}(y_i;\ell,h) - p_i^{\mathrm{cf}}(y_i), \\ e_{\ell,h}^{\mathrm{dmg},i} &= p_i^{V}(y_i) - p_i^{V\leftarrow \mathrm{cf}}(y_i;\ell,h). \end{aligned} \tag{7}$$

Here, $p_i^{\mathrm{cf}\leftarrow V}$ is the candidate distribution after inserting the original representation of the target head into the counterfactual branch; $p_i^{V\leftarrow \mathrm{cf}}$ denotes the reverse replacement. A positive Recovery score means that the original representation raises correct-answer support in the counterfactual branch. A positive Damage score means that inserting the counterfactual representation lowers correct-answer support in the original branch. Together, the two directions summarize how swapping this head changes support for the correct answer.

We average Recovery and Damage over a support-evaluation subset $\mathcal{P} \subseteq D_{\mathrm{discovery}}$ to obtain the bidirectional support change:

$$c_{\ell,h} = \frac{1}{2|\mathcal{P}|} \sum_{i\in\mathcal{P}} \left(e_{\ell,h}^{\mathrm{rec},i} + e_{\ell,h}^{\mathrm{dmg},i} \right). \tag{8}$$

$c_{\ell,h}$ estimates the average change in original-answer support under the specified swaps. Temporal sensitivity $a_{\ell,h}$ identifies candidates, whereas $c_{\ell,h}$ refines their ordering. Qwen3 uses 32 support samples per dataset; SmolVLM2 and VideoLLaMA3 use all 100/200 discovery samples on TempCompass/VidHalluc, respectively.

We use $\mathcal{C}_{\mathrm{eligible}} \subseteq \mathcal{C}_a$ to mark heads that satisfy a support-change stability criterion. On the sensitivity candidate pool $\mathcal{C}_a$, we construct a deterministic discovery-time ranking using bidirectional support change, with stability information incorporated according to each setting. Depending on the setting, stability information serves as a ranking priority, a selection criterion, or an auxiliary marker. The criterion and its role are fixed during discovery for each setting; Supplement A gives the corresponding implementations.

The frozen discovery-defined family provides the prefixes

$$\mathcal{H}_{\mathrm{TEH}}(K) = \mathrm{TopK}(\mathcal{R}_{\mathrm{TEH}}). \tag{9}$$

For the setting that uses stability as a selection criterion, eligible candidates define the admissible prefix family. An answer-aligned margin change supplies a secondary ordering signal, with deterministic layer/head indices resolving remaining ties. Validation selects the prefix size K and intervention strength η from this discovery-defined family. The selected heads and strength remain fixed during evaluation.

3.4 How Should Selected Heads Be Adjusted per Input?

Once the heads are selected, TEH computes an adjustment direction for each input. Videos differ in content, action stage, and question semantics, so we share the head locations and intervention strength across samples but recompute the direction from the current video. For each selected head, the difference between the original and counterfactual representations defines

$$\begin{aligned} d_{\ell,h} &= h_{\ell,h}(V,q) - h_{\ell,h}(V^{\mathrm{cf}},q), \\ \widetilde{h}_{\ell,h} &= h_{\ell,h}(V,q) + \eta\, d_{\ell,h}, \\ &\qquad (\ell,h) \in \mathcal{H}_{\mathrm{TEH}}(K). \end{aligned} \tag{10}$$

In implementation, $h_{\ell,h}$ is the per-head representation before the attention output projection, and V^{cf} is constructed by block-pair reversal. LocalContrast uses the raw, unnormalized original–counterfactual difference with backbone-specific execution. For SmolVLM2 and VideoLLaMA3,

the original and counterfactual representations are cached in separate forwards and $d_{\ell,h}$ is held fixed during the intervention forward. For Qwen3, the counterfactual representation is cached first, while the original-branch representation in Eq. (10) is obtained online at each selected location during the intervention forward; downstream selected layers can therefore incorporate effects from preceding interventions. In all cases, the head set and η remain fixed after validation, while the adjustment is specific to the current input.

We call this update LocalContrast. For heads in $\mathcal{H}_{\mathrm{TEH}}(K)$, LocalContrast moves the original representation by η times the current sample's direction $d_{\ell,h}$; all other heads remain unchanged. The head set and η are fixed after validation, while $d_{\ell,h}$ is recomputed for every input. The same head locations can therefore be used across videos, with an adjustment direction determined from each video's own representation difference.

During inference, each backbone derives the adjustment from the current video and its counterfactual using the execution above. The base-model parameters remain frozen, and every adjustment is derived from the two branches of the current sample. Supplement A records the implementation and frozen configurations. Supplement C tests an alternative temporal permutation with the head sets and intervention settings fixed, and Supplement K reports inference cost.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. We evaluate TEH on TempCompass Event-Order Yes/No (Liu et al., 2024) and VidHalluc-TSH (Li et al., 2025). The 382 TempCompass samples are split into discovery100, validation50, and heldout232; VidHalluc-TSH is split into discovery200, validation100, and test300. Discovery data locate and rank candidate heads, and validation data select the intervention size and strength. The resulting configuration is fixed before held-out or test evaluation. To assess cross-dataset transfer, we also apply the configuration selected on TempCompass directly to VidHalluc full600 without target-dataset tuning.

Models and baselines. Our main experiments use Qwen3-VL-2B (Qwen Team, 2025), SmolVLM2-2.2B (Marafioti et al., 2025), and VideoLLaMA3-2B (Zhang et al., 2025a). We compare against Vanilla, VCD (Leng et al., 2024), MotionCD (Kong et al., 2025), TA-AE (Cai et al., 2025), DINO-HEAL (Li et al., 2025), and SEASON (Wu et al., 2026). We adapt TA-AE and DINO-HEAL from public implementations, using the latter's additive fusion, and reimplement SEASON following its published formulation. All methods share held-out/test splits and forced-choice scoring; validation selects method-specific hyperparameters where applicable. Supplement G gives the implementation and selection details.

Evaluation and implementation. The main metric is strict forced-choice accuracy, with exact ties counted as incorrect. We also report the accuracy change relative to Vanilla and paired flips as W→C/C→W. TEH requires no parameter training; all base-model parameters remain frozen. The TEH candidate ranking is estimated from discovery data; the final number of heads and intervention strength are selected on validation data and remain fixed during evaluation.

Controlled comparisons and statistics. The fixed-parameter head-selection comparison in Table 3 uses the same head budget and intervention strength within each setting, along with matched sampling, prompt, scoring, and LocalContrast implementation. Supplement B provides the configuration details and paired statistics. TempCompass confidence intervals use paired video-cluster bootstrap over 70 held-out videos (232 questions), with 10,000 replicates and seed 20260908. VidHalluc test300 has one video per item. Supplement B reports complete paired intervals, exact item-level McNemar tests, and flips.

