KNOWING WHERE IS NOT KNOWING WHAT: EVIDENCE-GROUNDING FAILURE IN SOCIAL GAZE

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding social gaze is essential for socially interactive agents, yet a look's function depends less on its direction than on the subsequent interaction. While multimodal large language models (MLLMs) are evaluated on gaze following, whether they ground functional judgments in this interaction remains largely unexplored. To address this gap, we present **EviGaze**, a benchmark specifically designed to diagnose evidence grounding in social gaze understanding. EviGaze annotates 5,120 gaze events from four social-video corpora with their function and evidence interval. Since this interval follows the gaze, removing it withdraws the grounds for judgment while keeping the look visible. We define evidence-grounding failure (EGF) as label retention after removal once the target and evidence are located. Extensive experiments are conducted to benchmark fifteen MLLMs on EviGaze. Experimental results reveal that EGF persists across all fifteen MLLMs, with evidence removal lowering model accuracy by only 16.0 percentage points against 64.4 for humans. Probing reveals that model states encode functional information that judgments underuse, a dissociation consistent across all seven analyzed models. Prompting models to describe the evidence before answering reduces the EGF rate beyond a matched placebo. The benchmark and code will be publicly released to support further research.

1 INTRODUCTION

A look across a table can request the salt, yield the floor, or check understanding. Its direction alone does not distinguish these functions; the interaction that follows determines what the look accomplishes (Kendon, 1967; Duncan, 1972; Sacks et al., 1974; Schegloff et al., 1977; Argyle & Cook, 1976; Kleinke, 1986; Clark & Brennan, 1991; Emery, 2000). Socially interactive agents must therefore interpret a look by its consequences, since the same glance can call for a reply, an object, or no action. Although multimodal large language models (MLLMs) are increasingly evaluated on gaze following and social gaze prediction (Wang et al., 2025a; 2026; Kim et al., 2026), it remains largely unexplored whether their functional judgments rest on the ensuing interaction. Judging gaze function is harder than locating a gaze target, since the deciding evidence arrives only after the look ends, often in another person's response. As shown in Figure 1, MLLMs can still commit to a function after this evidence is removed, inventing an exchange the edited video no longer contains.

Though gaze has been studied extensively in computer vision and MLLM evaluation, existing benchmarks suffer from three issues: (1) **Geometric scope:** gaze benchmarks annotate fixation targets, shared attention, social predicates, or typed events (Recasens et al., 2015; Chong et al., 2020; Fan et al., 2018; 2019; Gupta et al., 2024a; Tafasca et al., 2023; Marín-Jiménez et al., 2019; Hu et al., 2023; Wang et al., 2025a; 2026; Kim et al., 2026; Madan et al., 2025; Gupta et al., 2024b), so a correct answer cannot separate locating a look from understanding its function; (2) **Inseparable evidence:** social-video benchmarks mark scene properties and answer-supporting segments (Zadeh et al., 2019; Lei et al., 2020; Niu et al., 2025; Mathur et al., 2025; Gavryushin et al., 2026; Khebour et al., 2024), but leave the interval that decides a look's function unannotated and entangled with the look itself; (3) **Unverified grounding:** since MLLMs can answer video questions from priors and infer gaze from head orientation (Rahmanzadehgervi et al., 2024; Plizzari et al., 2025; Guo et al., 2023; Zhang et al., 2026), accuracy alone cannot show whether a correct judgment rests on the observed interaction. As a result, they cannot reveal whether a model that locates the right target also understands what the look accomplished, a competence socially interactive agents require.

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.

