# CIRCUITPROBE: TRACING VISUAL TEMPORAL EVIDENCE FLOW IN VIDEO LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Autoregressive large vision–language models (LVLMs) process video by projecting video features into the LLM's embedding space as continuous visual token embeddings. Where temporal evidence is represented along this pathway, and how it causally shapes decoding, remains unclear. We present **CIRCUITPROBE**, a circuit-level analysis framework that dissects the end-to-end video-language pathway in two stages. *Visual Auditing* localizes object semantics within the projected video-token sequence and establishes their causal necessity through targeted ablations and controlled substitutions. *Semantic Tracing* applies logit-lens probing to track the layer-wise emergence of object and temporal concepts, and uses temporal frame interventions to measure sensitivity to temporal structure. The analysis locates a consolidation interval in which visual content becomes language-aligned, together with a set of temporally specialized attention heads. Amplifying these heads in a single window fixed inside the consolidation interval (Layers 25–30), without any parameter update, improves all seven evaluated benchmarks, with significant held-out gains of $+4.0\%$ absolute on EgoSchema, $+2.6\%$ on VideoMME, and $+2.4\%$ on NExTQA, and smaller gains on the development benchmark TempCompass. These results provide interventional evidence that the identified layer interval and temporal-routing heads are functionally relevant to LVLM performance on temporal video benchmarks.

## 1 INTRODUCTION

Large vision–language models (LVLMs) integrate visual inputs (images or videos) with natural language to produce free-form textual outputs (Liu et al., 2023b;a; Maaz et al., 2023; Zhang et al., 2024b). Most LVLMs couple a pre-trained visual encoder with a pre-trained large language model (LLM) through a trainable adapter, or projector (Li et al., 2022; Moon et al., 2023; Chen et al., 2023), which converts visual features into continuous token embeddings (soft prompts) compatible with the LLM's input space, enabling cross-modal alignment and downstream reasoning (Merullo et al., 2023; Liu et al., 2024a; Yu et al., 2025; Pan et al., 2025).

Understanding *how* video evidence influences language decoding matters for two reasons: it clarifies what information is actually transferred through the visual–language interface, which directly affects alignment quality (Radford et al., 2021; Zhang et al., 2024a), and it offers mechanism-level guidance for LVLM design beyond benchmark-driven improvements (Pang et al., 2024; Woo et al., 2024; Park et al., 2024; Wang et al., 2025a).

Recent interpretability work has advanced our understanding of image-based LVLM representations and visual–text integration (Neo et al., 2024), but how these tools extend to *video* inputs remains less clear. Video understanding requires aggregating evidence across frames and integrating cues distributed over time, yet video-derived soft prompts do not correspond directly to discrete language tokens, making it difficult to attribute model outputs to specific temporal evidence or to isolate where temporal reasoning succeeds or fails (Merullo et al., 2023; Yin et al., 2023).

We introduce **CIRCUITPROBE**, a circuit-based tracing framework with two complementary analytic probes. *Visual auditing* localizes task-critical semantics within the projected video-token sequence and tests their causal necessity through targeted token ablations and controlled substitutions. *Semantic tracing* tracks the layer-wise emergence of language-aligned object and action concepts with logit-lens–style (Nostalgebraist, 2020) probing and shows how temporal structure affects semantic



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.

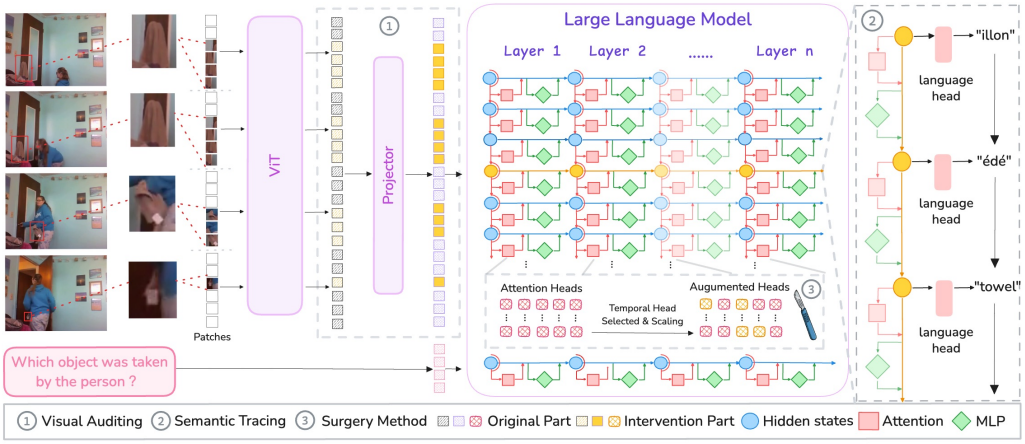


Figure 1: CIRCUITPROBE consists of two circuits: ① *Visual Auditing*, which localizes task-critical object semantics within the projected video-token sequence, and ② *Semantic Tracing*, which tracks the layer-wise emergence of language-aligned concepts. Surgical Recovery amplifies *temporal attention heads* within the identified layers, yielding held-out gains on long-video benchmarks.

consolidation. Guided by these observations, we design targeted interventions that test whether the identified structures are causally responsible for temporal reasoning. Our contributions are:

- We show that video semantics remain strongly localized to object-aligned token positions after projection: ablating object tokens causes substantially larger performance degradation than removing matched control tokens.
- We find that object and action concepts become reliably language-aligned primarily in mid-to-late LLM layers, defining a consolidation interval where video-conditioned semantics become decodable, in line with recent layer-wise analyses of VideoLLMs (Kim et al., 2026); at the inspected token positions, decoded concepts track frame-wise states with a lag, consistent with an evidence-aggregation account that motivates targeted temporal interventions.
- Using attention-based temporal metrics, we isolate temporally specialized attention heads and test their causal role through *bidirectional* interventions. Amplifying them in one window fixed inside the consolidation interval yields Holm-significant held-out gains on EgoSchema, VideoMME, and NExTQA without parameter updates (sufficiency); suppressing or ablating them causes dose-dependent degradation on temporal video benchmarks that is minimal on image benchmarks and under matched random-head controls (necessity), so the identified heads causally contribute to performance on the evaluated temporal video tasks.

## 2 RELATED WORK

**Interpretability of LLM.** LLM interpretability has largely focused on text-only models (Singh et al., 2024; Wang et al., 2025b), using circuit analysis, which reduces models to sparse circuits (Hanna et al., 2023; Yao et al., 2024), causal tracing, which tracks information flow to attribute module contributions and edit factual knowledge (Meng et al., 2022; Yu & Ananiadou, 2024; Sharma et al., 2024), and unembedding-space projection, which probes internal states in token space (Geva et al., 2021; Dar et al., 2023).

**Interpretability of LVLMs.** Existing work examines how frozen transformers process visual tokens (Neo et al., 2024; Kaduri et al., 2024), how misalignment causes hallucinations (Jiang et al., 2025), and how modality spaces are bridged (Pang et al., 2024; Woo et al., 2024; Park et al., 2024); other studies expose shortcomings on static images (Tong et al., 2024; Verma et al., 2024), extract visual concept subgraphs (Rajaram et al., 2024), or reverse-engineer ViTs (Vilas et al., 2023). Recent work addresses video data, identifying temporal reasoning bottlenecks (Li et al., 2024c) and developing new interpretability tools (Joseph et al., 2025). Most closely related, Kim et al. (2026) apply attention knockout and logit lens to VideoLLMs and report that cross-frame interaction starts in early-to-middle layers, video–language integration proceeds in middle layers, and answers form in middle-to-late layers; our layer-wise readouts at object-aligned video-token positions agree with this

picture, and we build on it by treating the consolidation interval as an *a priori* target for head-level causal interventions rather than the endpoint of the analysis. Han et al. (2026) trace temporal signals across encoder, projector, and LLM and attribute arrow-of-time failures mainly to the projector, complementing our focus on the LLM backbone; token-pruning methods (Huang et al., 2025) likewise show that retaining object-relevant video tokens preserves performance. Head-level analyses attribute image-to-text information flow to specific attention heads (Kim et al., 2025) and probe head specialization to enhance or suppress targeted concepts (Basile et al., 2026), while Li et al. (2026) trace temporal-logic inconsistency to poorly discriminative cross-modal heads and sharpen them by training. The term *temporal heads* was coined by Park et al. (2025) for text-only heads recalling time-specific facts; we reuse the name for a different object, heads that route attention across video frames. Closest to ours, Cai et al. (2025) reduce hallucination in VideoLLMs by injecting offset vectors into probe-selected heads at inference; we instead select heads by attention statistics, train no probe, and scale them in both directions as a causal test within the interval located by our tracing analysis.

# 3 CircuitProbe

## 3.1 Preliminary Setup

We evaluate three 7B LVLM backbones covering both image- and video-trained variants: the `LLaVA-NeXT` family (Li et al., 2024b), the `LLaVA-OneVision` family (Li et al., 2024a), and `Qwen2.5-VL` (Bai et al., 2025). The suffixes `-I` and `-V` denote *image*- and *video*-trained variants, respectively.

**Model input representation.** Given a video sampled into $T$ frames and a textual query $Q$, the input embedding sequence is $\mathbf{X}_0 = [\mathbf{X}_{\text{sys}}; \mathbf{X}_V; \mathbf{X}_Q] \in \mathbb{R}^{N \times d}$, where $N$ is the total number of tokens, $d$ is the hidden dimension of the LLM, $\mathbf{X}_{\text{sys}}$ denotes system tokens, $\mathbf{X}_Q$ the query tokens, and $\mathbf{X}_V \in \mathbb{R}^{N_V \times d}$ the video-conditioned tokens injected into the LLM, obtained by encoding each frame into $P$ visual patches with the vision encoder $f_C$ followed by the projector. The resulting $N_V = T \times P$ video tokens are indexed by $\mathcal{I}_V = \bigcup_{t=1}^{T} \mathcal{I}_{V,t}$, where $\mathcal{I}_{V,t}$ denotes the tokens originating from frame $t$. All non-video token indices (system and query tokens) form $\mathcal{I}_Q$.

**Circuit-based analysis.** Visual auditing (Circuit ①, §3.2) and semantic tracing (Circuit ②, §3.3) both manipulate subsets of video tokens and trace their transformation across layers, testing concrete predictions about information localization and semantic emergence. Unless otherwise specified, the analyses focus on the prefill stage of autoregressive decoding and on video-token positions, where visual information is first injected into the LLM.

**Dataset curation and probing setup.** To probe temporal understanding, we curate a subset of the STAR benchmark (Wu et al., 2021) covering two tasks: *Interaction* (action-object binding) and *Sequence* (chronological dependencies). To isolate visual reasoning from language priors, we apply three filters: (1) *Indispensability*: we retain only samples whose answers require integrating video information, discarding those solvable from a single frame or common sense. (2) *Visual Clarity*: we select key frames clearly depicting the target actions or objects. (3) *Counterfactual Validity*: we keep samples whose unperturbed prediction is correct but flips when the annotated object is masked; this subset is used for controlled circuit tracing, and Appendix A.3 reports the same ablation trend on the unfiltered Correct and Full subsets. We employ two complementary probing formats: *open-ended* questions ("Which object...") with prefix-forcing to trace specific semantic concepts, and *close-ended* ("Yes/No") questions that register prediction flips under intervention. By default, Circuit ① reports results under both formats, while Circuit ② uses the open-ended format for concept-level tracing. Example frames, further dataset and probing details, and stability across prompt formats are given in Appendix A.2, Appendix A, and Appendix E.1.

## 3.2 Circuit ①: Visual Auditing

As illustrated in Fig. 1, Circuit ① uses controlled manipulations of projected video tokens to identify where task-critical visual semantics reside after projection into the LLM input space, and in what form they are used for downstream language decoding.

### 3.2.1 Where does task-critical visual information reside?

We begin by asking whether object-level visual semantics are diffusely encoded across the projected video-token sequence $X_V$. Inspired by Neo et al. (2024), we follow the notation of §3.1 and let $\mathcal{I}_V$ denote the index set of all projected video tokens in $X_V$. We de-

Table 1: Accuracy drop (↓) from visual token ablation, relative to the *No Ablation* baseline. $_{bf}$ denotes the buffer radius of spatially adjacent tokens included with object tokens. Ablating object tokens causes far larger drops than random tokens, verifying semantic localization.

| | | LLaVA-NeXT-I | | LLaVA-NeXT-V | | LLaVA-OV-I | | LLaVA-OV-V | |
|---|---|---|---|---|---|---|---|---|---|
| **Type** | **Avg Tokens** | Open | Close | Open | Close | Open | Close | Open | Close |
| No Ablation | 0 | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ |
| Register | 13 | $96.6_{\downarrow 3.4}$ | $98.3_{\downarrow 1.7}$ | $98.0_{\downarrow 2.0}$ | $98.2_{\downarrow 1.8}$ | $84.7_{\downarrow 15.3}$ | $98.2_{\downarrow 1.8}$ | $90.3_{\downarrow 9.7}$ | $96.9_{\downarrow 3.1}$ |
| Random | 100 | $97.7_{\downarrow 2.3}$ | $99.1_{\downarrow 0.9}$ | $97.7_{\downarrow 2.3}$ | $98.4_{\downarrow 1.6}$ | $85.2_{\downarrow 14.8}$ | $98.0_{\downarrow 2.0}$ | $90.5_{\downarrow 9.5}$ | $97.1_{\downarrow 2.9}$ |
| | 500 | $94.0_{\downarrow 6.0}$ | $98.6_{\downarrow 1.4}$ | $93.3_{\downarrow 6.7}$ | $98.7_{\downarrow 1.3}$ | $84.9_{\downarrow 15.1}$ | $97.7_{\downarrow 2.3}$ | $89.4_{\downarrow 10.6}$ | $96.5_{\downarrow 3.5}$ |
| | 900 | $89.4_{\downarrow 10.6}$ | $98.0_{\downarrow 2.0}$ | $89.3_{\downarrow 10.7}$ | $98.4_{\downarrow 1.6}$ | $85.3_{\downarrow 14.7}$ | $97.2_{\downarrow 2.8}$ | $90.4_{\downarrow 9.6}$ | $97.1_{\downarrow 2.9}$ |
| Object | 354 | $18.1_{\downarrow 81.9}$ | $45.9_{\downarrow 54.1}$ | $17.3_{\downarrow 82.7}$ | $62.3_{\downarrow 37.7}$ | $40.5_{\downarrow 59.5}$ | $67.9_{\downarrow 32.1}$ | $60.8_{\downarrow 39.2}$ | $78.8_{\downarrow 21.2}$ |
| Object$_{(bf=1)}$ | 476 | $10.0_{\downarrow 90.0}$ | $34.5_{\downarrow 65.5}$ | $12.5_{\downarrow 87.5}$ | $49.3_{\downarrow 50.7}$ | $33.6_{\downarrow 66.4}$ | $53.3_{\downarrow 46.7}$ | $52.6_{\downarrow 47.4}$ | $67.1_{\downarrow 32.9}$ |
| Object$_{(bf=2)}$ | 634 | $9.3_{\downarrow 90.7}$ | $27.8_{\downarrow 72.2}$ | $7.4_{\downarrow 92.6}$ | $44.9_{\downarrow 55.1}$ | $30.9_{\downarrow 69.1}$ | $45.9_{\downarrow 54.1}$ | $51.7_{\downarrow 48.3}$ | $63.0_{\downarrow 37.0}$ |

fine an object-centric index set $\mathcal{I}_O \subset \mathcal{I}_V$ consisting of tokens whose corresponding image patches are fully contained within the ground-truth bounding box of the target object. The bounding-box source and robustness to box precision are detailed in Appendix C.6. The remaining video tokens $\mathcal{I}_V \setminus \mathcal{I}_O$ serve as a pool from which control subsets are constructed.