4.2 Results

Table 1 summarizes the main results on TempCompass heldout232 and VidHalluc-TSH test300. TEH outperforms the corresponding Vanilla baseline for all three Video-MLLMs in both settings. The gains range from 11.21 to 18.53 percentage points on TempCompass and from 12.33 to 13.33 points

Table 1: Main comparison on TempCompass heldout232 and VidHalluc-TSH test300. Values are strict forced-choice accuracy (%). Method-specific hyperparameters are selected using validation data where applicable; implementation details are provided in Supplement G.

	TempCompass heldout232			VidHalluc-TSH test300		
Method	Qwen3	SmolVLM2	VideoLLaMA3	Qwen3	SmolVLM2	VideoLLaMA3
Vanilla	59.48	55.17	53.45	82.00	82.00	82.67
VCD	56.47	56.90	53.88	81.33	83.67	82.33
MotionCD	69.83	**68.10**	**77.59**	84.67	94.00	93.67
TA-AE	60.78	55.60	53.45	80.67	83.33	82.33
DINO-HEAL	50.43	53.02	45.69	52.00	60.67	53.00
SEASON	59.91	56.03	53.45	76.00	83.00	79.33
TEH	**78.02**	66.38	69.40	**95.33**	**94.33**	**95.67**

Table 2: Sensitivity stability and sensitivity–support agreement across three models and two datasets. ρ_{sens} compares split-half sensitivity over all heads; J@32 measures their Top-32 Jaccard overlap. $\rho(a, c)$ and Top-4 overlap compare sensitivity with pure bidirectional support change within the frozen sensitivity Top-32, before eligibility policies. Discovery sizes and support subsets are given in Supplement A.

Dataset	Model	ρ_{sens}	J@32	$\rho(a, c)$	Top-4 overlap
TempCompass	Qwen3	0.9982	0.8286	-0.2031	1/4
TempCompass	SmolVLM2	0.9925	0.8286	-0.1015	1/4
TempCompass	VideoLLaMA3	0.9851	0.7297	0.4032	3/4
VidHalluc	Qwen3	0.9998	0.9394	0.2335	3/4
VidHalluc	SmolVLM2	0.9996	0.8824	0.1067	1/4
VidHalluc	VideoLLaMA3	0.9994	0.9394	0.4260	3/4

on VidHalluc-TSH. These results indicate that modifying a small set of selected attention heads can improve judgments of event order and action stage.

Comparison with inference-time baselines. Table 1 compares TEH with five inference-time methods. MotionCD provides a strong baseline, achieving the highest TempCompass accuracy on SmolVLM2 and VideoLLaMA3, while TEH performs best on Qwen3. On VidHalluc-TSH, TEH achieves the highest accuracy across all three backbones. The additional TA-AE, DINO-HEAL, and SEASON comparisons retain this overall pattern: TA-AE and SEASON remain close to their corresponding Vanilla baselines in most settings, whereas DINO-HEAL yields lower accuracy in this evaluation. These results show that the choice of where and how temporal evidence enters the current prediction matters alongside the temporal contrast itself.

4.3 Why Temporal Evidence Heads?

To understand how temporal sensitivity relates to answer support, we compare their rankings across models and datasets. Sensitivity rankings are highly repeatable across the discovery halves. Table 2 covers all three main models on both datasets. Within the sensitivity Top-32 pool, $\rho(a, c)$ ranges from -0.203 to 0.426, and the sensitivity and pure-c Top-4 sets overlap by one to three heads. Figure 3(a) illustrates the SmolVLM2–VidHalluc setting. These results show that answer-support measurements provide a distinct ordering of temporally sensitive heads. Support-estimate stability is discussed separately in Supplement D.

Recovery and Damage measure the effects of both replacement directions. In Figure 3(b), the original representation raises original-answer support in the counterfactual branch, whereas the counterfactual representation lowers it in the original branch. The displayed Damage curve is $-e^{\text{dmg}}$, so its negative values are consistent with the positive damage magnitude in Eq. (7).

Controlled head selection. To examine whether answer-support-based ranking provides additional value beyond temporal sensitivity, we compare sensitivity-only selection with TEH at matched K and η within each setting (Table 3). Across all six model–dataset settings, TEH improves accuracy in five and matches sensitivity-only selection in the remaining one. The gains are largest on TempCompass: +27.59 points for Qwen3, +9.05 for SmolVLM2, and +4.31 for VideoLLaMA3, with

video-cluster 95% intervals of [19.05, 35.78], [3.70, 14.29], and [1.38, 7.58] points, respectively. On VidHalluc, TEH improves Qwen3 and SmolVLM2 accuracy and matches VideoLLaMA3 sensitivity-only selection.

Table 3: Controlled sensitivity-only versus TEH selection. Accuracy is in %; Δ is TEH minus sensitivity-only in percentage points. Each row uses matched K, η at evaluation; full protocols and paired statistics appear in Supplement B.

Dataset	Model	Sensitivity-only	TEH	Δ
TempCompass	Qwen3	50.43	78.02	+27.59
TempCompass	SmolVLM2	57.33	66.38	+9.05
TempCompass	VideoLLaMA3	65.09	69.40	+4.31
VidHalluc	Qwen3	93.67	95.33	+1.67
VidHalluc	SmolVLM2	92.33	94.33	+2.00
VidHalluc	VideoLLaMA3	95.67	95.67	0.00

Independent Tuning Ablation. The preceding fixed-parameter ablation isolates the effect of changing the selected head set at a common operating point, whereas this experiment asks whether the comparison persists when each selection strategy is independently tuned. Sensitivity-only selection ranks heads directly by the temporal-sensitivity score $a_{\ell,h}$, without the support-based re-ranking used by TEH. We therefore allow sensitivity-only selection and TEH to independently choose the head budget K and intervention strength η from the same validation search grid. Each method retains its own discovery-defined head ranking, and configurations are selected by strict validation accuracy, with ties resolved by smaller K and then smaller η. The selected configurations are frozen before final evaluation.

Table 4: Independently tuned head-selection comparison on TempCompass. Sensitivity-only and TEH independently select (K, η) from the same validation search grid. The higher final accuracy within each row is shown in bold.

Model	Sens. (K, η)	Sens. Acc. (%)	TEH (K, η)	TEH Acc. (%)
Qwen3	$(16, 2)$	61.64	$(8, 4)$	**78.45**
SmolVLM2	$(8, 4)$	63.79	$(2, 4)$	63.79
VideoLLaMA3	$(16, 2)$	55.60	$(4, 4)$	**69.40**

As shown in Table 4, independently tuned sensitivity-only selection reaches 61.64% accuracy on Qwen3, whereas TEH reaches 78.45%. A similar pattern is observed for VideoLLaMA3, where TEH achieves 69.40% compared with 55.60% for sensitivity-only selection. SmolVLM2 reaches the same accuracy under both selection strategies. The paired video-cluster 95% confidence intervals are [9.21, 24.14] and [9.24, 18.52] percentage points for Qwen3 and VideoLLaMA3, respectively, excluding zero in both cases. Under independent tuning, TEH remains higher on Qwen3 and VideoLLaMA3, while SmolVLM2 reaches the same accuracy under both selection strategies.