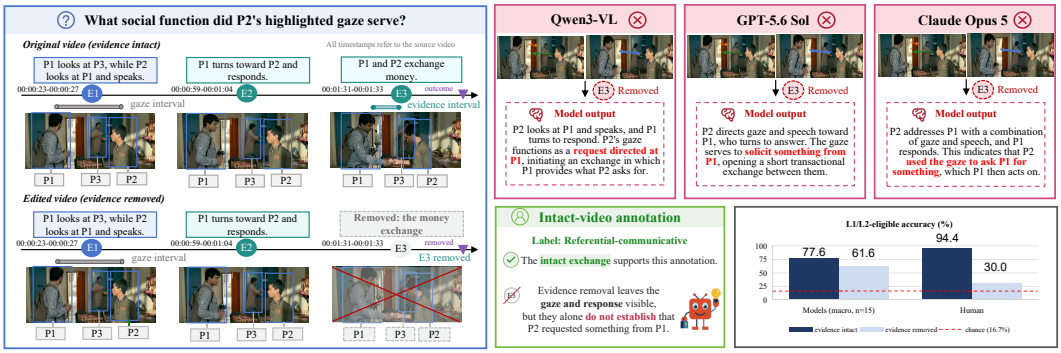

Figure 1: **A visible gaze with its deciding evidence removed.** The edit removes the money exchange while preserving the earlier gaze and response; timestamps refer to the source video. The intact-video annotation provides a reference, while red phrases highlight unsupported request inferences in the three post-removal outputs. Bars show six-way accuracy on intact-side L1/L2-eligible cases for fifteen models and a separate between-subjects human reference (Table 1); chance is 16.7%. $\otimes$ marks a label retained after its deciding evidence is removed.

To address these gaps, we introduce **EviGaze**, a novel benchmark for diagnosing evidence grounding in MLLM judgments of social gaze. Starting from $7{,}260$ candidates in four public social-video corpora (Fan et al., 2019; Chong et al., 2020; Fan et al., 2018; Gupta et al., 2024a), we retain $5{,}120$ events after screening out those with ambiguous functions, poor visibility, or insufficient temporal context. Three qualified annotators independently label each event with one of seven functions, from referential-communicative gaze to no observable uptake, each defined by an executable criterion. Their function labels reach substantial agreement (Fleiss' $\kappa = 0.76$) before a senior adjudicator resolves the remaining disagreements. Each event is queried for its gaze target, determining evidence, and function, separating functional errors from perceptual and localization errors.

Unlike previous benchmarks that annotate where a look lands or which segment supports an answer, EviGaze annotates each look's function jointly with the evidence interval that determines it. Since this interval begins only after the gaze interval has closed, its removal withdraws the grounds for a judgment while keeping the look visible. We term label retention after this removal evidence-grounding failure (EGF), computing its rate on events where a model identifies the target and localizes the evidence for the ground-truth target. Benchmarking fifteen MLLMs on EviGaze reveals pervasive EGF and a wide gap from humans, with evidence removal costing humans 64.4 percentage points of accuracy but models only 16.0, each on its own eligible set (Table 1). For current MLLMs, the judgment of a look's function is thus largely settled before the deciding interaction unfolds. Our three main contributions are summarized as follows:

- We introduce **EviGaze**, which pairs $5{,}120$ gaze events with function-defining evidence intervals, three diagnostic questions, and six controlled interventions.
- We benchmark fifteen MLLMs on EviGaze against human references, and all fifteen exhibit EGF on more than 60% of eligible events. Even naturally paired looks with different functions receive the same label from models in 34.7% of pairs.
- Our analysis favors evidence use over perception and localization as the primary bottleneck. In seven open-weight models, combined ground-truth target and interval hints leave 55.9% of baseline functional errors unrepaired, and linear probes recover gaze function from internal states that the final judgments only partly reflect. An evidence-focused prompt reduces the EGF rate beyond a structure-matched placebo.

2 Related Work

Social Video Understanding Benchmarks. Social-video benchmarks score the inferences a model draws about an interaction, from social situations and mental states to common ground and collaboration (Zadeh et al., 2019; Lei et al., 2020; Niu et al., 2025; Khebour et al., 2024; Gavryushin et al., 2026; Lee et al., 2024). Social Genome (Mathur et al., 2025) goes furthest toward grounding,

annotating reasoning traces that cite visual, verbal, and vocal evidence. Because answers are not re-queried without the cited evidence, however, such evaluation cannot expose EGF. In contrast, EviGaze records the interval licensing each look's function, so questions can be re-posed without it.