**Intervention.** To causally test which subsets of video tokens carry task-critical semantics, we remove information from selected token subsets by *replacement rather than deletion*. Specifically, for any subset $\mathcal{S} \subset \mathcal{I}_V$, we replace the corresponding token embeddings in $X_V$ with an *uninformative visual embedding* $x_{\text{null}}$, defined as the mean visual embedding computed over all visual tokens extracted from 10,000 images randomly sampled from the ImageNet (Deng et al., 2009) validation set. This preserves embedding scale and distribution while selectively removing semantic content, avoiding out-of-distribution artifacts; all non-video tokens ($X_{\text{sys}}$, $X_Q$) and remaining video tokens are left unchanged. We consider four ways of selecting the ablated subset $\mathcal{S}$: (1) *Register tokens*: following Darcet et al. (2024), within each key frame we select tokens whose $L_2$ norms are more than two standard deviations above the mean token norm, and ablate these tokens as a control subset; (2) *Random tokens*: as a baseline, we uniformly sample a subset $\mathcal{I}_R \subset \mathcal{I}_V$ of fixed size $|\mathcal{I}_R| \in \{100, 500, 900\}$, spanning from fewer to substantially more tokens than $\mathcal{I}_O$, which controls for the effect of removing video tokens without targeting any specific semantic content; (3) *Object tokens*: the object-centric set $\mathcal{I}_O$ defined above; (4) *Object tokens with buffer*: to account for spatial imprecision and contextual spillover, we expand $\mathcal{I}_O$ to include spatially adjacent tokens, evaluating two buffer sizes: 1-Buffer, which includes immediately neighboring tokens, and 2-Buffer, which includes all tokens within a two-position radius of any object token.

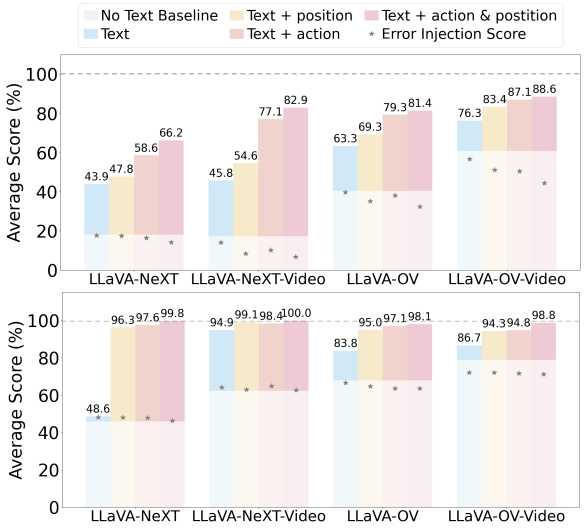


Figure 2: ***Text injection experiment***. Accuracy when object-aligned visual tokens are replaced by embedded text: the label alone (*Text*), or the label with the object's position and/or action appended; *No Text Baseline* is object ablation without injection, and *Error Injection* injects an incorrect label. Left: open-ended; right: close-ended.

**Observation.** As shown in Table 1, ablating object-aligned tokens induces a far larger degradation in performance than ablating any of the control token subsets. This causal asymmetry persists even when the random control set contains substantially more tokens than the object set. For example, on `LLaVA-NeXT-V` (Li et al., 2024b), replacing the 354 object tokens in $\mathcal{I}_O$ leads to an 82.7% drop in open-ended accuracy, rising to 92.6% when a 2-token spatial buffer is included (634 tokens), whereas replacing 900 randomly selected tokens in $\mathcal{I}_R$ reduces performance by only 10.7%. Register tokens produce drops comparable to random tokens, buffered object sets deepen the drop, and the asymmetry

holds under both evaluation formats, across the `LLaVA-NeXT` and `LLaVA-OneVision` (Li et al., 2024a) variants, and on `LLaVA-NeXT-34B` (Appendix C.5).

#### 3.2.2 What representation does the LLM use for object semantics?

Having established *where* task-critical visual semantics are localized in the input sequence, we now ask in what representational form the LLM most effectively uses them: must object-level semantics be conveyed through native visual token embeddings, or can equivalent information be supplied directly in the language embedding space?

**Intervention.** Instead of masking, we perform a controlled substitution by replacing the visual embeddings at the object-aligned indices $\mathcal{I}_O$ with the embedding of the object's textual label (e.g., replacing the visual tokens of a towel with the embedding of the word "towel"), leaving token positions, ordering, system tokens $X_{\text{sys}}$, and query tokens $X_Q$ unchanged. Beyond the label alone (*Text*), we inject richer descriptions that append the object's action and/or its spatial position to the label (*Text + action*, *Text + position*, *Text + action & position*), and an *error injection* control that injects an incorrect label.

**Observation.** As shown in Fig. 2, replacing object-aligned visual tokens with their textual embeddings only partially recovers the performance loss caused by object ablation, indicating that native multimodal evidence carries complementary cues beyond the discrete label. In the open-ended setting on `LLaVA-NeXT-V`, examples that collapse under object ablation (17.3% accuracy) recover to 45.8% with the label alone, and to 82.9% when the action and position descriptions are injected as well. Injecting an incorrect label instead degrades performance, so the recovery comes from the correct semantic content rather than from the substitution itself. This behavior is consistent across LVLM variants and both evaluation formats.

### 3.3 Circuit ②: Semantic Tracing

Circuit ① showed that task-critical visual information enters the LLM through a compact set of object-aligned video tokens, and that supplying the object label in the language embedding space partially recovers the ablated evidence. We build on these constraints and ask the process-level question: *How do object and temporal semantics carried by video tokens evolve across LLM layers before influencing decoding?* The answer locates where temporal reasoning can be causally intervened upon: if visual semantics become usable for decoding only after a particular processing stage, then head-level interventions at that stage should have the largest effect on temporal understanding, a prediction that §4 tests by targeted surgery.

#### 3.3.1 Logit-lens probing: measuring concept evolving across layers

To trace when video-conditioned semantic concepts become expressed in the language space, we employ the logit-lens method (Nostalgebraist, 2020), projecting the intermediate hidden states $h_i^{(\ell)}$ of visual tokens into the vocabulary space using the frozen output head (formal details in Appendix C.1). We quantify this emergence on the object-centric visual tokens $\mathcal{I}_O$ (identified in Circuit ①) using two metrics relative to the target concept $s_c$:

1. *Correspondence Rate ($C_R^{(\ell)}$).* A hard metric: the fraction of tokens whose top-ranked vocabulary projection matches the target concept:

$$C_R^{(\ell)} \triangleq \frac{1}{|\mathcal{I}_O|} \sum_{i \in \mathcal{I}_O} \mathbb{I}\left( \arg\max_{w \in \mathcal{V}} P(w \mid h_i^{(\ell)}) \in \mathcal{T}(s_c) \right). \tag{1}$$

2. *Answer Probability ($A_P^{(\ell)}$).* A soft metric: the average probability mass assigned to the target concept:

$$A_P^{(\ell)} \triangleq \frac{1}{|\mathcal{I}_O|} \sum_{i \in \mathcal{I}_O} P(s_c \mid h_i^{(\ell)}). \tag{2}$$

#### 3.3.2 When do video-conditioned semantics become language-aligned?

Fig. 3 plots $C_R^{(\ell)}$ and $A_P^{(\ell)}$ as a function of layer depth. Across models and concepts, both metrics remain low in early layers, despite the presence of object-aligned visual tokens in the input, and then rise sharply and consistently within the mid-to-late layer interval. Projection alone thus does not render visual information immediately usable for decoding; object and action semantics undergo depth-dependent transformation inside the LLM, becoming stably expressed in the vocabulary space

only after sufficient processing. Random visual-token positions outside $\mathcal{I}_O$ show no comparable rise, and Tuned Lens (Belrose et al., 2023) and per-layer linear probes reproduce the same mid-to-late profile with higher absolute decodability (Appendix C.2); the profile is therefore a property of the object-aligned states, not of the frozen unembedding. Absolute correspondence on the LLaVA-OV backbones stays below 2%, since a top-1 vocabulary match on an individual visual-token state is a strict criterion; $A_P^{(\ell)}$ gives the graded complement, and all comparisons use the layer-wise profile. The attention knockout and scaling experiments (Appendix C.8 & C.5) show the same localization. We call the interval from the onset of this rise to the final layer the *consolidation interval* $\mathcal{L}_{\text{crit}}$. On the 28-layer backbones in Fig. 3 the onset lies at layer 20 under the rule of Appendix C.3 (19 on LLaVA-OV-I), so $\mathcal{L}_{\text{crit}}$ spans Layers 20 to 28, which §4 uses to place the intervention windows.

#### 3.3.3 How does temporal structure affect semantic consolidation?

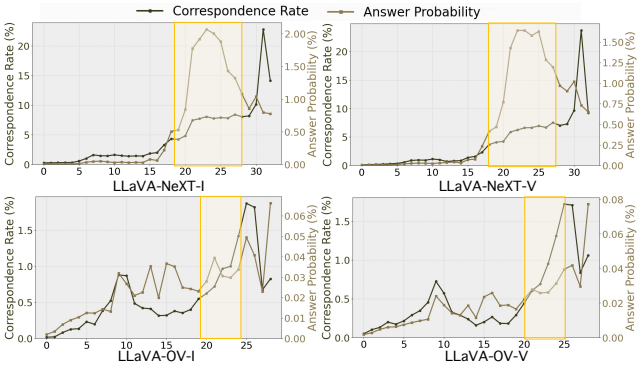


Figure 3: Quantitative analysis of semantic tracing. Both metrics show a sharp increase in the mid-to-late layers, indicating that abstract semantic concepts are consolidated deep within the network.

To test whether consolidation depends on temporal structure, we perturb inter-frame ordering via reversal, local pairwise swap, and global shuffling while preserving per-frame content. Table 9 (Appendix C.7) shows accuracy drops that scale with disruption severity: reversal causes the largest degradation ($-45.6$ average accuracy for `LLaVA-OV-V`, $-46.4$ for the image-trained `LLaVA-OV-I`, $-20.6$ for `Qwen2.5-VL`), local swaps the smallest ($-3.0$ for `LLaVA-OV-V`). Two features of this pattern constrain its interpretation: the image-trained variant is as sensitive to reversal as its video-trained counterpart, and Interaction questions, whose answers do not depend on event order, drop as much as Sequence questions ($-47.1$ vs. $-43.0$ on `LLaVA-OV-V`). The outputs are therefore highly sensitive to frame order, but this sensitivity is tied neither to video training nor to order-dependent questions, which points to position-dependent aggregation of frame evidence rather than explicit modeling of temporal relations; §3.3.4 examines how the evidence is aggregated.

#### 3.3.4 What does semantic tracing reveal about temporal reasoning?

Beyond aggregate layer-wise metrics, we inspect how decoded concepts evolve over time at fixed token positions. As shown in Fig. 10, the decoded concepts at a fixed token position evolve in tandem with the instantaneous visual appearance of each frame: in the standing-up example, "sitting"-related tokens persist into frames where the action has already begun, and only gradually give way to "standing"-related tokens. At the inspected token positions, this lagged, per-frame update reflects a frame-by-frame description strategy: decoded concepts track momentary visual states without expressing the transition itself as a temporal relation; whether such relations are encoded at query or answer positions is not tested here. Together with the degradation under temporal reversal, these behaviors suggest an evidence-aggregation account of temporal understanding in current LVLMs rather than explicit modeling of state transitions, which is the working hypothesis behind the head-level interventions in §4. More interactive examples are provided in Appendix C.9, along with an anonymous website demo.

## 4 CircuitProbe-derived Surgery

Circuits ① and ② identify two prerequisites for effective temporal reasoning in LVLMs: (1) task-critical semantics enter the model through object-aligned video tokens, and (2) these semantics become language-aligned only within a mid-to-late consolidation interval. We turn these observations into a two-stage intervention: identifying *temporal attention heads* using interpretable temporal routing statistics, then selectively amplifying their contributions to test whether they causally support performance on temporal video tasks.

### 4.1 Identifying Temporal Attention Heads

Circuit ② suggests that temporal reasoning depends not on early visual processing but on how evidence from multiple frames is aggregated once concepts become language-aligned. This motivates a testable mechanism: *if* temporal evidence becomes useful mainly after semantic consolidation, *then* heads that both route strongly to video tokens and distribute attention across multiple temporal slices are candidate carriers of cross-frame evidence. We call such heads *temporal heads* and model each as a two-stage router: it first decides whether to attend to video tokens at all, then how to distribute that attention across temporal slices.

**Video routing.** We first quantify whether a head consumes video evidence at all. Following §3, let $q$ denote the answer position, from which the first answer token is decoded, and $\alpha_{q\to j}^{(\ell,h)}$ the attention weight from $q$ to token $j$ at layer $\ell$, head $h$. The attention mass routed from $q$ into a token subset $\mathcal{S}$ is $\mathcal{A}_q^{(\ell,h)}(\mathcal{S}) \triangleq \sum_{j\in\mathcal{S}} \alpha_{q\to j}^{(\ell,h)}$. With video tokens $\mathcal{I}_V = \bigcup_{t=1}^{T} \mathcal{I}_{V,t}$ grouped by frame (or temporal slice) $t$ and non-video tokens $\mathcal{I}_Q$, the *video attention ratio* measures how strongly head $(\ell, h)$ routes attention to the video stream:

$$m^{(\ell,h)} \triangleq \frac{\mathcal{A}_q^{(\ell,h)}(\mathcal{I}_V)}{\mathcal{A}_q^{(\ell,h)}(\mathcal{I}_V \cup \mathcal{I}_Q) + \varepsilon}. \tag{3}$$

$m^{(\ell,h)}$ acts as a gating score, near zero when the head largely ignores video tokens. Only heads with a sufficiently large ratio enter the temporal analysis.

**Temporal distribution across frames.** We next characterize how each video-attending head spreads attention across frames, since a head locked onto one frame, however strongly, cannot integrate temporal evidence. We define the normalized attention mass assigned by head $(\ell, h)$ to slice $t$ as $p_t^{(\ell,h)} \triangleq \frac{\mathcal{A}_q^{(\ell,h)}(\mathcal{I}_{V,t})}{\mathcal{A}_q^{(\ell,h)}(\mathcal{I}_V)+\varepsilon}$, $\sum_{t=1}^{T} p_t^{(\ell,h)} = 1$. A single-frame head yields a highly peaked $p_t$, whereas cross-frame integration yields a broader distribution, which we characterize by its *breadth* (how many frames receive non-trivial attention) and its *temporal span* (how far it extends along time). Breadth is the effective number of slices, the Hill number of order 2 (inverse Simpson index) (Somerfield et al., 2008):

$$n_{\text{eff}}^{(\ell,h)} \triangleq \frac{1}{\sum_{t=1}^{T} \left(p_t^{(\ell,h)}\right)^2 + \varepsilon}. \tag{4}$$

where $n_{\text{eff}} \approx 1$ indicates focus on a single slice and larger values indicate attention spread across frames. Span is the second central moment over normalized slice indices $\tau_t \triangleq \frac{t-1}{T-1} \in [0, 1]$, comparable across $T$:

$$\sigma_\tau^{(\ell,h)} = \sqrt{\sum_{t=1}^{T} p_t^{(\ell,h)} \, (\tau_t - \mu^{(\ell,h)})^2}, \text{ where } \mu^{(\ell,h)} = \sum_{t=1}^{T} p_t^{(\ell,h)} \, \tau_t. \tag{5}$$

Unlike $n_{\text{eff}}$, $\sigma_\tau$ separates local multi-slice smoothing from long-range temporal integration.