To further examine support-based head ranking, we compare TEH with a Damage-only variant on Qwen3–VidHalluc. Damage-only ranking reaches 92.00%, while TEH reaches 95.33%, showing that how support changes are used to rank heads can affect downstream accuracy. The SmolVLM2 scope control reaches 93.00% when all 96 heads in the selected layers are patched, compared with 94.33% for TEH Top-4. Its opposite-direction control reaches 8.00%, motivating the direction analysis below.

4.4 Sample-Specific Intervention

To examine the role of sample-specific intervention directions, we keep the selected head locations and intervention size fixed and vary only the source of the adjustment direction. For the fixed-mean baseline, we average representation differences across discovery samples and select the intervention strength on validation data. LocalContrast derives its direction from the current video and its temporal counterfactual. On Qwen3-VL-2B, SmolVLM2-2.2B, and VideoLLaMA3-2B, the fixed-mean direction achieves 81.33%, 82.33%, and 83.00%, close to the corresponding Vanilla results of 82.00%, 82.00%, and 82.67%. LocalContrast reaches 95.33%, 94.33%, and 95.67%, respectively. These results support sharing the selected head locations across samples while determining the intervention direction from the current video.

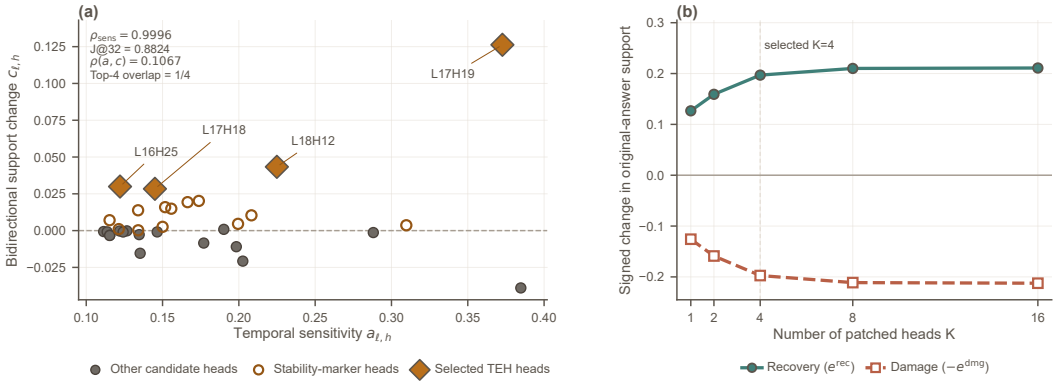

Figure 3: SmolVLM2 on VidHalluc discovery200. (a) The 32 sensitivity candidates are compared with $c_{\ell,h}$. Split-half eligibility serves as an auxiliary stability marker in this setting; diamonds identify the final selected heads. Top-4 overlap compares sensitivity with pure-c ranking. (b) Signed patching effects: Recovery is e^{rec} and the negative Damage curve is $-e^{\text{dmg}}$. Both refer to the answer of the original video.

Cached-direction versus online execution. In Qwen3, earlier interventions can affect later representations during the online forward. We compare this execution with cached-direction LocalContrast on Qwen3–TempCompass heldout232, keeping the selected heads fixed (Supplement J). At matched $\eta = 4$, fixed mean reaches 57.33%, cached direction 74.14%, and online execution 77.59%. Cached direction gains +16.81 points over fixed mean (paired video-cluster 95% CI [10.50, 22.77]); online execution adds +3.45 points (95% CI [0, 7.08]). Input-specific direction thus provides a substantial benefit in this setting, with a smaller observed increment from online execution.

Further direction and location controls appear in Supplement J; counterfactual, transfer, cost, and large-model results appear in Supplements C and K.

5 Conclusion

We presented TEH, a method that connects temporal sensitivity with answer-support changes to select attention heads for temporal reasoning. TEH ranks candidate heads using bidirectional representation swapping and applies a sample-specific original–counterfactual direction at inference. The method requires no parameter training and combines discovery-based head selection with input-dependent intervention.

Across TempCompass and VidHalluc, TEH improves over Vanilla on three Video-MLLMs and compares favorably with recent inference-time methods. Matched head-selection comparisons favor answer-support ranking in five of six settings, with one tie. On Qwen3–TempCompass, cached input-specific directions substantially outperform a matched shared mean, while online execution adds a smaller increment. Fixed-head counterfactual results retain gains under another temporal construction. Together, these findings distinguish temporal responsiveness, answer relevance, and input-specific intervention.

Our evaluation focuses on forced-choice temporal reasoning, and query-level semantic validity of temporal counterfactuals remains to be systematically evaluated. Cross-dataset transfer varies across models, and the counterfactual branch adds inference cost. Future work may extend TEH to open-ended video understanding, evaluate counterfactual semantics at the query level, and develop more efficient intervention directions.

AI Use Statement

Generative AI tools were used to assist with research discussion, literature retrieval, and manuscript editing for clarity. All research decisions, experimental procedures, results, citations, and claims were

reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Jianfeng Cai, Jiale Hong, Zongmeng Zhang, Wengang Zhou, Nianji Zhan, and Houqiang Li. Mitigating hallucination in videoLLMs via temporal-aware activation engineering. In *Advances in Neural Information Processing Systems*, 2025.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-STaR: Benchmarking video-LLMs on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9155–9164, 2026.

Wey Yeh Choong, Yangyang Guo, and Mohan S. Kankanhalli. VidHal: Benchmarking temporal hallucinations in vision LLMs. *CoRR*, abs/2411.16771, 2024. URL https://arxiv.org/abs/2411.16771.

Linfeng Fan, Yuan Tian, Ziwei Li, and Zhiwu Lu. STEAR: Layer-aware spatiotemporal evidence intervention for hallucination mitigation in video large language models. *arXiv preprint arXiv:2604.03045*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. MHBench: Demystifying motion hallucination in video LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4401–4409, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024. doi: 10.18653/v1/2024.findings-acl.517.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. SECOND: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. In *Proceedings of the International Conference on Machine Learning*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. SEASON: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11096–11105, 2026.

Songyuan Yang, Guijian Tang, Kun Hu, Haotian Wang, Shixuan Liu, Wenjing Yang, Long Lan, and Huibin Tan. Unstitching the chimera: Frame-level risk and train-free mitigation for video hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4188–4198, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

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.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, and Jingjing Chen. EventHallusion: Diagnosing event hallucinations in video LLMs. *arXiv preprint arXiv:2409.16597*, 2024a.