Gaze Following and Social Gaze. Gaze research has progressed from static and 3D gaze targets (Recasens et al., 2015; Hu et al., 2023) to video attention targets, shared attention, mutual gaze, and typed gaze-communication events (Chong et al., 2020; Fan et al., 2018; Marín-Jiménez et al., 2019; Tafasca et al., 2023; Fan et al., 2019), unified by multi-person temporal models (Gupta et al., 2024a). As these tasks move into MLLMs, VL4Gaze (Wang et al., 2025a) casts gaze following as visual question answering, while others use MLLMs as cue extractors (Gupta et al., 2024b; Madan et al., 2025). EyeVLM (Wang et al., 2026) evaluates gaze points and binary relations such as mutual gaze and shared attention, but on static images, where no subsequent interaction can reveal a look's function. GRASP (Kim et al., 2026) comes closest, composing identity-consistent gaze trajectories and deictic gestures into social events. Its questions, however, center on identifying the participants of each interaction rather than the role a look plays. In contrast, EviGaze separates target perception from functional judgment and defines each function by the interaction that follows the look.

Grounding Diagnostics for MLLMs. High accuracy need not reflect grounding in the visual input. MLLMs fail elementary visual tests (Rahmanzadehgervi et al., 2024; Xu et al., 2025), exploit answer biases in social and egocentric video QA (Guo et al., 2023; Plizzari et al., 2025), and mistake head orientation for gaze direction (Zhang et al., 2026). Paired or counterfactual inputs test whether answers change when their grounds do (Han et al., 2026; Sepehri et al., 2025). Probing (Alain & Bengio, 2016; Hewitt & Liang, 2019) reveals that information decodable from internal states can go unused by the final answer (Fu et al., 2025; Orgad et al., 2025). In contrast, EviGaze compares decodability, answer sensitivity, and human sensitivity under controlled removal of annotated evidence, using matched probe and judgment readouts and a between-subjects human reference.

3 The EviGaze Benchmark

3.1 Overview

In this section, we describe how EviGaze is constructed, from event collection and annotation to diagnostic tasks and interventions. EviGaze contains 5,120 gaze events from four social-video corpora, yielding 15,360 core questions and 24,700 intervention variants (Figure 2).

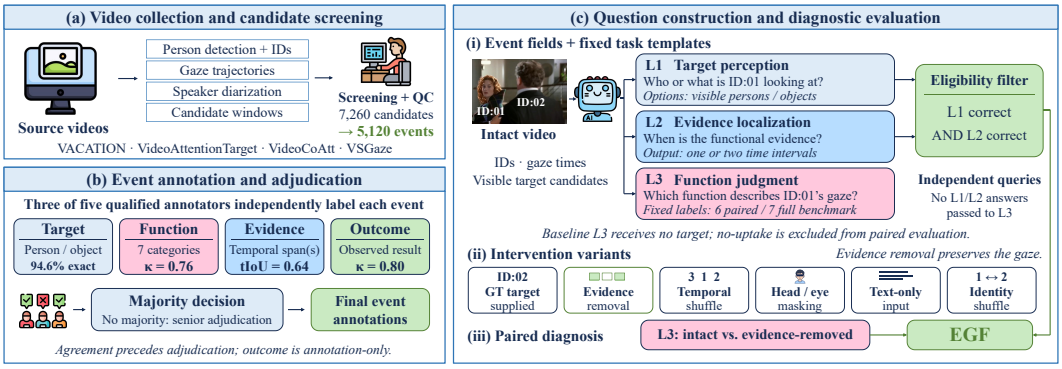

Figure 2: **EviGaze construction and evaluation.** (a) Screening retains 5,120 of 7,260 candidate events from four corpora. (b) Three annotators label each event, followed by majority decision or senior adjudication; agreement precedes adjudication and outcome is annotation-only. (c) Independent L1 to L3 queries, six interventions, and paired diagnosis. L2 receives the annotated target; baseline L3 receives no target or L1/L2 answers. The EGF rate measures label retention after removal of annotated evidence on events with correct L1 and L2.