**Temporal head selection.** We designate a head $(\ell, h)$ as *temporal* only if it (i) routes non-trivial attention mass to the video stream, (ii) draws support from multiple frames, (iii) spans a sufficiently wide temporal range, and (iv) does so consistently across probe samples; concretely, we require $\bar{m}^{(\ell,h)} \geq \tau_m$, $\overline{n_{\text{eff}}}^{(\ell,h)} \geq \tau_{\text{eff}}$, $\bar{\sigma}_\tau^{(\ell,h)} \geq \tau_\sigma$, with this joint criterion holding on at least a fraction $\tau_{\text{occ}}$ of probe samples. Thresholds $(\tau_m, \tau_{\text{eff}}, \tau_\sigma, \tau_{\text{occ}})$ and metric ablations are given in Appendix D; Fig. 4 illustrates the selection pipeline. The selected set $\mathcal{S}_{\text{temp}}$ is stable: its Jaccard overlap remains $\geq 0.70$ across probe-sample splits, prompt templates, frame counts, and probing datasets (Appendix E.2), and count-matched top video-attention or top-entropy head sets do not reproduce its gains (Appendix D.3).

### 4.2 Temporal Head Amplification

Given the set $\mathcal{S}_{\text{temp}}$ selected above (124 of the 784 heads of Qwen2.5-VL; 56 of the 140 heads in Layers 20–25), we test whether strengthening it causally improves performance on temporal video tasks by amplifying the selected heads within each evaluated layer window $\mathcal{W}$. For an attention head $h$ at layer $\ell$, let $\mathbf{O}^{(\ell,h)} \in \mathbb{R}^{d_k}$ denote its output vector after attention but before the final linear

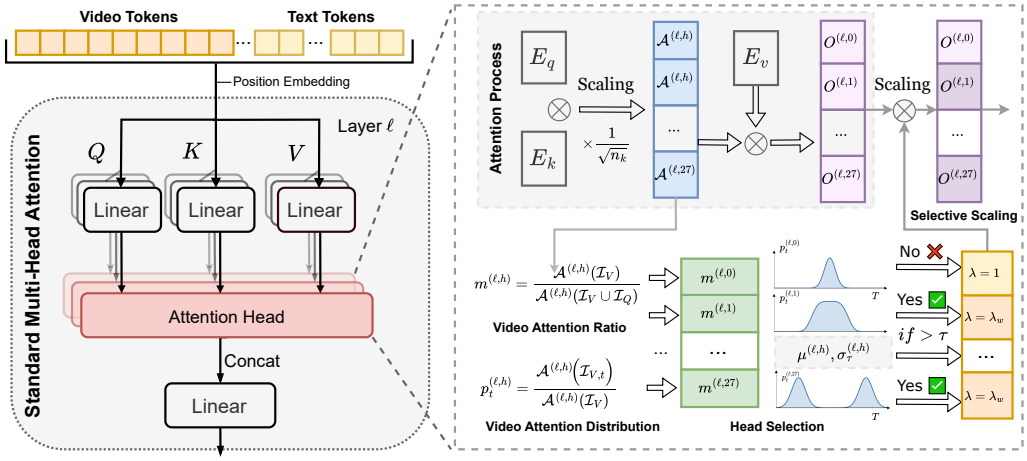


Figure 4: **Surgery framework.** We identify *temporal heads* with high video attention and broad temporal dispersion, then amplify them to test their causal role on temporal video tasks.

projection $\mathbf{W}_O$. We apply a scaling factor $\lambda$ to the identified temporal heads:

$$\tilde{\mathbf{O}}^{(\ell,h)} = \begin{cases} \lambda \cdot \mathbf{O}^{(\ell,h)}, & \text{if } (\ell,h) \in \mathcal{S}_{\text{temp}} \text{ and } \ell \in \mathcal{W} \\ \mathbf{O}^{(\ell,h)}, & \text{otherwise} \end{cases}$$

We fix $\lambda = 1.75$ throughout, selected on TempCompass via the sweep in Fig. 5 and applied unchanged to every other benchmark (Appendix D.1). Circuit ② located $\mathcal{L}_{\text{crit}}$ on the LLaVA backbones (Layers 20 to 28 of 28, §3.3) before any intervention was run; the onset rules of Appendix C.3 were formalized afterwards. On Qwen2.5-VL the onset lies at layer 21 under the primary rule, one layer above LLaVA-OV-V (Appendix C.4), and 88 of the 124 heads passing the temporal criteria lie in Layers 20–27 (Appendix D.10). We evaluate five-layer windows (the last runs to the final layer): Layers 25–30 lies inside $\mathcal{L}_{\text{crit}}$ under every onset rule, Layers 20–25 inside it under the primary rule on the LLaVA-OV backbones and from its second layer onward on Qwen2.5-VL, Layers 15–20 before the onset on Qwen2.5-VL and LLaVA-OV-V, and Layers 0–15 are earlier-layer controls. Scaling changes only the *magnitude* of temporal evidence entering the residual stream, not the head's attention pattern.

### 4.3 Experiments and Validation of Findings

**Experiment results.** Table 3 and Table 2 give two views of the same intervention: a sweep over three LVLMs and two frame counts on TempCompass (Liu et al., 2024b), and an evaluation of Qwen2.5-VL across four capability categories: temporal (TempCompass and NExTQA (Xiao et al., 2021)), intent (IntentQA (Li et al., 2023)), long-video reasoning (EgoSchema (Mangalam et al., 2023) and VideoMME (Fu et al., 2025)), and hallucination resistance (VidHalluc (Li et al., 2025) and VideoHallucer (Wang et al., 2024)).[1] With a single window fixed inside the consolidation interval (Layers 25–30), amplifying the identified temporal heads improves all seven benchmarks without any parameter update, with held-out gains of +4.0 on EgoSchema, +2.6 on VideoMME, and +2.4 on NExTQA (each significant, Appendix D.9) and a smaller, non-significant change on the development benchmark TempCompass (+0.4). Under the official Qwen2.5-VL frame budget the same window gains +2.0 on VideoMME and +2.8 on EgoSchema, both significant (Appendix D.8). The other windows form a sensitivity analysis: the largest per-benchmark gains reach +3.0 on VideoMME (Layers 20–25) and +3.6 on VidHalluc (Layers 15–20), both significant, whereas EgoSchema decreases by 1.2 in Layers 20–25 (not significant). The relevant temporal capacity thus already resides in the model and can be strengthened at inference time. A count-matched random head set yields no gain on VideoMME or EgoSchema (Appendix D.5), and the image benchmarks MMBench and SEED-Bench are unchanged under amplification (Appendix D.6).

**Necessity via suppression.** Amplification establishes only *sufficiency*. To test *necessity*, we scale the same head set with $\lambda \in \{0.5, 0\}$ (Appendix D.6). Suppression yields a dose–response pattern on Qwen2.5-VL: TempCompass drops by 2.7 points at $\lambda$=0.5 and 6.3 points at $\lambda$=0 (NExTQA: −2.2 and −5.1), whereas the non-temporal MMBench and SEED-Bench decline by at most 0.9

[1] †/‡: $p_{\text{holm}} < 0.05/< 0.01$ under Holm-corrected exact McNemar tests; unmarked changes are not significant. Protocol, confidence intervals, and remaining test outcomes: Appendix D.9.

Table 2: Temporal Head Amplification across capability categories on Qwen2.5-VL (32 frames) for the window below the onset (Layers 15–20) and the two windows inside the consolidation interval (Layers 20–25, 25–30); Layers 25–30 provides the headline results and the other two rows a sensitivity analysis. Subscripts: absolute gain (↑) or loss (↓) over the baseline; bold: each benchmark's largest gain.

| **Amplification** | **Temporal** | | **Intent** | **Long-video Reasoning** | | **Hallucination** | |
|---|---|---|---|---|---|---|---|
| **Windows** | TempCompass | NExTQA | IntentQA | EgoSchema | VideoMME | VidHalluc | VideoHallucer |
| *Baseline (None)* | 74.6 | 71.0 | 87.0 | 55.4 | 42.2 | 74.0 | 73.7 |
| Layers 15–20 | 74.8 $_{\uparrow 0.2}$ | 73.0 $_{\uparrow 2.0}^{\dagger}$ | 86.8 $_{\downarrow 0.2}$ | 54.8 $_{\downarrow 0.6}$ | 43.0 $_{\uparrow 0.8}$ | 77.6 $_{\uparrow \mathbf{3.6}}^{\dagger}$ | 73.4 $_{\downarrow 0.3}$ |
| Layers 20–25 | 75.4 $_{\uparrow \mathbf{0.8}}$ | 71.0 $_{\uparrow 0.0}$ | 86.8 $_{\downarrow 0.2}$ | 54.2 $_{\downarrow 1.2}$ | 45.2 $_{\uparrow \mathbf{3.0}}^{\ddagger}$ | 74.8 $_{\uparrow 0.8}$ | 73.9 $_{\uparrow 0.2}$ |
| Layers 25–30 | 75.0 $_{\uparrow 0.4}$ | 73.4 $_{\uparrow \mathbf{2.4}}^{\dagger}$ | 88.4 $_{\uparrow \mathbf{1.4}}$ | 59.4 $_{\uparrow \mathbf{4.0}}^{\ddagger}$ | 44.8 $_{\uparrow 2.6}^{\dagger}$ | 75.0 $_{\uparrow 1.0}$ | 74.8 $_{\uparrow \mathbf{1.1}}$ |

points, and ablating a count-matched random set of video-attending heads costs only 0.8 points on TempCompass. The evaluated temporal tasks thus depend causally on the identified heads rather than on video-attending heads in general. Since the two benchmark groups also differ in modality (multi-frame video vs. single image), this contrast establishes task-selective dependence; separating temporal-order computation from general multi-frame integration needs same-video controls (Appendix D.6). The asymmetry between modest amplification gains and large ablation losses suggests a near-saturated mechanism: the heads already operate near their effective regime, yet removing them impairs cross-frame routing.

**Validation of CircuitProbe findings.** On the two video-trained LVLMs (Qwen2.5-VL and LLaVA-OV-V), the positive changes in Table 3 concentrate in Layers 20–30; Layers 15–20, before the onset on both backbones, is neutral on TempCompass, and lower windows yield negligible or slightly negative effects. The LLaVA-OV-V changes lie within $\pm 1.0$, and only Layers 25–30 at $T{=}16$ is significant (Appendix D.9). The image-trained LLaVA-OV-I shows no such pattern: video-trained models contain 2.2–2.6× more heads meeting the full temporal criteria (106–124 vs. 48), and the image-trained variant's heads, though in the same layers, do not improve accuracy when amplified (Appendix D.10). On the held-out benchmarks of Table 2 the layer specificity is weaker: six of the seven benchmarks reach their largest gain in Layers 20–30, where all significant EgoSchema and VideoMME gains lie, yet Layers 15–20, entirely before the Qwen2.5-VL onset and containing 20 of the selected heads, already yields significant gains on NExTQA (+2.0) and VidHalluc (+3.6, its largest gain). The consolidation interval thus marks where amplification helps most consistently, not a sharp boundary below which the selected heads have no effect (qualitative cases: Appendix D.4).

Table 3: Temporal Head Amplification on TempCompass with 16- and 32-frame inputs, without retraining (protocols in §3.1). Shaded rows lie inside the consolidation interval.

| **Amplification** | **Qwen 2.5 VL** | | **LLaVA-OV-V** | | **LLaVA-OV-I** | |
|---|---|---|---|---|---|---|
| **Windows** | $T=16$ | $T=32$ | $T=16$ | $T=32$ | $T=16$ | $T=32$ |
| *Baseline (None)* | 73.6 | 74.6 | 72.2 | 72.2 | 63.6 | 63.6 |
| Layers 0–5 | 73.6 $_{\downarrow 0.0}$ | 74.2 $_{\downarrow 0.4}$ | 72.0 $_{\downarrow 0.2}$ | 71.4 $_{\downarrow 0.8}$ | 63.8 $_{\uparrow 0.2}$ | 63.6 $_{\downarrow 0.0}$ |
| Layers 5–10 | 73.4 $_{\downarrow 0.2}$ | 74.4 $_{\downarrow 0.2}$ | 72.2 $_{\downarrow 0.0}$ | 72.4 $_{\uparrow 0.2}$ | 63.6 $_{\downarrow 0.0}$ | 63.8 $_{\uparrow 0.2}$ |
| Layers 10–15 | 73.2 $_{\downarrow 0.4}$ | 74.2 $_{\downarrow 0.4}$ | 72.6 $_{\uparrow 0.4}$ | 72.2 $_{\downarrow 0.0}$ | 63.2 $_{\downarrow 0.4}$ | 63.8 $_{\uparrow 0.2}$ |
| Layers 15–20 | 73.4 $_{\downarrow 0.2}$ | 74.8 $_{\uparrow 0.2}$ | 71.8 $_{\downarrow 0.4}$ | 72.2 $_{\downarrow 0.0}$ | 62.6 $_{\downarrow 1.0}$ | 63.2 $_{\downarrow 0.4}$ |
| Layers 20–25 | 76.0 $_{\uparrow \mathbf{2.4}}$ | 75.4 $_{\uparrow 0.8}$ | 72.2 $_{\uparrow 0.0}$ | 72.6 $_{\uparrow 0.4}$ | 62.8 $_{\downarrow 0.8}$ | 62.6 $_{\downarrow 1.0}$ |
| Layers 25–30 | 74.0 $_{\uparrow 0.4}$ | 75.0 $_{\uparrow 0.4}$ | 73.2 $_{\uparrow \mathbf{1.0}}^{\ddagger}$ | 72.2 $_{\downarrow 0.0}$ | 63.0 $_{\downarrow 0.6}$ | 63.2 $_{\downarrow 0.4}$ |

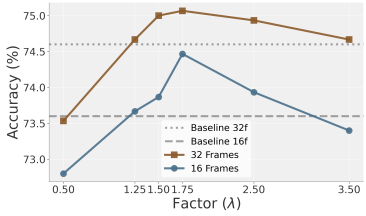


Figure 5: $\lambda$ sensitivity of Temporal Head Amplification on TempCompass, averaged over the three windows in Layers 15–30.

## 5 Conclusion and Limitation

CircuitProbe mechanistically dissects video LVLMs through (i) *Visual Auditing*, which localizes task-critical object-aligned video tokens and causally verifies their necessity, and (ii) *Semantic Tracing*, which shows that video semantics become language-aligned mainly in mid-to-late layers, where decoded concepts track frame-wise evidence rather than explicit temporal-state transitions. Bidirectional interventions on the temporally dispersed, video-routing heads we identify show that they causally contribute to performance on temporal video benchmarks: amplification within the consolidation interval yields Holm-significant gains on held-out EgoSchema and VideoMME for Qwen2.5-VL, with smaller changes on the development benchmark and on LLaVA-OV-V, while suppression produces dose-dependent degradation. The intervention evidence rests on two 7B video-trained backbones, and the LLaVA-OV-V gains are significant in one of four TempCompass cells, so the head-level effect is established mainly on Qwen2.5-VL; larger models, other architectures, and amplification side effects such as calibration shift remain untested.

### AI USE STATEMENT

We used LLMs solely for language polishing in this work. Specifically, after completing the initial draft entirely through human effort, we employed LLM assistance exclusively for grammatical refinement and improving the clarity of English expression to meet academic writing standards. We did not use generative AI tools for research ideation, method design, experiment execution, data analysis, code generation, or the production of any reported results or figures. All intellectual contributions, from conceptualization to initial manuscript preparation, were performed by the human authors, and all AI-assisted language edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts.