Yiming Zhang, Zhuokai Zhao, Chengzhang Yu, Kun Wang, Zhendong Chu, Qiankun Li, Zihan Chen, Yang Liu, Zenghui Ding, Yining Sun, and Qingsong Wen. CircuitProbe: Tracing visual temporal evidence flow in video language models. *arXiv preprint arXiv:2507.19420*, 2025b.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, April 2024b.

Zefan Zhang, Kehua Zhu, Shijie Jiang, Hongyuan Lu, Shengkai Sun, and Tian Bai. VERHallu: Evaluating and mitigating event relation hallucinations in video large language models. *arXiv preprint arXiv:2601.10010*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

SUPPLEMENTARY MATERIAL

A DISCOVERY, ELIGIBILITY, AND FROZEN CONFIGURATIONS

All six main settings score every attention head by discovery-time temporal sensitivity and retain $M = 32$ candidates. Sensitivity split-half statistics measure the repeatability of candidate discovery, while support-change stability is computed from bidirectional patching effects on the support subset $\mathcal{P}$. TEH requires no parameter training. Discovery uses original-answer labels to estimate support changes; validation performance selects the prefix length K and intervention strength η from the discovery-defined family. Evaluation uses the frozen configuration.

Fixed stability criteria and their roles. Let c_A and c_B denote support changes on the two discovery support halves. Criterion Q is

$$c > 0 \quad \wedge \quad c_A > 0 \quad \wedge \quad c_B > 0 \quad \wedge \quad (\mathrm{rank}_A(c) \leq 16 \ \vee \ \mathrm{rank}_B(c) \leq 16).$$

Criterion P is $c_A > 0 \wedge c_B > 0$. The marker set $\mathcal{C}_{\mathrm{eligible}}$ records which candidate heads satisfy the relevant criterion. Table S1 specifies its operational role. Each setting uses a fixed stability criterion and role in the discovery-defined ranking.

Table S1: Discovery protocols. N_a is the sensitivity sample count and $N_c = |\mathcal{P}|$ the support sample count. All candidate pools have 32 heads. Auxiliary markers describe stability alongside the support-based ranking.

Dataset	Model	N_a	N_c	Criterion	Role in the frozen family
TempCompass	Qwen3	100	32	Q	Eligible-first ranking priority
TempCompass	SmolVLM2	100	100	P	Recorded marker
TempCompass	VideoLLaMA3	100	100	P	Recorded marker
VidHalluc	Qwen3	200	32	P	Discovery-defined filter
VidHalluc	SmolVLM2	200	200	P	Recorded marker
VidHalluc	VideoLLaMA3	200	200	P	Recorded marker

For Qwen3–TempCompass, all 32 heads remain in the ranking, with eligible heads first, followed by descending c, descending answer-aligned margin effect, and ascending layer/head indices. The final 16-head prefix contains five eligible and eleven non-eligible heads. For Qwen3–VidHalluc, the full-pool ranking is ordered by c, margin, and layer/head; a discovery-defined eligibility filter then determines the admissible prefix family. The filter is applied at the validation entry point using the frozen discovery fields. For the four auxiliary-marker settings, the full candidate pool is ordered by c, margin, and layer/head, with stability recorded alongside the ranking. These setting-specific rules define the prefix families used for validation.

Qwen3–TempCompass assigns support halves by hashed video identity, yielding 17/15 samples from 10/10 videos. Qwen3–VidHalluc uses the first 32 discovery IDs, split into 16/16 video-disjoint items. The other settings use the first and second halves of their ordered discovery lists; these halves are video-disjoint (50/50 samples on TempCompass; 100/100 on VidHalluc). The same setting-specific half construction is retained for the split-half analyses.

Table S2: Validation search families and final TEH configurations. All strength grids are $\eta \in \{1, 2, 4\}$; opposite-direction controls are separate conditions. Final configurations are selected on validation data.

Dataset	Model	Prefix grid K	Final K	Final η
TempCompass	Qwen3	$\{1, 2, 4, 8, 16\}$	16	4
TempCompass	SmolVLM2	$\{1, 2, 4, 8, 16\}$	4	4
TempCompass	VideoLLaMA3	$\{4, 8, 16\}$	8	4
VidHalluc	Qwen3	$\{2, 4, 8, 16\}$	8	2
VidHalluc	SmolVLM2	$\{1, 2, 4, 8, 16\}$	4	4
VidHalluc	VideoLLaMA3	$\{1, 2, 4, 8, 16\}$	4	4

Table S3 identifies the selected head locations using zero-based layer/head indices. The controlled TempCompass comparison uses the same VideoLLaMA3 head set, with simultaneous intervention at these locations.

Scoring and intervention. The intervention targets the per-head representation before the attention output projection at the candidate-scoring position. TempCompass uses the Yes/No candidate protocol. VidHalluc

Table S3: Frozen TEH head locations used for evaluation.

Setting	(Layer, head)
TempCompass / Qwen3	(19,6), (23,6), (25,6), (24,9), (25,7), (18,15), (17,0), (17,1), (26,8), (27,7), (24,6), (26,15), (21,13), (25,11), (23,14), (20,10)
TempCompass / SmolVLM2	(15,14), (15,18), (15,16), (16,17)
TempCompass / VideoLLaMA3	(18,11), (17,0), (19,11), (18,7), (18,0), (17,2), (24,4), (21,3)
VidHalluc / Qwen3	(21,1), (21,13), (23,6), (17,1), (17,0), (18,15), (19,6), (21,12)
VidHalluc / SmolVLM2	(17,19), (18,12), (16,25), (17,18)
VidHalluc / VideoLLaMA3	(21,11), (21,9), (18,3), (22,7)

scores the normal and A/B-swapped action prompts, maps their probabilities back to the original actions, and averages them before the final decision. The two prompt-specific head differences are used by LocalContrast. Exact ties are incorrect. For secondary ranking, Qwen3–VidHalluc uses its answer-aligned probability margin, while the other implementations use their signed candidate-score/logit margin.

Data protocol. Discovery and validation determine the head ranking and intervention configuration, which remain fixed for evaluation. VidHalluc test300 follows the discovery200/validation100/test300 split. The full600 dataset was also used in earlier experiments.

B Controlled Comparisons and Paired Statistics

To examine the value of answer-support-based ranking, Table 3 compares TEH with sensitivity-only selection in all six main model–dataset settings. For the five comparisons other than SmolVLM2–VidHalluc, K and η are inherited from the frozen TEH configuration. For SmolVLM2–VidHalluc, $K = 4$ is fixed and sensitivity-only selects $\eta \in \{1, 2, 4\}$ on validation100 (ties favor smaller η); it selects $\eta = 4$, matching TEH. Its sensitivity control uses RMS attribution over the normal/swapped prompt differences. Within each evaluation pair, the head budget, realized strength, model, frame sampling, preprocessing, prompts, scoring, and LocalContrast are matched. Only the selected head set differs.

Qwen3–VidHalluc Damage-only provides an additional comparison of bidirectional and single-direction support ranking.

Unified paired uncertainty. For each item we compute the binary correctness difference $d_i = B_i - A_i$ between the named conditions. We resample video IDs with replacement and include all associated questions from each selected video. Each bootstrap replicate reports $100 \sum_i d_i / N_{\text{resampled}}$. We use 10,000 replicates, Python's random generator with seed 20260908, and the sorted replicate entries at $\lfloor 0.025(9999) \rfloor$ and $\lfloor 0.975(9999) \rfloor$ for the percentile interval. Video clusters follow first appearance in the evaluation row order. TempCompass has 70 clusters for 232 questions. VidHalluc test300 has 300 unique video IDs, so its video-cluster resampling reduces to item resampling. The exact two-sided McNemar test uses item-level discordances, and its unadjusted p-values are reported alongside the video-cluster intervals. All comparisons use this bootstrap procedure.

Table S4: Complete paired head-selection statistics. A is sensitivity-only except in the final Damage-only row; B is TEH. Δ and CI are percentage points. Flips are A-wrong/B-correct and A-correct/B-wrong. The final row is a separate support-ranking variant.

Setting	K, η	A correct	B correct	Δ	Video-cluster 95% CI	Flips	McNemar p
TempCompass / Qwen3	16, 4	117/232	181/232	+27.59	[19.05, 35.78]	79/15	1.15×10^{-11}
TempCompass / SmolVLM2	4, 4	133/232	154/232	+9.05	[3.70, 14.29]	28/7	5.08×10^{-4}
TempCompass / VideoLLaMA3	8, 4	151/232	161/232	+4.31	[1.38, 7.58]	12/2	0.01294
VidHalluc / Qwen3	8, 2	281/300	286/300	+1.67	[-0.67, 4.00]	9/4	0.26685
VidHalluc / SmolVLM2	4, 4	277/300	283/300	+2.00	[-0.67, 4.67]	11/5	0.21011
VidHalluc / VideoLLaMA3	4, 4	287/300	287/300	+0.00	[0.00, 0.00]	0/0	1.00000
VidHalluc / Qwen3 (Damage-only)	8, 2	276/300	286/300	+3.33	[1.33, 5.67]	10/0	0.00195

All three TempCompass sensitivity-only comparisons favor TEH, with video-cluster confidence intervals excluding zero. The Qwen3 and SmolVLM2 VidHalluc sensitivity-only comparisons show smaller gains whose confidence intervals include zero. VideoLLaMA3 uses different head sets but has identical correctness on all 300 items, yielding a zero-width empirical interval. Qwen3's Damage-only comparison differs in one of eight heads and yields 10 improvements with no regressions relative to that control.

Table S5: Paired TempCompass comparisons using the same video-cluster bootstrap. Differences and flips follow the displayed arrow. P0 refers to the default frozen-head construction. Each comparison is shown once.

Comparison	Δ (pp)	Video-cluster 95% CI	Flips	McNemar p
Qwen3 P0: Sensitivity → TEH	+27.59	[19.05, 35.78]	79/15	1.15×10^{-11}
Qwen3 P0: Vanilla → TEH	+18.53	[11.30, 25.20]	55/12	1.03×10^{-7}
Qwen3 P0: Vanilla → Sensitivity	-9.05	[-14.60, -3.78]	14/35	0.00380
VideoLLaMA3: Sensitivity → TEH	+4.31	[1.38, 7.58]	12/2	0.01294
VideoLLaMA3: Vanilla → TEH	+15.95	[11.11, 20.98]	41/4	9.33×10^{-9}
VideoLLaMA3: Vanilla → Sensitivity	+11.64	[6.76, 16.67]	29/2	4.63×10^{-7}
Qwen3 P1: Sensitivity → TEH	+20.69	[11.98, 29.46]	68/20	2.77×10^{-7}
Qwen3 P1: Vanilla → TEH	+15.09	[7.98, 21.96]	46/11	3.31×10^{-6}
Qwen3: P0 TEH → P1 TEH	-3.45	[-8.78, 1.28]	7/15	0.13380
Qwen3: P0 Sensitivity → P1 Sensitivity	+3.45	[-0.88, 7.87]	25/17	0.27996
Qwen3 P1: Vanilla → Sensitivity	-5.60	[-11.48, 0.00]	16/29	0.07245

C Fixed-Head Counterfactual Construction

To examine sensitivity to the temporal construction, we evaluate Qwen3-VL-2B-Instruct on the same TempCompass heldout232 IDs. Both methods retain their respective P0-frozen head sets, $K = 16$, $\eta = 4$, identical model checkpoint, 16-frame sampling and preprocessing, prompt, candidate scoring, and LocalContrast. Only the temporal permutation changes. Using zero-based frame positions,

$$\pi_{P0} = [14, 15, 12, 13, 10, 11, 8, 9, 6, 7, 4, 5, 2, 3, 0, 1],$$

$$\pi_{P1} = [8, 9, 10, 11, 12, 13, 14, 15, 0, 1, 2, 3, 4, 5, 6, 7].$$

The original video and evaluation label remain unchanged. P1 reuses the sensitivity-only and TEH head sets, K, and η selected with P0. The frozen sensitivity-only locations are (26,9), (25,0), (23,7), (18,15), (23,6), (21,12), (27,5), (19,7), (25,9), (25,6), (25,11), (24,2), (22,7), (23,14), (17,1), and (17,8). The TEH set is the Qwen3–TempCompass row in Table S3.

P0 reaches 117/232 for sensitivity-only and 181/232 for TEH; P1 reaches 125/232 and 173/232, respectively. Vanilla is 138/232 under both conditions. P0 uses the same evaluation results as the Qwen3 head-selection comparison. P1 retains the same items, videos, sampled frame indices, and original-branch baseline scores.

Table S6: Fixed-head counterfactual construction on Qwen3–TempCompass heldout232. The two frozen head sets, $K = 16$, $\eta = 4$, 16-frame sampling, prompt, scoring, and LocalContrast are all fixed; only the P0/P1 temporal permutation changes. Values are correct/232 and accuracy (%). Vanilla is 138/232 (59.48%) for both.

Permutation	Frozen sensitivity-only	Frozen TEH
P0: block-pair reversal	117 (50.43)	181 (78.02)
P1: block-half swap	125 (53.88)	173 (74.57)

Table S5 reports paired video-cluster intervals for the construction comparisons. P1 retains a positive TEH advantage over both Vanilla and frozen sensitivity-only selection, showing that the selected heads remain effective under block-half swap.

D Support-Change Stability

To examine the repeatability of answer-support measurements, we correlate the two half-specific support-change vectors across the frozen sensitivity Top-32 pool. This analysis uses pure c values and complements the sensitivity-stability results in Table 2.