### ETHICS STATEMENT

This work analyzes and intervenes on publicly released vision–language models using publicly available video benchmarks (STAR, TempCompass, NExTQA, IntentQA, EgoSchema, VideoMME, VidHalluc, VideoHallucer, MMBench, SEED-Bench); no human subjects, private data, or personally identifiable information are involved. Our interventions modify model behavior only at inference time and are studied for understanding and improving temporal reasoning; potential societal impacts of the underlying models and of interpretability-guided interventions are discussed in Appendix F.

### REPRODUCIBILITY STATEMENT

Dataset curation and probing protocols are specified in §3.1 and Appendix A, including the counterfactual-validity filtering criteria. The logit-lens setup, probe cross-validation details, the onset rule, and the Qwen2.5-VL tracing are given in Appendix C.1, Appendix C.2, Appendix C.3, and Appendix C.4. Temporal-head selection thresholds, head-set sizes and per-window counts, the amplification factor $\lambda$ and its calibration protocol, and compute resources are documented in Appendix D, Appendix D.1, and Appendix B; the evaluation frame budget and its official-budget replication are stated in Appendix D.2 and Appendix D.8. Statistical procedures (paired bootstrap and exact McNemar tests on per-sample predictions) are described in Appendix D.9.

## REFERENCES


Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers, 2026. URL https://arxiv.org/abs/2510.21518.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

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 (NeurIPS)*, 2025. URL https://openreview.net/forum?id=7mTECPRtll. arXiv:2505.12826.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning, 2023. URL https://arxiv.org/abs/2310.09478.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Peitao Han, Fei Cheng, Lis K. Pereira, Qianying Liu, and Shigeru Kitazawa. Tracing the arrow of time: Diagnosing temporal information flow in Video-LLMs, 2026. URL https://arxiv.org/abs/2605.07568.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025. doi: 10.18653/v1/2025.findings-acl.1024. URL https://aclanthology.org/2025.findings-acl.1024/.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations, 2025. URL https://arxiv.org/abs/2410.02762.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. *arXiv preprint arXiv:2504.19475*, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv.org/abs/2411.17491.

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision-language models, 2025. URL https://arxiv.org/abs/2509.17588.

Minji Kim, Taekyung Kim, and Bohyung Han. Map the flow: Revealing hidden pathways of information in VideoLLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.13251. arXiv:2510.13251.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

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.

Chengzhi Li, Heyan Huang, Ping Jian, Zhen Yang, Yaning Tian, and Zhongbin Guo. Understanding temporal logic consistency in video-language models through cross-modal attention discriminability, 2026. URL https://arxiv.org/abs/2510.08138.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024b.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11963–11974, October 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. *arXiv preprint arXiv:2410.06166*, 2024c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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? *arXiv preprint arXiv: 2403.00476*, 2024b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, 2023. URL https://arxiv.org/abs/2306.05424.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

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

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space, 2023. URL https://arxiv.org/abs/2209.15162.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model, 2023. URL https://arxiv.org/abs/2309.16058.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models, 2024. URL https://arxiv.org/abs/2410.07149.

Nostalgebraist. Interpreting GPT: The logit lens - AI Alignment Forum. https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=t0FI3Q66K5.

Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridging vision and language spaces with assignment prediction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=lK2V2E2MNv.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. Does time have its place? Temporal heads: Where language models recall time-specific information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16616–16643, 2025. doi: 10.18653/v1/2025.acl-long.812. URL https://aclanthology.org/2025.acl-long.812/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Achyuta Rajaram, Neil Chowdhury, Antonio Torralba, Jacob Andreas, and Sarah Schwettmann. Automatic discovery of visual circuits, 2024. URL https://arxiv.org/abs/2404.14349.

Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. *arXiv preprint arXiv:2404.03646*, 2024.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *ArXiv*, abs/2402.01761, 2024. URL https://api.semanticscholar.org/CorpusID:267412530.

PJ Somerfield, KR Clarke, and RM Warwick. Simpson index. In *Encyclopedia of ecology*, pp. 3252–3255. Elsevier, 2008.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://doi.org/10.1109/CVPR52733.2024.00914.

Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:268032058.

Martina G. Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space, 2023. URL https://arxiv.org/abs/2310.18969.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, et al. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*, 2025a.

Ruochen Wang, Si Si, Felix Yu, Dorothea Wiesmann Rothuizen, Cho-Jui Hsieh, and Inderjit S Dhillon. Large language models are interpretable learners. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=hTphfqtafO.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models, 2024. URL https://arxiv.org/abs/2405.17820.

Bo Wu, Shoubin Yu, Zhenfang Chen, and Joshua B Tenenbaum. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *arXiv preprint arXiv:2405.17969*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models, 2023. URL https://arxiv.org/abs/2306.13549.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *arXiv preprint arXiv:2503.19900*, 2025.

Zeping Yu and Sophia Ananiadou. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3293–3306, 2024.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. *ArXiv*, abs/2404.09387, 2024a. URL https://api.semanticscholar.org/CorpusID:269149157.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. Beyond training: Dynamic token merging for zero-shot video understanding. *ArXiv*, abs/2411.14401, 2024b. URL https://api.semanticscholar.org/CorpusID:274166248.

# APPENDIX

## A ANALYSIS SETUP

### A.1 QUESTION PROBING SETUP

Two complementary question formats probe object and action understanding under token-level interventions with controlled decoding:

1. *Open-ended questions.* We formulate questions targeting objects and actions depicted in the video (e.g., "Which object did the person throw in the video?"), each answerable from the provided key frames alone. We pre-fill the model's response with a fixed prefix (e.g., "The object is") and compare the next generated token before and after intervention, so that every ablation condition is scored at the same position.
2. *Close-ended questions.* We prompt the model with binary questions (e.g., "Is there a/an [object] in this video?") and compare the model's next-token prediction before and after intervention. A flip from "Yes" to "No" after the intervention indicates that the manipulated tokens were necessary for the prediction.

### A.2 FRAMES VISUALIZATION

Figure 6 shows sample input keyframes; each row is one video sequence fed to the LVLMs. The examples cover everyday actions, objects, and environments, with changes in viewpoint, object scale, and partial occlusion.



Figure 6: Visualization of Input Keyframes. Each row displays a sequence of frames provided to the model as input for a specific video. The red bounding boxes mark the ground-truth object referenced by the question (e.g., the object being picked up, kicked, or taken); they are shown for clarity and were not provided to the model during inference.

### A.3 COUNTERFACTUAL-VALIDITY FILTERING AND SUBSET DEFINITIONS

We partition the STAR analysis pool into three nested subsets:

- **Full dataset.** The complete STAR subset used in our analysis ($N = 1720$).
- **Correct subset.** Examples from the full dataset that the model answers correctly under the unperturbed input ($N = 731$ for `LLaVA-OneVision-V`).

- **Counterfactual-valid subset.** Examples in the Correct subset whose answers are *not* trivially recoverable from language priors, dataset shortcuts, or semantic leakage. This subset, used in the main paper, isolates cases where the visual stream is causally necessary.

Table 4: Object-token ablation on the **Correct subset** and the **Full dataset**, complementing the counterfactual-valid results in the main paper. The same degradation pattern holds on all three subsets.

| Ablation Type | LLaVA-OneVision-V | | LLaVA-OneVision-I | |
|---|---|---|---|---|
| | Correct Set | Full Dataset | Correct Set | Full Dataset |
| *Baseline* | 100.0 | 42.5 | 100.0 | 38.1 |
| Object | 68.9 | 34.5 | 52.9 | 21.2 |
| Object (buffer 1) | 63.1 | 31.3 | 45.7 | 18.1 |
| Object (buffer 2) | 61.4 | 30.3 | 42.6 | 16.7 |

Table 4 reports the Circuit ① ablation on the Correct subset and the Full dataset; the same monotone degradation pattern (Object → buffer 1 → buffer 2) holds without the counterfactual filter.

# B Experiments Compute Resources

All experiments were run on a single compute node with 8 NVIDIA A100 GPUs. One experiment, parallelized across the 8 GPUs, takes about 1.5 hours on average, for the circuit-probing analyses (Circuits ① and ②) as well as the temporal-head amplification surgery.

# C Supplementary Experiments and Details: Circuit ① and ②

## C.1 Logit lens Setup

The analysis reads the internal representations of projected visual tokens as they propagate through the LLM backbone, before any autoregressive generation. Let $\mathcal{I}_O$ denote the set of object-centric visual token indices identified in Circuit ①, and let $n_O \triangleq |\mathcal{I}_O|$. For each token position $i \in \mathcal{I}_O$, we denote by $h_i^{(\ell)} \in \mathbb{R}^d$ the hidden state at layer $\ell$. At each layer, we project this hidden state into the vocabulary space using the frozen language-model output projection $W_{LM}$, defining a layer-wise token distribution

$$P(\cdot \mid h_i^{(\ell)}) = \mathrm{softmax}(W_{LM} h_i^{(\ell)}). \tag{6}$$

For each sample, we define a target semantic concept $s_c$ corresponding to the ground-truth object or action referenced by the question (e.g., "`microwave`"). If $s_c$ is tokenized into multiple sub-word tokens, we define the target token set $\mathcal{T}(s_c) \subseteq \mathcal{V}$ accordingly, and use

$$P(s_c \mid h_i^{(l)}) \triangleq \sum_{w \in \mathcal{T}(s_c)} P(w \mid h_i^{(\ell)}). \tag{7}$$

When $s_c$ maps to a single token, $\mathcal{T}(s_c) = \{w_c\}$ and the definition reduces to $P(w_c \mid h_i^{(\ell)})$.

## C.2 Probe Cross-Validation: Tuned Lens, Linear Probes, and Position Controls

The frozen unembedding head is trained on final-layer language-token states; applied to visual-token states at intermediate layers, it could decode vocabulary that reflects the projection rather than the representation. To test whether the mid-to-late semantic consolidation in Fig. 3 survives calibrated readouts, we repeat the layer-wise tracing on `LLaVA-OV-V` with two alternative probes and one position control:

- **Tuned Lens** (Belrose et al., 2023): a per-layer affine translator trained to match the final-layer distribution, correcting for representational drift across depth. Translators are trained on a held-out set of videos disjoint from the evaluation set.
- **Linear probes**: a per-layer linear classifier trained to predict the target concept $s_c$ from $h_i^{(\ell)}$, with train/test splits partitioned *by video* to prevent leakage.
- **Random-position control**: the original logit lens applied to randomly chosen visual-token positions *outside* the object-aligned set $\mathcal{I}_O$. If decoded concepts were artifacts of the unembedding projection itself, this control should exhibit the same layer-wise pattern.

As shown in Table 5, both calibrated probes increase the hit counts at every depth, so early layers carry partially decodable information that the raw logit lens underestimates, yet the profile is unchanged:

Table 5: Correspondence-hit counts within layer bands on `LLaVA-OV-V`: the number of probed token positions (three to five object-aligned positions per sample) whose top-ranked projection matches the target concept, summed over the layers in each band. These are raw counts over the probed subset, not the per-layer rates $C_R^{(\ell)}$ of Fig. 3, which normalize by the full $|\mathcal{I}_O|$. The bands have unequal width (the last two are narrower than the first two), so counts are compared across bands after normalizing by band width, as in the text. Calibrated probes (Tuned Lens, linear probe) raise decodability at every depth and preserve the profile of the logit lens: few hits before layer 20 and many from layer 20 onward, while the random-position control stays near zero at all depths.

| Probe | Layers 0–10 | Layers 10–20 | Layers 20–25 | Layers 25+ |
|---|---|---|---|---|
| Logit lens (original) | 2 | 8 | 39 | 33 |
| Tuned Lens | 8 | 21 | 58 | 49 |
| Linear probe | 15 | 34 | 71 | 66 |
| Logit lens @ random positions | 0 | 2 | 4 | 3 |

hits rise sharply from Layers 20 onward (Tuned Lens: $21 \to 58$ between the second and third bands; linear probe: $34 \to 71$) and, normalized by band width, per-layer decodability is highest in the final band for all three probes, matching the consolidation interval $\mathcal{L}_{\text{crit}}$ (Layers 20 to 28) defined in the main text. The random-position control never exceeds 4 hits in any band, an order of magnitude below the object-aligned counts at the same depth. Normalized by band width, its rate also increases with depth, from 0.2 hits per layer in Layers 10–20 to 0.8 in Layers 20–25, the pattern that the full random-position curves on Qwen2.5-VL show as well (Appendix C.4); the depth at which decodability rises is thus partly shared by visual-token states in general, whereas the magnitude of the rise is specific to the object-aligned positions. The unembedding matrix therefore does not decode task-relevant vocabulary from arbitrary visual tokens at a comparable rate, and the consolidation interval reflects video-conditioned semantic emergence at the object-aligned positions rather than a property of the logit-lens readout.

### C.3 Onset Rule for the Consolidation Interval

The main text places the onset of $\mathcal{L}_{\text{crit}}$ at layer 20 on the 28-layer backbones of Fig. 3. This section defines the onset as a function of the layer-wise curve, so that it can be recomputed on any backbone. The rules were formalized after the interventions of §4 had been run, and the baseline window and floor of the primary rule R3′ were chosen after inspecting the LLaVA-OV curves; we therefore report the onset under four rules, so that no claim rests on a single choice of rule. All rules act on the correspondence-rate curve $C_R^{(\ell)}$ at object-aligned positions over the block outputs $\ell = 0, \ldots, L-1$; the final unembedding point plotted in Fig. 3 is excluded from the maximum. Let $\mu_{a\text{–}b}$ and $\sigma_{a\text{–}b}$ denote the mean and population standard deviation of $C_R^{(\ell)}$ over Layers $a$ to $b$ inclusive.

- **R1 (half-maximum):** the first layer at which $C_R^{(\ell)} \geq \frac{1}{2} \max_{\ell'} C_R^{(\ell')}$.
- **R2 (steepest rise):** $\arg\max_{\ell} [C_R^{(\ell)} - C_R^{(\ell-1)}]$.
- **R3 (early-layer baseline):** the first layer with $C_R^{(\ell)} > \mu_{0\text{–}10} + 3\sigma_{0\text{–}10}$.
- **R3′ (mid-layer baseline, primary):** the first layer after the baseline window ($\ell \geq 19$) with $C_R^{(\ell)} > \mu_{10\text{–}18} + 3\sigma_{10\text{–}18}$. The baseline window starts after the transient bump at layer 9 that the LLaVA-OV backbones and Qwen2.5-VL show (Fig. 3, Fig. 7); under R3 this bump inflates $\sigma_{0\text{–}10}$.