Support-change stability is setting-dependent. Near-zero half agreement for Qwen3 and SmolVLM2 on TempCompass means that low sensitivity–support correlation there can partly reflect uncertain support estimates. On VidHalluc, high c stability coexists with limited sensitivity–support agreement, particularly for Qwen3 and SmolVLM2. These results characterize how sensitivity and support measurements vary across discovery subsets and datasets.

Table S7: Split-half stability of support-change estimates on discovery data. Support sample counts are shown for each setting.

Dataset	Model	Support samples	Split-half ρ_c
TempCompass	Qwen3	32	0.0312
TempCompass	SmolVLM2	100	-0.0290
TempCompass	VideoLLaMA3	100	0.5634
VidHalluc	Qwen3	32	0.9439
VidHalluc	SmolVLM2	200	0.9872
VidHalluc	VideoLLaMA3	200	0.9934

E FIXED TEMPCOMPASS SPLITS

TempCompass uses a fixed video-disjoint partition of its 382 Event-Order Yes/No questions. The calibration set contains 150 questions and is divided by video identity into discovery100 and validation50; the remaining 232 questions form heldout232. The three sets contain 20, 10, and 70 videos, respectively, with zero pairwise overlap in both question IDs and video IDs. Discovery100 contains 50 Yes and 50 No answers, validation50 contains 26 Yes and 24 No answers, and heldout232 contains 116 of each. All three main backbones use this partition.

The exact split IDs are provided with the supplementary source in `supplementary_source/tempcompass/`: `discovery_ids.txt`, `validation_ids.txt`, and `heldout_ids.txt`. The accompanying `split_items.csv` maps each question ID to its split, video ID, question, and answer. These files specify the partition used for discovery, validation, and evaluation.

F PROMPTS AND CANDIDATE SCORING

TempCompass prompts. Qwen3 and SmolVLM2 use the following text, with the dataset question substituted for {`question`}:

```
Answer the question based on the video.
Question: {question}
Answer with yes or no only.
```

VideoLLaMA3 uses:

```
Answer the following question based on the video. Reply with yes
↪  or no only.
Question: {question}
```

VidHalluc prompts. Qwen3 uses:

```
Below are two actions from the video:
A. {action_a}
B. {action_b}

Which action happened first in the video? Answer with A or B only.
```

SmolVLM2 uses:

```
Below are two actions in the video:
Action A. {action_a}
Action B. {action_b}
Which action happens first? Answer with A or B only.
```

VideoLLaMA3 uses:

```
Answer the question using the video. Reply with A or B only.
Below are two actions in the video:
Action A. {action_a}
Action B. {action_b}
Which action happens first?
```

Each text is supplied as the text component of a user message alongside the sampled video, using the backbone's processor chat template and an assistant generation prefix. Qwen3 and SmolVLM2 use `add_generation_prompt=True`; VideoLLaMA3 also uses `add_system_prompt=True`. The sampled frames and their temporal order are supplied explicitly to the processor. The supplementary source includes the backbone-specific prompt and input-construction implementations in `supplementary_source/prompt_implementations/`.

Candidate scores for the three main backbones. Let z_t be the next-token logit for token t at the final prompt position. For Yes/No scoring, the candidate strings are `yes`, `Yes`, `YES` and `no`, `No`, `NO`, each with and without a leading space. For A/B scoring, they are `A`, `a` and `B`, `b`, again with and without a leading space. Each string is tokenized with `add_special_tokens=False`; single-token encodings are retained and duplicate token IDs are removed. For the resulting token set T_y of candidate y, the score and probability are

$$s_y = \max_{t \in T_y} z_t, \qquad p(y) = \frac{\exp(s_y)}{\sum_{y' \in \mathcal{Y}} \exp(s_{y'})}.$$

TempCompass uses $\mathcal{Y} = \{\text{Yes}, \text{No}\}$ and the higher-scoring candidate. Strict accuracy counts ties as incorrect.

VidHalluc evaluates both the normal action assignment and the assignment with the two action descriptions swapped. Each prompt uses a two-candidate softmax over A/B. The probabilities are mapped back to the original actions and averaged:

$$p(A_{\text{original}}) = \tfrac{1}{2}\big[p(A \mid \text{normal}) + 1 - p(A \mid \text{swapped})\big], \qquad p(B_{\text{original}}) = 1 - p(A_{\text{original}}).$$

Strict correctness requires the probability of the correct original action to exceed 0.5. LocalContrast computes a separate original–counterfactual head difference for each prompt assignment before combining the resulting answer probabilities.

G Baseline Implementations, Search Spaces, and Selected Configurations

G.1 VCD and MotionCD

VCD uses the 1,000-step sigmoid diffusion noise schedule on the model-ready visual tensor. We search noise steps $t \in \{200, 500, 700, 999\}$ and contrast strengths $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$, with $\beta = 0.1$. MotionCD reverses the complete frame sequence and searches the official strength $\alpha_{\text{official}} \in \{10, 20, 40\}$, with $\beta = 0.1$. Its contrastive logits are $(1 + \alpha_{\text{official}}/2)z_{\text{orig}} - (\alpha_{\text{official}}/2)z_{\text{reverse}}$; VCD uses $(1+\alpha)z_{\text{orig}} - \alpha z_{\text{noisy}}$. Here β is the candidate-plausibility threshold parameter. For VidHalluc, the same visual-noise realization is used for the normal and swapped prompts.

Each model–dataset setting selects one configuration per method on validation50 for TempCompass or validation100 for VidHalluc. Selection maximizes strict accuracy, then mean correct-answer probability, with condition-name order breaking remaining ties. Table S8 gives the selected configurations used for the main comparisons.

Table S8: Validation-selected VCD and MotionCD configurations. VCD entries are (t, α, β); MotionCD entries are $(\alpha_{\text{official}}, \beta)$.

Dataset	Model	VCD	MotionCD
TempCompass	Qwen3	$(500, 1.0, 0.1)$	$(20, 0.1)$
TempCompass	SmolVLM2	$(700, 0.75, 0.1)$	$(10, 0.1)$
TempCompass	VideoLLaMA3	$(500, 1.0, 0.1)$	$(40, 0.1)$
VidHalluc	Qwen3	$(999, 1.0, 0.1)$	$(40, 0.1)$
VidHalluc	SmolVLM2	$(999, 1.0, 0.1)$	$(20, 0.1)$
VidHalluc	VideoLLaMA3	$(200, 0.5, 0.1)$	$(40, 0.1)$

G.2 TA-AE

We adapt the public TA-AE implementation to each backbone. Its model-specific activation and direction artifacts are constructed from the TA-AE ShareGPT4Video action subset, using an interval-sampled negative with temporal sampling ratio 4. Following the public code, we capture the output of each decoder `self_attn.o_proj`, reshape it into head-indexed slices, rank locations using a two-fold logistic-regression classifier, and add a mean activation direction scaled by its projection standard deviation at the last text position. Each backbone has its own discovery artifact. This is the `official_literal` hook variant used for all six TA-AE results.