Table 6 lists the outputs. On the 32-layer LLaVA-NeXT backbones the half-maximum and steepest-rise rules agree on layer 20 and R3′ gives 20 (R3: 17). On the 28-layer LLaVA-OV backbones the onset is layer 20 under R3′ (19 on OV-I), the value used for $\mathcal{L}_{\text{crit}}$ in the main text, and layers 22–23 under the half-maximum rule R1; R2 and R3 place it at 24–25, because $C_R$ on these backbones keeps rising up to layer 25 and reaches half of its maximum only in the upper part of the rise. On the 60-layer 34B model the rules give layers 41 to 45, in line with the rise from around layer 40 in Appendix C.5. On Qwen2.5-VL, whose curves are given in Appendix C.4, the onset is 21 under R3′ and 22 under R1; between this backbone and LLaVA-OV-V, the two 28-layer video-trained backbones, the onset differs by at most two layers under every rule (R1: 1, R2: 0, R3: 2, R3′: 1), so the agreement between the two does not depend on the choice of rule. The headline window of §4 does not depend on this spread: on the three 28-layer backbones on which the interventions are run, Layers 25–30 lies inside the consolidation interval under every rule. Layers 20–25 lies inside the interval under R3′ on the LLaVA-OV backbones, begins one layer before the R3′ onset on

Table 6: Onset of the consolidation interval under the four rules, computed on the $C_R^{(\ell)}$ curves of Fig. 3, Fig. 8, and Fig. 7 over block outputs $0..L{-}1$. The last column applies R1/R2/R3 to $A_P^{(\ell)}$ instead; on the LLaVA-OV and Qwen2.5-VL backbones the half-maximum and steepest-rise rules are then triggered by the layer-9 bump. Between the two 28-layer video-trained backbones, Qwen2.5-VL and LLaVA-OV-V, the onset differs by at most two layers under every rule (R1: 1, R2: 0, R3: 2, R3′: 1).

| **Backbone** | $L$ | **R1** | **R2** | **R3** | **R3′ (primary)** | **R1/R2/R3 on $A_P$** |
|---|---|---|---|---|---|---|
| LLaVA-NeXT-I | 32 | 20 | 20 | 17 | 20 | 20/20/17 |
| LLaVA-NeXT-V | 32 | 20 | 20 | 17 | 20 | 20/20/17 |
| LLaVA-OV-I | 28 | 22 | 25 | 24 | 19 | 9/9/21 |
| LLaVA-OV-V | 28 | 23 | 25 | 24 | 20 | 9/9/24 |
| LLaVA-NeXT-34B-V | 60 | 45 | 44 | 41 | 41 | 46/44/40 |
| Qwen2.5-VL | 28 | 22 | 25 | 22 | 21 | 9/9/23 |

Qwen2.5-VL, straddles the onset under R1 and R3, and precedes it under R2. Layers 15–20 ends at the R3′ onset on LLaVA-OV-I and one to two layers before it on LLaVA-OV-V and Qwen2.5-VL, and precedes the onset under the other three rules.

### C.4 Semantic Tracing on Qwen2.5-VL

The layer-wise tracing of §3.3 used the LLaVA backbones. We repeat it on `Qwen2.5-VL-7B-Instruct`, the 28-layer backbone on which the interventions of §4 are evaluated, with the same STAR counterfactual-valid subset, open-ended format, and prefix forcing as Fig. 3 ($N{=}312$ samples whose boxes map onto the token grid), so that the consolidation interval on this backbone is located by the tracing itself rather than transferred from the LLaVA backbones. Qwen2.5-VL merges $2{\times}2$ spatial patches and two consecutive frames into one token, so each token covers a $28{\times}28$-pixel region in two frames. A merged token belongs to $\mathcal{I}_O$ only if its footprint lies fully inside the STAR bounding box in both frames (mean $|\mathcal{I}_O|{=}168$). A majority-overlap variant, which admits tokens with at least half of the footprint inside the box, gives a mean $|\mathcal{I}_O|$ of 262; all curves below use the strict definition. The probe is the logit lens of Appendix C.1, and the random-position control of Appendix C.2 draws positions from $\mathcal{I}_V \setminus \mathcal{I}_O$. Tuned Lens was reserved for two cases: random-position hits in Layers 20 and beyond reaching one third of the object-aligned hits, which would indicate a readout artifact, or the primary and secondary onset rules disagreeing by more than two layers, which would indicate a curve too noisy for a rule-based onset. Neither occurs here (ratio 0.08; onsets 21 under R3′ and 22 under R1), so no translators were trained for this backbone.

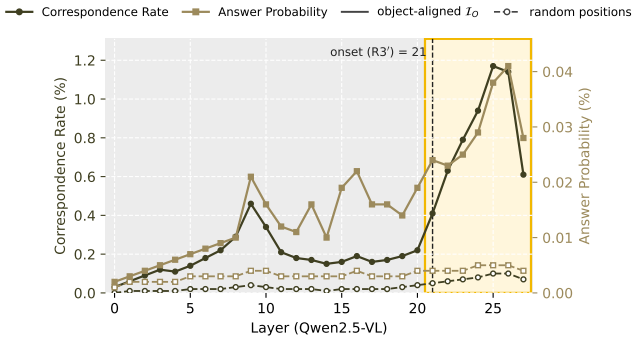


Figure 7: Semantic tracing on `Qwen2.5-VL` (32 frames, $N{=}312$): $C_R^{(\ell)}$ (left axis) and $A_P^{(\ell)}$ (right axis) at object-aligned positions (solid) and at random visual-token positions outside $\mathcal{I}_O$ (dashed), over block outputs $0..27$. The vertical line marks the onset under R3′ (layer 21); the box spans the consolidation interval from that onset to the last layer. $C_R$ at object-aligned positions peaks at $1.17\%$ (layer 25); the random-position curve rises at the same depth but peaks at $0.10\%$.

Fig. 7 shows the same profile as the LLaVA backbones: both metrics stay low through the early layers apart from the transient bump at layer 9, then rise from layer 21 onward, with $C_R$ climbing from $0.22\%$ at layer 20 to $1.17\%$ at layer 25. The onset under the primary rule R3′ is layer 21,

Table 7: Correspondence-hit counts within layer bands on `Qwen2.5-VL`, same protocol as Table 5: counts over the probed subset (three to five object-aligned positions per sample, $N$=312), summed over the layers in each band; bands have unequal width. With four probed positions per sample on average, the implied per-layer hit rates, 0.66% in Layers 20–25 and 0.96% in Layers 25+, match the average $C_R^{(\ell)}$ over the full $\mathcal{I}_O$ in Fig. 7 (0.60% and 0.97%).

| Probe | Layers 0–10 | Layers 10–20 | Layers 20–25 | Layers 25+ |
|---|---|---|---|---|
| Logit lens (object-aligned) | 3 | 10 | 41 | 36 |
| Logit lens @ random positions | 0 | 1 | 3 | 3 |

Table 8: Accuracy (%) under visual-token ablation for the 34B models; the ↓ subscript gives the drop from the baseline.

| Ablation Type | Tokens Number | LLaVA-NeXT-34B-I Open | LLaVA-NeXT-34B-I Close | LLaVA-NeXT-34B-V Open | LLaVA-NeXT-34B-V Close |
|---|---|---|---|---|---|
| *Control Groups (Low Tokens)* | | | | | |
| Baseline | 0 | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ | $100.0_{\downarrow 0}$ |
| Register | 13 | $55.3_{\downarrow 44.68}$ | $96.9_{\downarrow 3.14}$ | $55.3_{\downarrow 44.68}$ | $98.2_{\downarrow 1.8}$ |
| *Object-based Ablation* | | | | | |
| Object | 304 | $9.3_{\downarrow 90.75}$ | $29.8_{\downarrow 70.16}$ | $14.6_{\downarrow 85.37}$ | $11.7_{\downarrow 88.35}$ |
| | 413 | $5.8_{\downarrow 94.24}$ | $21.1_{\downarrow 78.88}$ | $10.7_{\downarrow 89.32}$ | $11.3_{\downarrow 88.72}$ |
| | 573 | $5.8_{\downarrow 94.24}$ | $18.0_{\downarrow 82.02}$ | $10.3_{\downarrow 89.74}$ | $11.8_{\downarrow 88.21}$ |
| *Control Groups (High Tokens)* | | | | | |
| Random | 100 | $54.3_{\downarrow 45.72}$ | $96.9_{\downarrow 3.14}$ | $93.1_{\downarrow 6.92}$ | $29.7_{\downarrow 70.26}$ |
| | 350 | $55.5_{\downarrow 44.5}$ | $96.5_{\downarrow 3.49}$ | $88.5_{\downarrow 11.54}$ | $31.0_{\downarrow 68.97}$ |
| | 500 | $54.5_{\downarrow 45.55}$ | $96.3_{\downarrow 3.66}$ | $83.1_{\downarrow 16.92}$ | $30.8_{\downarrow 69.23}$ |
| | 900 | $59.9_{\downarrow 40.14}$ | $96.8_{\downarrow 3.17}$ | $78.3_{\downarrow 21.67}$ | $29.7_{\downarrow 70.33}$ |

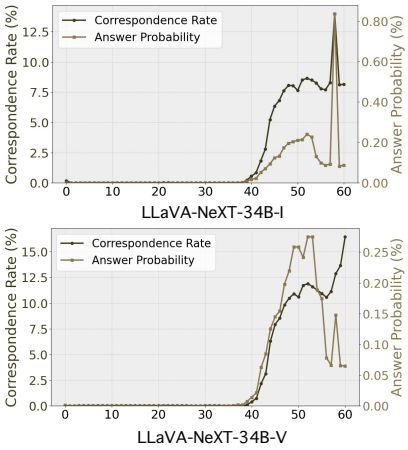


Figure 8: Semantic tracing on the 34B backbone.

one layer above the LLaVA-OV-V onset under the same rule, and the two 28-layer video-trained backbones agree to within two layers under every rule (Table 6). The intervention windows of §4 are therefore placed relative to an onset measured on Qwen2.5-VL itself: Layers 25–30 lies inside the interval under every rule, Layers 20–25 begins one layer before the R3′ onset, and Layers 15–20 precedes it (Appendix C.3). Table 7 gives the band-wise hit counts: 77 hits in Layers 20 and beyond at object-aligned positions against 6 at random positions (ratio 0.08), the same profile as on LLaVA-OV-V (Table 5). The random-position curve nevertheless rises at the same depth (onset 20 under R3′, 21 under R1), at about one tenth of the amplitude: $C_R$ peaks at 0.10% against 1.17% and $A_P$ at 0.005% against 0.041%. As on LLaVA-OV-V (Appendix C.2), the depth of the rise is partly generic to visual-token states on this backbone, while its magnitude is specific to the object-aligned positions that locate the interval.

### C.5 Scaling Experiments

We repeat the core analyses on the LLaVA-NeXT-34B variants. In the visual token ablation (Table 8), ablating object tokens degrades performance far more than removing a larger number of random tokens, matching the 7B pattern.

Semantic tracing on the 34B backbone (Fig. 8) shows the same profile: both Correspondence Rate and Answer Probability stay negligible through the early layers and rise sharply and concurrently from around layer 40. Semantic localization and late-layer consolidation therefore hold at both scales tested.

### C.6 Bounding-Box Source and Robustness to Box Precision

All bounding boxes used in Circuit ① are taken from the official annotations of the STAR benchmark (Wu et al., 2021); no external detector is used. The degradation pattern does not depend on tight box boundaries: as Table 1 of the main paper shows, expanding the object region by one or two tokens of buffer (buffer 1, buffer 2) preserves the pattern and produces larger drops on the expanded regions.

### C.7 Temporal Perturbation Results

Table 9 gives the full temporal-robustness results referenced in §3.3.3 of the main text.

Table 9: **Temporal Robustness Analysis.** Results are reported as the **Accuracy (%)** on the correct-only subset ($T$=16 frames; $N$=517 for `LLaVA-OV-V`, 478 for `LLaVA-OV-I`, 608 for `Qwen2.5-VL`) across three temporal perturbation types: Reverse, Local Swap, and Shuffle. Higher values indicate better stability against temporal perturbations. $\downarrow \Delta$ denotes the performance drop relative to the unperturbed baseline. Interaction questions do not depend on event order; the image-trained `LLaVA-OV-I` is as sensitive to reversal as the video-trained `LLaVA-OV-V`.

| Model | Reverse | | | Local Swap | | | Shuffle | | |
|---|---|---|---|---|---|---|---|---|---|
| | Interaction | Sequence | Avg | Interaction | Sequence | Avg | Interaction | Sequence | Avg |
| LLaVA-OV-V | 52.9 $_{\downarrow 47.1}$ | 57.0 $_{\downarrow 43.0}$ | 54.4 $_{\downarrow 45.6}$ | 96.4 $_{\downarrow 3.6}$ | 98.2 $_{\downarrow 1.8}$ | 97.0 $_{\downarrow 3.0}$ | 75.5 $_{\downarrow 24.5}$ | 85.5 $_{\downarrow 14.5}$ | 79.3 $_{\downarrow 20.7}$ |
| LLaVA-OV-I | 54.1 $_{\downarrow 45.9}$ | 52.6 $_{\downarrow 47.4}$ | 53.6 $_{\downarrow 46.4}$ | 96.6 $_{\downarrow 3.4}$ | 98.7 $_{\downarrow 1.3}$ | 97.4 $_{\downarrow 2.6}$ | 79.7 $_{\downarrow 20.3}$ | 87.0 $_{\downarrow 13.0}$ | 82.4 $_{\downarrow 17.6}$ |
| Qwen2.5-VL | 80.7 $_{\downarrow 19.3}$ | 77.3 $_{\downarrow 22.7}$ | 79.4 $_{\downarrow 20.6}$ | 88.3 $_{\downarrow 11.7}$ | 92.0 $_{\downarrow 8.0}$ | 89.6 $_{\downarrow 10.4}$ | 87.2 $_{\downarrow 12.8}$ | 92.4 $_{\downarrow 7.6}$ | 89.1 $_{\downarrow 10.9}$ |

### C.8 Attention Knockout

Circuit ① localizes task-relevant object semantics in the video embedding, but it does not show *how* that evidence is routed through the LLM backbone when the answer is produced, or whether object-centric evidence alone supports the answer. This section characterizes the *attention flow* into the *first generated token* and gives a second line of evidence for the layer localization of Circuit ②.

**Setup and masking protocol.** We conduct all experiments on the `LLaVA-NeXT` family. To control information flow during prediction of the first output token, we apply an attention knockout mask that blocks attention edges from a specified source token set to the final answer token, restricted to a chosen layer window. We partition layers into five windows (Early, Early-to-Mid, Mid, Mid-to-Late, Late), and apply the same mask to *all heads* within the selected window.

We consider two complementary masking regimes: (i) **object-centric masking** blocks attention from object-region tokens and buffers (O, O+1, O+2) to the final token; (ii) **contextual masking** blocks attention from all non-object tokens and their complements (C, C+1, C+2) to the final token. For each configuration, we report (a) **accuracy drop** (Table 10) and (b) **answer probability drop** for the correct token (Fig. 9).

**Analysis.** The knockout asks a different question from Circuits ①–②: *which* information source (object vs. context) the LLM backbone consumes when forming the first answer token, and *where* in depth this consumption matters most. Table 10 and Fig. 9 show a consistent separation between the roles of contextual versus object-centric information for first-token prediction.

**From localization to utilization.** Circuit ① shows that ablating object-region tokens severely harms object identification, so object semantics are present and localized. For the first token, however, *blocking access to non-object context* degrades prediction far more than blocking access to the object region (Table 10, Fig. 9). Object semantics are thus necessary and localized but *not sufficient for the model's initial reasoning state*, which draws mainly on the broader context.

**Depth-specific division of labor.** The layer-window results give a depth-resolved counterpart to Circuit ②, whose logit-lens tracing places the emergence of semantically decodable representations in the mid-to-late layers. Contextual masking is most harmful in the early-to-mid windows, whereas object-centric masking becomes relatively more harmful in the mid-to-late windows. This ordering is consistent with a two-stage computation: early-to-mid layers integrate contextual cues into a coarse reasoning state, and mid-to-late layers incorporate fine-grained object evidence as its representation becomes language-aligned, shaping the answer distribution.

The attention-flow analysis thus links the *spatial localization* of object evidence (Circuit ①) and the *depth-wise emergence* of language-aligned semantics (Circuit ②) to a routing process that determines which evidence is accessed at which depth during generation.