For each model–dataset setting, we search head counts $\{4, 8, 16, 32\}$ and strengths $\{2, 4, 8\}$ on TempCompass validation50 or VidHalluc validation100. Selection maximizes strict forced-choice accuracy, with ties resolved by fewer heads and then smaller strength. The selected configuration is frozen before heldout232 or test300 evaluation. Table S9 gives all six selections.

Table S9: Validation-selected TA-AE configurations. All rows search head counts $\{4, 8, 16, 32\}$ and strengths $\{2, 4, 8\}$; action mode and temporal sampling ratio 4 are fixed.

Dataset	Model	Heads	Strength
TempCompass	Qwen3	8	8
TempCompass	SmolVLM2	32	4
TempCompass	VideoLLaMA3	4	2
VidHalluc	Qwen3	8	8
VidHalluc	SmolVLM2	32	8
VidHalluc	VideoLLaMA3	4	2

G.3 DINO-HEAL

We adapt the public DINO-HEAL implementation using final-layer DINOv2 CLS-to-patch attention averaged over attention heads. Its per-frame saliency is normalized, bilinearly aligned to the backbone's visual feature grid, and passed through a sigmoid. We follow the public implementation's additive fusion: globally standardized visual features receive weight 0.3 and the aligned saliency receives weight 0.7. This differs from the multiplicative reweighting described in the paper. The weights are fixed; validation evaluates this single configuration before it is frozen for each heldout/test split.

The fusion point follows each backbone's visual pathway: Qwen3 uses the post-merger visual output, SmolVLM2 uses the connector output before multimodal insertion, and VideoLLaMA3 uses the vision-encoder output before the multimodal projector and native dynamic compression. Saliency is aligned to Qwen3's merged visual-token sequence, SmolVLM2's $16 \times 9 \times 9$ connector grid, or VideoLLaMA3's $16 \times 8 \times 8$ pre-compression grid. The VideoLLaMA3 projector and native compression then proceed unchanged.

G.4 SEASON

We reimplement SEASON following its published formulation. Each input uses eight uniformly sampled frames and three branches: original, spatial negative, and temporal negative. The spatial branch adds Gaussian noise of standard deviation 0.1 to the processed visual input; the temporal branch mixes each patch representation with its across-frame mean at every vision layer. Decoder attention at layers 20–23 yields per-frame diagnostic distributions and Jensen–Shannon weights, which combine the two negative-branch scores with the original score. Candidate scoring uses the same forced-choice answer protocol as the main comparison.

Qwen3's eight logical frames form four temporal tubelet groups in its vision encoder; decoder visual tokens from each tubelet are associated with their two source frames. SmolVLM2 uses the processor's image-token blocks and eager decoder attention to recover frame-wise diagnostics. VideoLLaMA3 uses the native post-compression kept-token map to associate decoder visual tokens with source frames while retaining its built-in compression. These mappings determine the visual tokens included in SEASON's frame diagnostics.

We search the two published (α, β) settings, $(1.0, 0.33)$ and $(0.5, 0.25)$, on TempCompass validation50 for each backbone. Selection maximizes strict accuracy, with ties resolved by smaller α and then smaller β. SmolVLM2 obtains 34/50 under both settings, so the tie rule selects $(0.5, 0.25)$. Qwen3 and VideoLLaMA3 select the same setting with 28/50 and 31/50, respectively. All three frozen configurations are used for TempCompass heldout232 and transferred directly to VidHalluc-TSH test300 without target-dataset reselection.

H Exploratory Large-Model Evaluation

To explore head intervention in a larger backbone, we evaluate LLaVA-NeXT-Video-7B (Zhang et al., 2024b) on VidHalluc full600 using within-dataset head discovery and strength selection. The experiment samples 16 frames uniformly, resizes them to 224×224 with bicubic interpolation, and constructs the counterfactual by reversing eight two-frame blocks. Head representations are taken at the final prompt position before the attention output projection.

Head discovery. Using baseline predictions on full600, we form three groups of 80 items: incorrect answers with the lowest correct-answer probabilities (F), correct answers closest to probability 0.5 (B), and correct answers with the highest correct-answer probabilities (C). For each head, we compute the L2 norm of its

original–counterfactual representation difference under the normal prompt and average it within each group, giving μ_F, μ_B, and μ_C. The attribution-lite ranking sorts heads by $(\mu_F - \mu_C) + (\mu_F - \mu_B)$, then by $\mu_F - \mu_C$, and then by the mean of the three group averages, all in descending order. Its Top-4 heads are $(19, 19)$, $(26, 15)$, $(15, 10)$, and $(19, 7)$, using zero-based layer/head indices.

Strength selection and update. We use 150 items from full600 for the strength sweep: the 100 incorrect answers closest to the decision boundary and the 50 correct answers closest to that boundary. With $K = 4$ fixed, the negative-offset strength is swept over $\alpha \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$. Strength 2 gives the highest accuracy in this sweep. The full600 evaluation includes strengths 1.5 and 2; the reported configuration uses 2. The update is

$$\widetilde{h}_{\ell,h} = h_{\ell,h}(V, q) - \alpha\big[h_{\ell,h}(V, q) - h_{\ell,h}(V^{\mathrm{cf}}, q)\big], \qquad \alpha = 2.$$

Thus, the exploratory experiment uses a negative-offset direction, with head discovery, strength selection, and evaluation drawn from the same full600 dataset.

Prompt and scoring. The experiment uses the Qwen3 VidHalluc text template in Supplement F, wrapped by the LLaVA processor. It chooses the first single-token A/B pair from `(A, B)`, the pair with a leading space, and the pair with a leading newline. A two-candidate softmax is computed for each prompt assignment and combined by the same normal/swapped probability average. The experiment's correctness rule is $p(\mathrm{correct}) \geq 0.5$.

Matched random-head controls and results. The random controls retain $K = 4$, negative-offset strength 2, and the selected layer distribution $[19, 26, 15, 19]$. For each seed in $0, \ldots, 29$, NumPy's `default_rng(seed)` samples two distinct heads from layer 19 excluding heads 19 and 7, one head from layer 26 excluding head 15, and one head from layer 15 excluding head 10. Sampling within layer 19 is without replacement. All configurations are evaluated on full600. The selected-head configuration increases accuracy from 63.33% to 67.33% (+4.00 points; W→C/C→W is 28/4). The 30 random configurations yield a mean gain of -0.56 points and a maximum gain of $+1.00$ point. These results describe the exploratory within-dataset setting.

I Inference Passes and Activation Caching

Table S12 counts model forward calls for each TempCompass question, including activation collection and the final intervention. The implementations use per-input head-activation caches as follows.