Table 10: **Accuracy drop** for `LLaVA-NeXT` models across various layer windows after attention knockout. Masking conditions: **O**, object tokens; **O+1/2**, object with 1/2 buffer; **C**, all tokens except the object; **C+1/2**, all tokens except the object with 1/2 buffer.

| Model | Condition (From*) | Early | Early-Mid | Mid | Mid-Late | Late | All Layers |
|---|---|---|---|---|---|---|---|
| *Object-centric masking* | | | | | | | |
| LLaVA-NeXT-I | O | 0.19 | 0.20 | 0.24 | **0.25** | 0.23 | 0.29 |
| | O+1 | 0.20 | 0.23 | 0.25 | **0.28** | 0.27 | 0.35 |
| | O+2 | 0.23 | 0.23 | 0.26 | **0.28** | 0.26 | 0.39 |
| LLaVA-NeXT-V | O | 0.25 | 0.26 | 0.27 | **0.30** | 0.27 | 0.32 |
| | O+1 | 0.25 | 0.26 | 0.29 | **0.33** | 0.29 | 0.38 |
| | O+2 | 0.25 | 0.26 | 0.31 | **0.33** | 0.32 | 0.42 |
| *Contextual masking* | | | | | | | |
| LLaVA-NeXT-I | C | 0.82 | **0.99** | 0.89 | 0.66 | 0.44 | 0.99 |
| | C+1 | 0.83 | **0.97** | 0.86 | 0.59 | 0.39 | 0.97 |
| | C+2 | 0.83 | **0.97** | 0.86 | 0.54 | 0.39 | 0.97 |
| LLaVA-NeXT-V | C | 0.49 | **0.91** | 0.89 | 0.92 | 0.62 | 0.97 |
| | C+1 | 0.47 | **0.90** | 0.86 | 0.88 | 0.56 | 0.94 |
| | C+2 | 0.46 | **0.90** | 0.83 | 0.87 | 0.52 | 0.92 |

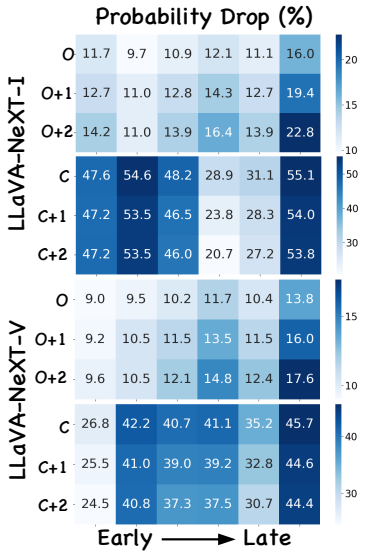


Figure 9: **Answer probability drop** under attention knockout across various layers.

## C.9 Interactive Examples

Fig. 10, Fig. 11, and Fig. 12 give qualitative examples of semantic tracing (Circuit ②) on questions about objects and actions. Each shows the input video frames, the question, the model's response, and a table of the decoded semantics of key tokens across layers. Further examples are available on the anonymous interactive demo website.

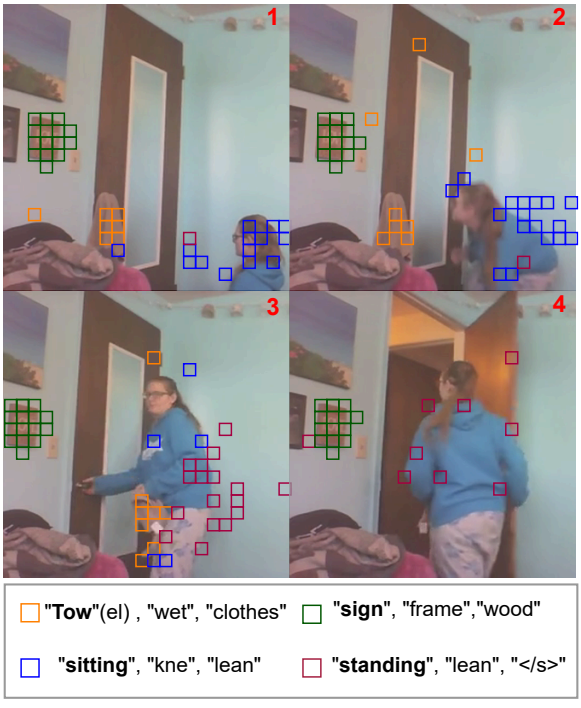


Figure 10: Qualitative example of semantic tracing illustrating how the model captures *temporal dynamics*. Panels 1 to 4 are consecutive frames; each colored box marks a token position whose top-3 most frequent decoded word groups are listed in the legend.



***Prompt***
USER: The input consists of a sequence of key frames from a video. Question: Which object was taken by the person?
ASSISTANT: The object is the

***Answer*:** Towel

***Response:*** Towel

| Token/Layer | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 |
|---|---|---|---|---|---|
| <s> | Institution | Архив | sierp | sierp | sierp |
| US | ona | Архив | pert | tap | pert |
| ER | Session | nyra | oded | 庄 | pel |
| : | Portail | nyra | <s> | <s> | Sav |
| | | nyra | Carter | 🙂 | 🙂 |
| <IMG001> | sak | sak | sak | sak | sak |
| <IMG002> | gresql | gresql | gresql | gresql | gresql |
| <IMG003> | olas | olas | olas | olas | olas |
| <IMG004> | << | << | << | << | << |
| <IMG005> | gresql | gresql | gresql | gresql | gresql |
| <IMG006> | sak | sak | sak | sak | sak |
| <IMG007> | yter | yter | yter | yter | yter |
| <IMG008> | nahm | nahm | nahm | nahm | nahm |
| <IMG009> | loop | loop | loop | loop | sak |
| <IMG010> | loop | loop | loop | loop | loop |
| <IMG011> | gresql | gresql | gresql | gresql | gresql |
| <IMG012> | yter | yter | yter | yter | yter |
| <IMG013> | izi | izi | Emp | Emp | conf |
| <IMG014> | alle | alle | alle | alle | conf |
| <IMG015> | yter | yter | yter | yter | yter |
| <IMG016> | alle | alle | alle | arab | anno |
| <IMG017> | yter | yter | yter | yter | yter |

Figure 11: Qualitative example of object identification. The user asks which object was taken by the person; the model answers "Towel", which is correct. The accompanying table shows the layer-by-layer semantic tracing for visual and text tokens.



***Prompt***
USER: The input consists of a sequence of key frames from a video. Question: Which object was thrown by the person?
ASSISTANT: The object is the

***Answer*:** Clothes

***Response:*** Green shirt that the person is throwing.

| Token/Layer | Layer 1 | Layer 2 | Layer 3 | Layer 4 | Layer 5 |
|---|---|---|---|---|---|
| <s> | Institution | Архив | sierp | sierp | sierp |
| US | ona | Архив | pert | tap | pert |
| ER | Session | nyra | oded | 庄 | pel |
| : | Portail | nyra | <s> | <s> | Sav |
| | | nyra | Carter | 🙂 | 🙂 |
| <IMG001> | yter | yter | yter | yter | yter |
| <IMG002> | yter | yter | yter | yter | yter |
| <IMG003> | middle | middle | middle | middle | middle |
| <IMG004> | yter | yter | yter | yter | yter |
| <IMG005> | yter | yter | yter | yter | yter |
| <IMG006> | kwiet | kwiet | kwiet | kwiet | kwiet |
| <IMG007> | onymes | onymes | onymes | onymes | stycz |
| <IMG008> | dek | dek | dek | dek | dek |
| <IMG009> | yter | yter | yter | yter | yter |
| <IMG010> | assen | assen | assen | assen | assen |
| <IMG011> | wheel | wheel | wheel | wheel | wheel |
| <IMG012> | assen | assen | assen | emann | emann |
| <IMG013> | alet | alet | alet | empio | empio |
| <IMG014> | gresql | gresql | gresql | gresql | gresql |
| <IMG015> | alet | alet | alet | alet | alet |
| <IMG016> | azon | azon | azon | azon | azon |
| <IMG017> | fle | fle | fle | fle | wi |

Figure 12: Qualitative example of identifying a thrown object. The model answers "Green shirt", identifying both the object and its color. The table shows the semantic trace of the visual information across layers.

# D Surgery Details and Setup

## D.1 Threshold Setting

This section gives the thresholding protocol used to identify *temporal heads* (the four criteria in §4.1) and the hyperparameters used for head amplification. All head-selection thresholds and the amplification factor $\lambda$ are calibrated on TempCompass, our development benchmark (referred to as the development split below). We hold these parameters fixed when evaluating the layer windows reported in Table 3 and Table 2. Results on the remaining benchmarks therefore measure transfer under each reported window.

**Threshold semantics.** The four thresholds $(\tau_m, \tau_{\text{eff}}, \tau_\sigma, \tau_{\text{occ}})$ carry distinct semantics and are *not directly comparable as raw numbers*:

- $\tau_m \in [0, 1]$ is an *absolute* lower bound on the video attention ratio $\bar{m}^{(\ell,h)}$.
- $\tau_{\text{eff}} \in [1, T]$ is an *absolute* lower bound on the effective number of slices $\overline{n_{\text{eff}}}^{(\ell,h)}$.

- $\tau_\sigma \in [0,1]$ is a *percentile-rank* threshold on temporal dispersion. Letting $\mathcal{H}$ denote the full set of attention heads, we compute the empirical percentile rank of $\bar{\sigma}_\tau^{(\ell,h)}$ on the development split,

$$\rho_\sigma^{(\ell,h)} \triangleq \frac{1}{|\mathcal{H}|}\left|\{(\ell',h') \in \mathcal{H} : \bar{\sigma}_\tau^{(\ell',h')} \le \bar{\sigma}_\tau^{(\ell,h)}\}\right| \in [0,1], \tag{8}$$

  and retain a head only if $\rho_\sigma^{(\ell,h)} \ge \tau_\sigma$. The criterion $\bar{\sigma}_\tau^{(\ell,h)} \ge \tau_\sigma$ in §4.1 should therefore be read as shorthand for this percentile-rank cut: by Popoviciu's inequality, $\bar{\sigma}_\tau \le 1/2$ for any distribution supported on $[0,1]$, so an absolute threshold with $\tau_\sigma > 0.5$ would have no feasible solution. Working with $\rho_\sigma$ instead yields a scale-free selection criterion.
- $\tau_{\text{occ}} \in [0,1]$ is the minimum sample-wise occurrence rate at which the joint criterion must hold for a head to be retained.

**Selected values.** We use percentile-based calibration to balance the number of selected heads against downstream gains under surgery. Across the explored configurations, the best-performing setting on the development split is $\tau_m = 0.1$, $\tau_{\text{eff}} = 2.0$, $\tau_\sigma = 0.8$ (equivalently, the top $20\%$ of heads by $\bar{\sigma}_\tau$), and $\tau_{\text{occ}} = 0.6$. The amplification factor is fixed to $\lambda = 1.75$ for all reported interventions, selected via the sensitivity sweep in Fig. 5 on TempCompass, where each point is the accuracy averaged over Layers 15–20, 20–25, and 25–30, and applied unchanged to all other benchmarks.

**Selection statistics and per-window head counts.** The head statistics $\bar{m}^{(\ell,h)}$, $\overline{n_{\text{eff}}}^{(\ell,h)}$, and $\bar{\sigma}_\tau^{(\ell,h)}$ are sample means over the STAR probe set (300 samples, $T$=32 frames), computed at the answer position $q$ with $\varepsilon$=$10^{-6}$ in Eq. (3) to Eq. (5). The thresholds above are applied to all 784 heads of each backbone (28 layers $\times$ 28 heads in the Qwen2.5-7B and Qwen2-7B language models underlying Qwen2.5-VL-7B and LLaVA-OneVision-7B, respectively). Table 11 gives the resulting set sizes and their layer distribution. The same set is used at both frame counts, and amplifying a window scales only the selected heads whose layer falls in that window, so the per-window counts are the numbers of heads scaled in each row of Table 3. LLaVA-OV-I has no selected head in Layers 0–5, so that row is a no-op; its $+0.2$ change at $T$=16 is numerical noise (Appendix D.9).

Table 11: Temporal-head selection details. Statistics are collected at the answer position on the STAR probe set (300 samples, $T$=32) with $\varepsilon$=$10^{-6}$ and the thresholds of this section. The per-window columns count the selected heads whose layer falls in each amplification window of Table 3; they sum to $|\mathcal{S}_{\text{temp}}|$. The census of Table 21 uses the same probe set and thresholds, so its counts coincide with $|\mathcal{S}_{\text{temp}}|$.

| Backbone | Total heads | $\lvert\mathcal{S}_{\text{temp}}\rvert$ | 0–5 | 5–10 | 10–15 | 15–20 | 20–25 | 25–end |
|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL | 784 | 124 | 3 | 5 | 8 | 20 | 56 | 32 |
| LLaVA-OV-V | 784 | 106 | 2 | 3 | 6 | 16 | 49 | 30 |
| LLaVA-OV-I | 784 | 48 | 0 | 1 | 2 | 6 | 24 | 15 |

### D.2 Evaluation Protocol

All video benchmarks use $T$ uniformly sampled frames ($T = 32$ unless stated), a smaller frame budget than the official Qwen2.5-VL evaluation, so that attention statistics, head selection, and intervention runs share one input length; the remaining inference settings are unchanged from the model defaults. Baseline accuracies are therefore below the values in the Qwen2.5-VL technical report. Every intervention condition is scored under the same configuration as its baseline, including the single-image controls MMBench and SEED-Bench, and the reported changes are paired differences under this one protocol (Appendix D.9). Table 12 lists, for each benchmark, the split and the number of questions $N$ that enter the paired tests, the subtitle setting, the frame count, the pixel budget, the prompt format, and the answer-parsing rule together with the fraction of outputs it fails to parse. Parsing failures are scored as wrong answers and stay below $0.5\%$ on every benchmark, which bounds their possible contribution to any reported change.

Appendix D.8 repeats the baseline and the Layers 25–30 amplification on VideoMME and EgoSchema under the official Qwen2.5-VL video-evaluation budget (2 fps sampling, up to 768 frames): the baselines then come within 1.6 points of the technical report, and the amplification gains remain significant.

Table 12: Evaluation protocol per benchmark under the 32-frame setting of Table 2 and Table 15. Video benchmarks use $T$ uniformly sampled frames at the default per-frame budget of 602,112 pixels; the single-image benchmarks use the default image budget. Subtitles: none = available but not used; n/a = the benchmark provides none. Multiple-choice prompts ask for the option letter only, and the first standalone option letter in the output is scored; VideoHallucer uses yes/no prompts scored by keyword. Outputs that yield no answer are scored as wrong and counted in the last column. The comparison with the technical report under the official frame budget is given in Appendix D.8.

| Benchmark | Split | $N$ | Subtitles | $T$ | max_pixels | Prompt | Parsing | Parse fail. |
|---|---|---|---|---|---|---|---|---|
| TempCompass | multiple-choice QA | 1,580 | n/a | 32/16 | 602,112 | option letter | first letter | 0.3% |
| NExTQA | val, multiple-choice | 4,996 | n/a | 32 | 602,112 | option letter | first letter | 0.2% |
| IntentQA | test | 2,134 | n/a | 32 | 602,112 | option letter | first letter | 0.1% |
| EgoSchema | public 500-question subset | 500 | n/a | 32 | 602,112 | option letter | first letter | 0.0% |
| VideoMME | full (short/medium/long) | 2,700 | none | 32 | 602,112 | option letter | first letter | 0.4% |
| VidHalluc | full, multiple-choice | 2,000 | n/a | 32 | 602,112 | option letter | first letter | 0.5% |
| VideoHallucer | full (basic + hallucinated pairs) | 1,800 | n/a | 32 | 602,112 | yes/no | yes/no keyword | 0.1% |
| MMBench | dev-en (v1.0) | 4,329 | n/a | image | 12,845,056 | option letter | first letter | 0.2% |
| SEED-Bench | image (dimensions 1–9) | 14,233 | n/a | image | 12,845,056 | option letter | first letter | 0.3% |

### D.3 Metric Ablation

The head-type ablation in Table 13 shows that partial criteria (e.g., using only $(m, n_{\text{eff}})$ or $(m, \sigma_\tau)$) do not reproduce the full gains, and amplifying a random head set does not help. This indicates that the improvement is not explained by a generic increase in attention capacity; rather, it depends on jointly satisfying (i) *video gating* and (ii) *broad, long-range temporal support*. To test the temporal criteria against single-statistic alternatives, we additionally amplify two count-matched head sets: the top-$|\mathcal{S}_{\text{temp}}|$ heads ranked by video attention ratio $m$ alone (*Top video-attention*), and the top heads ranked by frame-attention entropy alone (*Top-entropy*). These two sets and the random set are drawn from all heads in the window without excluding $\mathcal{S}_{\text{temp}}$, so each shares some heads with it; the random controls of Appendix D.6 and Appendix D.7 are instead disjoint from $\mathcal{S}_{\text{temp}}$. The two controls change accuracy by at most 0.3 points ($+0.3/-0.1$ and $+0.1/-0.2$ at 16/32 frames, respectively) and the random set by at most 0.4, whereas the full three-criterion set gains $+2.4/+0.8$. High video attention or broad attention dispersion alone therefore does not reproduce the gains; the temporal-span criterion is necessary. Among partial criteria, $(m, \sigma_\tau)$ performs best, so $\sigma_\tau$ is the most informative single statistic, though insufficient on its own.

Table 13: Ablation of the head-selection criteria on TempCompass (Qwen2.5-VL, amplification in Layers 20–25 with $\lambda$=1.75): partial criteria, a count-matched random set, and count-matched sets ranked by video-attention ratio or frame-attention entropy alone, against the full three-criterion set.

| Head Type | Frames=16 | Frames=32 |
|---|---|---|
| Baseline | 73.6 | 74.6 |
| (mass, $n_{\text{eff}}$) | 73.6 | 74.2 |
| (mass, $\sigma$) | 73.7 | 74.8 |
| ($n_{\text{eff}}$, $\sigma$) | 73.4 | 74.1 |
| Random | 73.2 | 74.2 |
| Top video-attention | 73.9 | 74.5 |
| Top-entropy | 73.7 | 74.4 |
| Temporal head | **76.0** | **75.4** |

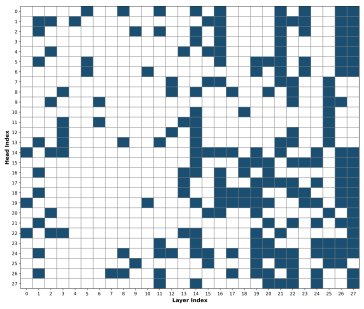


Figure 13: Layer head distribution heatmap.

### D.4 Qualitative Examples

Fig. 14 and Fig. 15 show key video frames, the input question, and the model's responses under two conditions: the *Baseline* (no intervention) and the *Intervened* model (temporal head amplification). In both cases the baseline answer is wrong about motion direction or event order and the intervened answer is correct.



**Question:** What is the direction of the camera motion relative to the aircraft carrier?

**Options:** A. counterclockwise B. downward C. upward D. clockwise

**Correct Answer: A**

**Qwen2.5VL:** The camera appears to be moving in a clockwise direction around the aircraft carrier, as we can see the carrier from different angles and perspectives throughout the video. Therefore, **the answer is D.**

**Qwen2.5VL+ Attention Intervention:** **The answer is A.** The aircraft carrier is moving in a counterclockwise direction. The aerial view shows the aircraft carrier from above, and the movement of the carrier can be observed by the changing position of the aircraft on the deck. The aircraft carrier is moving in a counterclockwise direction as it is shown

Figure 14: **Qualitative comparison on Visual Dynamics (Camera Motion).** The question asks for the direction of camera motion relative to the aircraft carrier. The **Baseline** model answers *clockwise*; the **Intervened** model answers *counterclockwise* (Option A), which matches the video.



Question: What is the order of events in the video?

**Options:**

A. An apple is drying out and then a big wave is shown

B. A big wave is shown and then an apple is drying out

C. At the same time, an apple is drying out and a big wave is shown

**Correct Answer: C**

**Qwen2.5VL:** **The correct answer is:\n\nA.** An apple is drying out and then a big wave is shown\n\nThe video shows a sequence where first, an apple is depicted in various stages of drying out, and then it transitions to showing a large wave.

**Qwen2.5VL**+ Attention Intervention: **The correct answer is:\n\nC.** At the same time, an apple is drying out and a big wave is shown\n\nThe video shows a half of an apple that appears to be drying out, with the background showing a large wave.

Figure 15: **Qualitative comparison on the Sequence Task (Event Ordering).** The question probes the chronological order of two visual events: an apple drying out and a wave. The **Baseline** misses the split-screen simultaneity and describes a sequential transition (*"...and then a big wave is shown"*); the **Intervened** model answers *"At the same time"* (Option C), which is correct.

### D.5 Random-Head Amplification Baseline

To separate the effect of the selected heads from a generic increase in attention capacity, we amplify a randomly selected, count-matched subset of heads at each candidate layer window. Table 14 reports the result on VideoMME and EgoSchema. Random amplification never improves over the baseline and lowers accuracy by 1.4 to 2.6 points in the Layers 15–20 and 25–30 windows, whereas the best temporal-head window reaches 45.2 on VideoMME and 59.4 on EgoSchema; the gains are tied to the selected head set.

Table 14: Random-head amplification control on VideoMME and EgoSchema.

| **Injection Layers** | **VideoMME** | **EgoSchema** |
|---|---|---|
| *Baseline* | 42.2 | 55.4 |
| Random heads, Layers 15–20 | 40.8 | 53.2 |
| Random heads, Layers 20–25 | 42.2 | 55.2 |
| Random heads, Layers 25–30 | 39.6 | 53.6 |
| *Temporal heads (max)* | **45.2** | **59.4** |

### D.6 Bidirectional Intervention: Necessity and Task Selectivity

Amplification alone establishes only that strengthening the identified heads is *sufficient* to improve temporal reasoning. To test whether these heads are also *necessary*, we apply the complementary negative intervention: scaling the outputs of the same head set $\mathcal{S}_{\text{temp}}$ within Layers 20–25 with $\lambda = 0.5$ (suppression) and $\lambda = 0$ (full ablation) on Qwen2.5-VL with 32-frame inputs. This window achieves the highest amplification accuracy on TempCompass in this setting. We evaluate on two temporal benchmarks (TempCompass, NExTQA) and two non-temporal, image-only benchmarks (MMBench, SEED-Bench), and include a count-matched control that ablates a random subset of heads drawn from those exceeding the video-attention threshold $\tau_m$ but *not* selected into $\mathcal{S}_{\text{temp}}$.

Table 15: Bidirectional intervention on $\mathcal{S}_{\text{temp}}$ (Qwen2.5-VL, 32 frames, Layers 20–25). Parentheses denote absolute change from baseline; $^\dagger$ marks changes significant under the paired exact McNemar protocol of Appendix D.9. Suppression produces dose-dependent, task-selective degradation: temporal benchmarks drop by up to 6.3 points under full ablation, while non-temporal benchmarks change by at most 0.9 points and a count-matched random video-head ablation costs only 0.8 points on TempCompass.

| | | **Temporal** | | **Non-temporal** | |
|---|---|---|---|---|---|
| **Intervention** | $\lambda$ | TempCompass | NExTQA | MMBench | SEED-Bench |
| *Baseline (none)* | 1.0 | 74.6 | 71.0 | 80.1 | 76.3 |
| Amplify $\mathcal{S}_{\text{temp}}$ | 1.75 | **75.4** (+0.8) | 71.0 (+0.0) | 80.0 (−0.1) | 76.2 (−0.1) |
| Suppress $\mathcal{S}_{\text{temp}}$ | 0.5 | 71.9 (−2.7)$^\dagger$ | 68.8 (−2.2)$^\dagger$ | 79.8 (−0.3) | 76.0 (−0.3) |
| Ablate $\mathcal{S}_{\text{temp}}$ | 0 | **68.3** (−6.3)$^\dagger$ | **65.9** (−5.1)$^\dagger$ | 79.2 (−0.9) | 75.6 (−0.7) |
| Ablate random video heads (matched) | 0 | 73.8 (−0.8) | 70.4 (−0.6) | 79.5 (−0.6) | 75.9 (−0.4) |

Three observations follow from Table 15. **(i) Dose–response necessity.** Degradation scales monotonically with intervention strength: TempCompass drops 2.7 points at $\lambda$=0.5 and 6.3 points at $\lambda$=0 (NExTQA: 2.2 and 5.1), so performance on the evaluated temporal benchmarks depends functionally on the identified heads. **(ii) Task selectivity.** Full ablation costs 6.3 points on TempCompass and 0.9 on MMBench, so the causal contribution of $\mathcal{S}_{\text{temp}}$ is concentrated on the multi-frame temporal benchmarks rather than on single-image benchmarks. These benchmark groups differ in modality and input length as well as in temporal demand, so this contrast establishes task selectivity; separating temporal-order computation from general multi-frame integration requires same-video controls that vary only the question type, which this paper does not include. **(iii) Head-set specificity.** Ablating a count-matched random set of video-attending heads costs 0.8 points on TempCompass, against 6.3 for $\mathcal{S}_{\text{temp}}$, so the loss is specific to the selected heads rather than to removing video-attending heads in general. The asymmetry between the modest amplification gain (+0.8) and the large ablation loss (−6.3) is consistent with a near-saturated mechanism: the temporal heads already operate close to their effective regime in the unmodified model, leaving limited headroom for amplification, while their removal substantially impairs cross-frame evidence routing.

### D.7 Order Sensitivity under Head Ablation

The frame-order perturbations of Appendix C.7 show that the outputs depend on frame order, and the ablations of Appendix D.6 show that the temporal benchmarks depend on $\mathcal{S}_{\text{temp}}$. This experiment links the two: if $\mathcal{S}_{\text{temp}}$ carries the order-dependent component of cross-frame routing, ablating it

should reduce the sensitivity to frame reversal, whereas ablating a count-matched random set of video-attending heads should not. We evaluate `Qwen2.5-VL` with 32 frames on the full STAR analysis pool ($N$=1720; Interaction 1090, Sequence 630) under three head conditions and three frame orders (original, reversed, shuffled). The head conditions are no intervention; $\mathcal{S}_{\text{temp}}$ ablated with $\lambda$=0 in Layers 20–25, the configuration of Table 15; and the count-matched random video-head set of Table 15 ablated in the same way. The full pool is used rather than the correct-only subset of Appendix C.7, because that subset is defined by the unperturbed prediction under $\lambda$=1 and would change with the head condition. For each head condition we define the reversal sensitivity $\text{Sens} = \text{Acc}_{\text{orig}} - \text{Acc}_{\text{rev}}$ and its relative form $\text{RelSens} = \text{Sens}/\text{Acc}_{\text{orig}}$, which guards against floor effects when an ablation lowers the original-order accuracy.

Table 16: Accuracy (%) under frame-order perturbation and head ablation (`Qwen2.5-VL`, 32 frames, Layers 20–25, full STAR analysis pool, $N$=1720; Avg is weighted by the number of questions per task). Drop is the decrease from the original order within the same head condition, computed from unrounded accuracies; RelSens is Drop divided by the original-order accuracy. Ablating $\mathcal{S}_{\text{temp}}$ reduces the reversal drop from 6.7 to 2.2 points, whereas ablating the count-matched random set leaves it at 6.3; the difference between the two changes in sensitivity is $-4.1$ points (paired bootstrap 95% CI $[-6.3, -1.9]$).

| Head set | $\lambda$ | Order | Interaction | Sequence | Avg | Drop | RelSens |
|---|---|---|---|---|---|---|---|
| None | 1 | original | 38.6 | 33.5 | 36.7 | – | – |
| None | 1 | reverse | 32.1 | 26.5 | 30.0 | 6.7 | 0.18 |
| None | 1 | shuffle | 34.2 | 29.8 | 32.6 | 4.1 | 0.11 |
| $\mathcal{S}_{\text{temp}}$ | 0 | original | 33.0 | 27.1 | 30.8 | – | – |
| $\mathcal{S}_{\text{temp}}$ | 0 | reverse | 30.6 | 25.2 | 28.6 | 2.2 | 0.07 |
| $\mathcal{S}_{\text{temp}}$ | 0 | shuffle | 31.5 | 26.3 | 29.6 | 1.2 | 0.04 |
| Random video heads (matched) | 0 | original | 37.9 | 32.6 | 36.0 | – | – |
| Random video heads (matched) | 0 | reverse | 31.8 | 25.9 | 29.6 | 6.3 | 0.18 |
| Random video heads (matched) | 0 | shuffle | 33.7 | 29.0 | 32.0 | 4.0 | 0.11 |

Without intervention, reversal lowers accuracy by 6.7 points (RelSens=0.18) and shuffling by 4.1 (Table 16). Ablating $\mathcal{S}_{\text{temp}}$ lowers the original-order accuracy by 5.9 points, far from the floor, and reduces the reversal drop to 2.2 (RelSens=0.07) and the shuffle drop to 1.2. Ablating the matched random set costs under one point in the original order (36.0 vs. 36.7) and leaves both drops unchanged (6.3 and 4.0; RelSens=0.18 and 0.11). Relative to the unmodified model, the change in reversal sensitivity is $-4.5$ points $[-6.4, -2.6]$ for $\mathcal{S}_{\text{temp}}$ and $-0.4$ $[-2.2, +1.4]$ for the random set. Their difference, $-4.1$ $[-6.3, -1.9]$, excludes zero (paired bootstrap over samples). The order-dependent part of the model's behavior therefore runs through $\mathcal{S}_{\text{temp}}$ rather than through video-attending heads in general. This characterizes $\mathcal{S}_{\text{temp}}$ as order-sensitive cross-frame routing, not as a computation of temporal relations: Interaction questions, whose answers do not depend on event order, lose about as much under reversal as Sequence questions in every head condition, so the sensitivity removed by the ablation is a position-dependent aggregation of frame evidence, the pattern already observed in Appendix C.7. Same-video controls that vary only the question type would be needed to separate the two (Appendix D.6).