Qwen3: two forwards. The first forward processes the block-pair-reversed video and caches its head representations. The second forward processes the original video with pre-output-projection hooks at the selected locations. At each selected head, the hook reads the current representation and the cached counterfactual representation, applies $h_{\mathrm{current}} + \eta(h_{\mathrm{current}} - h_{\mathrm{cf,cache}})$, and continues the forward to obtain the answer logits. Current representations are obtained during this intervention forward, including any effects of interventions in earlier layers.

SmolVLM2 and VideoLLaMA3: three forwards. The first two forwards separately process the original and counterfactual videos and cache their head representations. Their difference is computed at the selected locations and held fixed for a third forward on the original video, where the corresponding hooks add η times the cached difference. The original and counterfactual inputs are prepared once per question and reused for these calls. The two activation-collection forwards and the intervention forward are included in the reported inference cost.

These caches contain activations for the current video–question pair and are recomputed for each input. The selected head locations and intervention strength are reused across inputs. The forward counts in Table S12 refer to the single Yes/No prompt per TempCompass question; VidHalluc additionally evaluates the normal and swapped prompt assignments.

J Direction and Location Controls

Figure 4(b) provides additional direction and location controls for SmolVLM2-2.2B. The opposite direction reaches 8.00%, sample mismatch 53.33%, the 30-seed mean of matched random heads 82.43%, fixed mean 82.33%, and the current sample's LocalContrast 94.33%. Together, these controls show that both the selected heads and the intervention direction matter.

Qwen3–TempCompass direction and execution protocol. Table S10 uses the frozen 16 Qwen3 TempCompass head locations, 16 uniformly sampled frames, the same block-pair-reversed counterfactual, and the same Yes/No scoring on heldout232. The fixed-mean direction averages original–counterfactual head differences over discovery100. Its validation-tuned strength is $\eta = 1$, selected from $\{1, 2, 4\}$ on validation50; the matched-strength condition uses $\eta = 4$. Cached-direction LocalContrast obtains the current video's original and counterfactual representations before the intervention forward and applies their fixed difference at $\eta = 4$. Online LocalContrast caches the counterfactual representation and obtains the original representation during the intervention forward, also at $\eta = 4$. Online LocalContrast reaches 77.59% in this matched ablation.

Table S10: Direction-source and execution ablation on Qwen3–TempCompass heldout232. Selected head locations are fixed. Validation-tuned selects the shared-direction strength independently on validation data; matched strength uses the LocalContrast strength $\eta = 4$.

Condition	Accuracy (%)
Vanilla	59.48
Fixed mean, validation-tuned	59.91
Fixed mean, matched strength	57.33
Cached-direction LocalContrast	74.14
Online LocalContrast	77.59

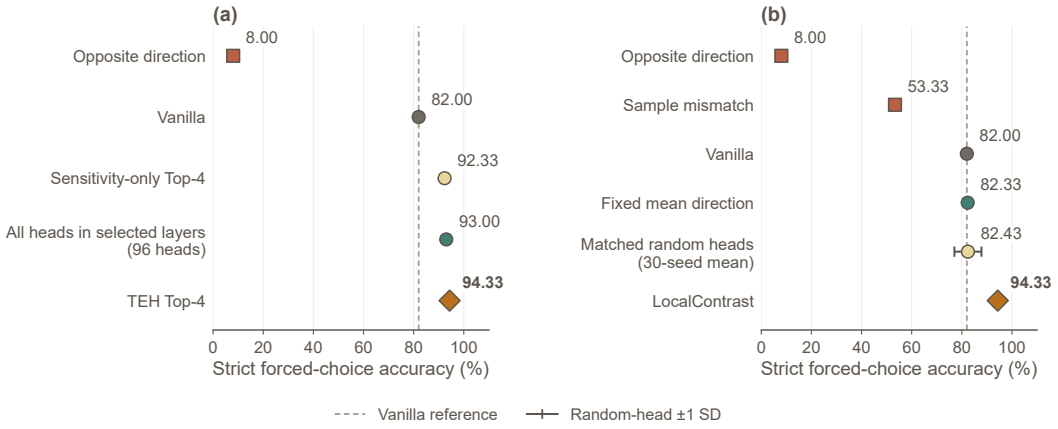

Figure 4: Results are shown for SmolVLM2-2.2B on VidHalluc-TSH test300. (a) Comparison of different head-selection scopes and the opposite direction. (b) Direction controls retain the selected head locations and budget, while matched random heads change the locations under the matching constraint. Fixed mean, sample mismatch, opposite direction, and LocalContrast compare different direction choices.

The paired video-cluster bootstrap uses 70 held-out videos, 10,000 replicates, and seed 20260908. Relative to matched fixed mean, cached-direction LocalContrast improves by 16.81 points (95% CI [10.50, 22.77]; W→C/C→W: 45/6). Online execution improves over cached-direction LocalContrast by a further 3.45 points (95% CI [0, 7.08]; flips: 12/4).

K Generalization and Efficiency

Cross-dataset transfer. We apply the TEH configuration selected on TempCompass directly to VidHalluc full600, without target-dataset tuning. As shown in Table S11, Qwen3-VL-2B and VideoLLaMA3-2B improve by 9.33 and 7.83 percentage points, while SmolVLM2-2.2B decreases by 0.67 points. The frozen configuration transfers well to Qwen3-VL-2B and VideoLLaMA3-2B, but not to SmolVLM2-2.2B. Transfer therefore depends on the backbone.

Table S11: Direct transfer of the TEH configuration selected on TempCompass to VidHalluc full600. Values are strict accuracy (%), and Flips denotes W→C / C→W.

Model	Vanilla	Frozen	Gain	Flips
Qwen3	82.50	**91.83**	+9.33	93 / 37
SmolVLM2	**82.00**	81.33	-0.67	19 / 23
VideoLLaMA3	84.33	**92.17**	+7.83	49 / 2

An exploratory evaluation of LLaVA-NeXT-Video-7B (Zhang et al., 2024b) with within-dataset head and strength selection is provided in Supplement H.

Inference cost. TEH modifies only a small set of attention heads, but it still requires the internal representations of a counterfactual video. The added cost therefore comes mainly from extra model forward passes. On

TempCompass full382, we measure end-to-end latency, peak memory, and forward count relative to Vanilla for the same model.

Table S12: Inference cost on TempCompass full382. Latency and VRAM are normalized to Vanilla for the same model. Head budgets refer to the TempCompass configuration; offline discovery and validation costs are excluded.

Model	Selected	Ratio	Latency	VRAM	Forwards
Qwen3-VL-2B	16/448	3.57%	1.426×	1.000×	2
SmolVLM2-2.2B	4/768	0.52%	2.364×	1.007×	3
VideoLLaMA3	8/336	2.38%	2.623×	1.001×	3

Table S12 shows that TEH modifies only 0.52%–3.57% of the attention heads and leaves peak memory close to Vanilla, while end-to-end latency rises to 1.426×–2.623×. The counterfactual branch accounts for most of this added latency; reducing its cost is the main route to more efficient TEH inference.