### D.8 Amplification under the Official Frame Budget

Table 17: Temporal-head amplification in Layers 25–30 under the official Qwen2.5-VL frame budget (2 fps, up to 768 frames) on VideoMME (without subtitles, $N$=2700) and EgoSchema (public subset, $N$=500). $\Delta = (n_{10} - n_{01})/N$; the 95% CI is the paired-binomial normal approximation; $p_{\text{holm}}$ comes from exact McNemar tests corrected within each two-test family (Family P: $\mathcal{S}_{32}$; Family S: $\mathcal{S}_{\text{off}}$). Under the 32-frame protocol the same window gives $+2.6$ $[+1.0, +4.2]$ on VideoMME and $+4.0$ $[+2.1, +5.9]$ on EgoSchema (Table 18). $^{\dagger}$: $p_{\text{holm}} < 0.05$.

| Benchmark | Head set | Baseline | Amplified | $\Delta$ | 95% CI | $n_{10}/n_{01}$ | $p_{\text{holm}}$ |
|---|---|---|---|---|---|---|---|
| VideoMME | $\mathcal{S}_{32}$ | 63.6 | 65.6 | $+2.0$ | $[+0.4, +3.6]$ | 270/216 | $0.019^{\dagger}$ |
| EgoSchema | $\mathcal{S}_{32}$ | 63.4 | 66.2 | $+2.8$ | $[+0.8, +4.8]$ | 20/6 | $0.019^{\dagger}$ |
| VideoMME | $\mathcal{S}_{\text{off}}$ | 63.6 | 65.7 | $+2.1$ | $[+0.5, +3.7]$ | 272/214 | $0.012^{\dagger}$ |
| EgoSchema | $\mathcal{S}_{\text{off}}$ | 63.4 | 66.4 | $+3.0$ | $[+1.0, +5.0]$ | 21/6 | $0.012^{\dagger}$ |

The 32-frame protocol of Appendix D.2 keeps one input length for attention statistics, head selection, and intervention, at the cost of baselines below the Qwen2.5-VL technical report (Bai et al., 2025). To test whether the amplification effect depends on this reduced input, we repeat the baseline and the Layers 25–30 amplification on VideoMME and EgoSchema under the official Qwen2.5-VL video-evaluation configuration: frames sampled at 2 fps up to 768 frames (360 frames for the 180-second EgoSchema clips), a per-frame budget of 602,112 pixels, and a total budget of 19,267,584 pixels per video, which lowers the per-frame resolution as the frame count grows. Prompt template, answer parsing, $\lambda$=1.75, and the benchmark splits are those of Table 2. Two head sets are amplified: $\mathcal{S}_{32}$, the set of the main paper selected at $T$=32, and $\mathcal{S}_{\text{off}}$, re-selected from attention statistics collected under the official budget on the same STAR probe set with unchanged thresholds ($|\mathcal{S}_{32}|$=124, $|\mathcal{S}_{\text{off}}|$=132, $|\mathcal{S}_{32} \cap \mathcal{S}_{\text{off}}|$=107, Jaccard 0.72). Amplifying both sets tests whether the result depends on which selection is used. The two $\mathcal{S}_{32}$ cells and the two $\mathcal{S}_{\text{off}}$ cells form two families, each Holm-corrected over its two exact McNemar tests (Appendix D.9).

Under the official budget the baselines are 63.6 on VideoMME and 63.4 on EgoSchema, 1.5 and 1.6 points below the 65.1 and 65.0 of the technical report. The gap between Table 2 and the report is therefore mostly due to the frame budget, with a residual of at most 1.6 points. Amplifying $\mathcal{S}_{32}$ in Layers 25–30 gains +2.0 on VideoMME and +2.8 on EgoSchema, both significant after Holm correction, and $\mathcal{S}_{\text{off}}$ gains +2.1 and +3.0 (Table 17). The gains are smaller than under the 32-frame protocol (+2.6 and +4.0) but remain significant, so the effect is not specific to a reduced input. The head set re-selected under the official budget overlaps the original set at Jaccard 0.72, comparable to the overlap across frame counts in Table 23 (0.74), and its gains lie within 0.2 points of those of $\mathcal{S}_{32}$. Only the Layers 25–30 window was run under the official budget; the per-benchmark peak windows discussed in §4 refer to the 32-frame protocol.

Table 18: Accuracy change (absolute %) of temporal-head amplification with 95% paired-bootstrap confidence intervals and Holm-corrected exact McNemar $p$-values (Qwen2.5-VL, 32 frames). $^{\dagger}$: $p_{\text{holm}} < 0.05$; $^{\ddagger}$: $p_{\text{holm}} < 0.01$, corrected over all 12 window–benchmark tests in this table; the bootstrap intervals are not adjusted for multiple comparisons.

| Benchmark | Window | Δ | 95% CI | $p_{\text{holm}}$ |
|---|---|---|---|---|
| EgoSchema | Layers 25–30 | +4.0 | [+2.1, +5.9] | $< 0.001^{\ddagger}$ |
| VideoMME | Layers 20–25 | +3.0 | [+1.4, +4.6] | $0.003^{\ddagger}$ |
| VideoMME | Layers 25–30 | +2.6 | [+1.0, +4.2] | $0.015^{\dagger}$ |
| IntentQA | Layers 25–30 | +1.4 | [+0.3, +2.5] | 0.12 |
| TempCompass | Layers 20–25 | +0.8 | [+0.1, +1.6] | 0.30 |
| TempCompass | Layers 15–20 | +0.2 | [−0.9, +1.3] | > 0.99 |
| TempCompass | Layers 25–30 | +0.4 | [−0.5, +1.3] | > 0.99 |
| EgoSchema | Layers 15–20 | −0.6 | [−1.8, +0.5] | > 0.99 |
| EgoSchema | Layers 20–25 | −1.2 | [−2.5, +0.1] | 0.54 |
| VideoMME | Layers 15–20 | +0.8 | [−0.4, +2.0] | 0.83 |
| IntentQA | Layers 15–20 | −0.2 | [−0.8, +0.4] | > 0.99 |
| IntentQA | Layers 20–25 | −0.2 | [−0.9, +0.5] | > 0.99 |

## D.9 Statistical Significance of Intervention Results

Table 19: Paired-test outcomes for the amplification cells not covered by Table 18 and for the suppression conditions of Table 15 (Qwen2.5-VL, 32 frames). Same paired exact McNemar protocol as Table 18, Holm-corrected within this family; Δ and corrected significance level are reported, without bootstrap intervals.

| Benchmark | Condition | Δ | $p_{\text{holm}}$ |
|---|---|---|---|
| NExTQA | Amplify, Layers 15–20 | +2.0 | $< 0.05^{\dagger}$ |
| NExTQA | Amplify, Layers 25–30 | +2.4 | $< 0.05^{\dagger}$ |
| VidHalluc | Amplify, Layers 15–20 | +3.6 | $< 0.05^{\dagger}$ |
| TempCompass | Suppress $\lambda$=0.5, Layers 20–25 | −2.7 | $< 0.05^{\dagger}$ |
| TempCompass | Ablate $\lambda$=0, Layers 20–25 | −6.3 | $< 0.05^{\dagger}$ |
| NExTQA | Suppress $\lambda$=0.5, Layers 20–25 | −2.2 | $< 0.05^{\dagger}$ |
| NExTQA | Ablate $\lambda$=0, Layers 20–25 | −5.1 | $< 0.05^{\dagger}$ |

All evaluations use greedy decoding with deterministic uniform frame sampling, so the dominant source of uncertainty is sample-level variation over the paired per-sample predictions, which is the error model used below. The remaining run-to-run variation comes from floating-point nondeterminism in GPU inference, which can flip a few near-tied predictions. The LLaVA-OV-I Layers 0–5 row of Table 3 indicates its size: no selected head lies in that window (Table 11), so amplification leaves the model unchanged, yet accuracy moves by 0.2 points at $T$=16. We treat changes of this magnitude as numerical noise. For each intervention condition that enters a test we therefore record, for every evaluation sample, the paired outcome $(\text{correct}_{\text{base}}, \text{correct}_{\text{interv}}) \in \{0,1\}^2$, and report: (i) the accuracy change Δ with a 95% confidence interval from 10,000 paired bootstrap resamples over samples (cluster-resampled by video when one video contributes multiple questions), and (ii) an exact McNemar test on the discordant pairs, i.e., samples *fixed* by the intervention ($n_{10}$: baseline wrong, intervention correct) versus samples *broken* by it ($n_{01}$: baseline correct, intervention wrong). Since $\Delta = (n_{10} - n_{01})/N$, the discordant counts determine the point estimate and show how many individual predictions the intervention changed. This procedure underlies every significance marker in

the paper. McNemar $p$-values for the amplification windows on EgoSchema, VideoMME, IntentQA, and TempCompass are Holm-corrected across the 12 window–benchmark tests in Table 18, which reports $\Delta$, the bootstrap interval, and the corrected $p$-value for each cell. The remaining amplification cells of Table 2 (NExTQA, VidHalluc, VideoHallucer) and the suppression conditions of Table 15 form a second family, tested with the same paired McNemar procedure and Holm-corrected within itself; Table 19 lists the cells of this family that reach $p_{\text{holm}} < 0.05$ with their $\Delta$. Bootstrap intervals were computed for the primary family only.

After Holm correction, the gains on EgoSchema (Layers 25–30) and VideoMME (Layers 20–25 and 25–30) remain significant, whereas the nominally significant TempCompass (Layers 20–25) and IntentQA (Layers 25–30) gains do not survive the correction, and changes of $+0.2$ and $+0.4$ are not distinguishable from zero.

**LLaVA-OV-V.** The per-sample predictions behind the `LLaVA-OV-V` columns of Table 3 were not stored, so we re-ran the four cells inside the consolidation interval (Layers 20–25 and 25–30 at $T \in \{16, 32\}$) on TempCompass ($N$=1580) and tested them with the same exact McNemar procedure, Holm-corrected over these four cells. The eight cells below layer 20 were not re-run and are not tested (their reported changes lie between $-0.8$ and $+0.4$). Table 20 lists the outcomes. Only Layers 25–30 at $T$=16 is significant ($+1.0$, $n_{10}/n_{01} = 19/3$, $p_{\text{holm}} = 0.003$). The same window at $T$=32 changes no net predictions ($16/16$), and neither Layers 20–25 cell is significant. TempCompass is the development benchmark on which thresholds and $\lambda$ were calibrated (Appendix D.1), so on `LLaVA-OV-V` the amplification effect is established for one in-interval cell and has no held-out counterpart.

Table 20: Paired-test outcomes for the re-run `LLaVA-OV-V` cells of Table 3 (TempCompass, $N$=1580; $\Delta = (n_{10} - n_{01})/N$, same format as Table 18). Exact McNemar on discordant pairs, Holm-corrected over the four in-interval cells; the eight windows below layer 20 were not tested. $^{\ddagger}$: $p_{\text{holm}} < 0.01$.

| Window | $T$ | $\Delta$ | $n_{10}/n_{01}$ | $p_{\text{holm}}$ |
|---|---|---|---|---|
| Layers 20–25 | 16 | 0.0 | 15/15 | $> 0.99$ |
| Layers 25–30 | 16 | $+1.0$ | 19/3 | $0.003^{\ddagger}$ |
| Layers 20–25 | 32 | $+0.4$ | 19/13 | $> 0.99$ |
| Layers 25–30 | 32 | 0.0 | 16/16 | $> 0.99$ |

### D.10 Temporal-Head Emergence: Video- vs. Image-Trained Models

The main text observes that amplification helps video-trained models but not the image-trained LLaVA-OV-I. To explain *why*, we profile all attention heads of each backbone with the same three temporal statistics $(m, n_{\text{eff}}, \sigma_\tau)$ and thresholds used for head selection, and count how many heads satisfy the full criteria.

Table 21: Census of heads passing the full temporal criteria across backbones (same probe set and thresholds as Table 11). Video-trained models contain 2.2–2.6$\times$ more qualifying heads than the image-trained variant. In all three backbones most qualifying heads lie in Layers 20–27, and the median head falls in the 20–25 window.

| Model | Training | #Heads passing | In Layers 20–27 |
|---|---|---|---|
| LLaVA-OV-V | video | 106 | 79 (75%) |
| Qwen2.5-VL | video | **124** | 88 (71%) |
| LLaVA-OV-I | image | 48 | 39 (81%) |

The video-trained backbones contain 106–124 heads satisfying the full temporal criteria, against 48 in the image-trained LLaVA-OV-I (Table 21). The layer distribution does not separate the backbones: in all three, 71–81% of the qualifying heads lie in Layers 20–27, the interval where Circuit ② localizes semantic consolidation, and fewer than 10% lie below layer 10. The census therefore separates video- from image-trained backbones by the number of qualifying heads, not by their location. This is consistent with the behavioral result in Table 3 but does not fully explain it: LLaVA-OV-I has 24 qualifying heads in Layers 20–25 and 15 in Layers 25–27, and amplifying either window lowers its TempCompass accuracy by 0.4 to 1.0 points, so the heads that pass the statistical criteria in the image-trained model do not carry the same function. A hypothesis consistent with this contrast is that temporally ordered training data increases the number of such heads and shapes what they route, rather than the heads being an architectural property of the LLM backbone. Because the two variants also differ in training stages and data mixtures, isolating this factor would require controlled training comparisons.

## E Additional Robustness Studies

### E.1 Stability Across Prompt Formats

To test sensitivity to prompt format, we re-run the analysis on `LLaVA-OneVision-V` with a reordered prompt template. The original format places the visual stream first ($\langle$video$\rangle$ + Question); the reordered template additionally prepends a system prompt (System Prompt + $\langle$video$\rangle$ + Question). The conclusions are unchanged. For Circuit ①, ablating object tokens still lowers accuracy from 100.0 to 58.9, and enlarging the object-centered region (buffer 1, buffer 2) lowers it further to 54.2 and 52.0, mirroring the main-paper pattern (Table 22). For Circuit ②, the layer-wise semantic-consolidation trend is qualitatively unchanged under the reordered prompt. The one constraint on the prompt is that the text preceding the video must *not* already expose the target semantics: under causal attention, video tokens can attend to preceding text tokens, and the traced signal would then reflect text leakage rather than the semantic evolution of the visual stream.

Table 22: Circuit ① ablation under the reordered prompt format on `LLaVA-OneVision-V`.

| Ablation Type | Accuracy |
|---|---|
| *Baseline* | 100.0 |
| Object | 58.9 |
| Object (buffer 1) | 54.2 |
| Object (buffer 2) | 52.0 |

### E.2 Stability of Temporal-Head Selection

$\mathcal{S}_{\text{temp}}$ is selected from attention statistics collected on a particular probe set, prompt template, and input configuration. We measure the Jaccard overlap $|A \cap B|/|A \cup B|$ between head sets selected independently under varied conditions on Qwen2.5-VL.

Across all four axes (probe-sample resampling, prompt reordering, frame count, and probing dataset) the overlap of the selected head sets remains at or above 0.70, with the highest consistency ($0.81 \pm 0.05$) under sample resampling. Together with the head-set controls in Appendix D.3 (random, top video-attention, and top-entropy sets do not reproduce the gains) and the suppression results in Appendix D.6, this identifies $\mathcal{S}_{\text{temp}}$ as a stable and functionally specific temporal routing structure.

Table 23: Jaccard overlap of $\mathcal{S}_{\text{temp}}$ under varied selection conditions. Overlap remains $\geq 0.70$ across all axes, so the selection is reproducible across probing configurations.

| Axis | Comparison | Jaccard |
|---|---|---|
| Probe samples | random half-splits of the probe set, selected independently ($\times 10$) | $0.81 \pm 0.05$ |
| Prompt template | original vs. reordered format with system prompt | 0.77 |
| Frame count | heads selected at $T$=16 vs. $T$=32 | 0.74 |
| Probing dataset | STAR probe set vs. TempCompass subset | 0.70 |

## F Broader Societal Impacts

This work analyzes the internal computations of LVLMs and proposes a training-free intervention, temporal-head amplification, that improves temporal reasoning on the evaluated benchmarks.

The mechanistic account (where object semantics are localized, when language-aligned concepts emerge, and which heads carry temporal evidence) helps practitioners and auditors understand why an LVLM produces a given answer and locate the depth and routing at which temporal or object-identity errors arise. Because the intervention needs no retraining, such corrections can be applied at low compute cost, including by groups without the resources to fine-tune large models.

The intervention reduces but does not eliminate temporal errors, so an improved benchmark score does not establish reliability for high-stakes video review, which still requires task-specific validation. Amplifying specific heads changes the model's routing and could strengthen biased associations inherited from training data; this effect is not measured here. Circuit-level knowledge of which heads carry temporal evidence could also inform adversarial perturbations against deployed LVLMs.