3.2 EVENT COLLECTION

Diagnosing gaze function requires multi-person scenes in which the recipient of a look remains observable. We therefore build on four public 2D social-video corpora, VACATION (Fan et al., 2019), VideoAttentionTarget (Chong et al., 2020), VideoCoAtt (Fan et al., 2018), and VSGaze (Gupta et al., 2024a), drawing only VSGaze's ChildPlay (Tafasca et al., 2023) and UCO-LAEO (Marín-Jiménez et al., 2019) portions to avoid double counting. Candidate events are proposed in four steps: YOLOv8 and DeepSORT (Jocher et al., 2023; Wojke et al., 2017) detect people and keep their identities consistent, GazeNet (Zhang et al., 2019) estimates each person's gaze trajectory, pyannote.audio (Bredin, 2023) marks speaker turns, and sliding-window heuristics with an $8\,\mathrm{s}$ window and a $2\,\mathrm{s}$ stride propose candidate events. These heuristics only generate proposals; human annotation and review determine the final gaze boundaries, function, evidence interval, and outcome (Appendix A.2). Manual screening and quality control retain $5{,}120$ of $7{,}260$ candidates (70.5%); most rejections stem from ambiguous function, poor visibility, or insufficient temporal context (Appendix A). Every retained look therefore has an observable recipient and enough subsequent context to adjudicate its function from the ensuing interaction.

3.3 FUNCTION AND EVIDENCE ANNOTATION

Instead of asking only where a look lands ("Who or what is ID:01 looking at?"), EviGaze asks what role it plays ("Which function best describes the role of this gaze in the interaction?"). Each gaze event is therefore annotated as a triple (f, E, o): the function $f \in \mathcal{F}$ is the look's interactional role, the evidence E is the set of observable temporal spans that determine f, and the outcome o records uptake as information acquired, alignment achieved, or action advanced. Evidence begins at the first observable signal that distinguishes candidate functions, such as a recipient responding or the gazer acting or searching, and ends at the earliest moment the function becomes decidable. Writing G for the gaze interval, in which the look toward its target is established, we require $E \cap G = \emptyset$ and allow at most two non-contiguous spans in E; events needing more are excluded. Outcome is annotated but never enters an evaluation question or model input.

The seven functions differ in where their evidence lies: in the recipient's response for referential-communicative, social monitoring, object access, and joint-attention building (Tomasello, 1995), and in the gazer's own subsequent action for action preparation. Exploratory-scanning is decided by the search that follows, and no observable uptake by an opportunity that draws no response. Each function is defined by an executable criterion that names its deciding observation (Appendix A.3). The two temporal criteria are quantified: no observable uptake requires the recipient to stay visible and available for at least $2.0\,\mathrm{s}$, and exploratory-scanning requires at least two gaze-target transitions across three or more targets. Five qualified annotators label every event independently in groups of three, and a senior adjudicator resolves cases without a majority while leaving the raw agreement statistics unchanged. Agreement is substantial for function and outcome (Fleiss' $\kappa = 0.76$ and 0.80) and moderate for evidence boundaries (mean pairwise tIoU 0.64) (Fleiss, 1971; Landis & Koch, 1977), whereas gaze-target agreement is near ceiling (94.6%). Disagreement about *what* a look does is therefore not inherited from disagreement about *where* it points.

3.4 DIAGNOSTIC TASKS AND INTERVENTIONS

Separating functional errors from perceptual and localization errors requires querying each stage on its own. Every event therefore carries three separately prompted tasks: **L1** identifies the gaze target (Wang et al., 2026; Zhang et al., 2026), **L2** localizes the evidence given the annotated target rather than L1's prediction, and **L3** judges the function without L1 or L2 outputs. For model m, let $\mathcal{S}_m$ contain the events on which m answers L1 correctly and reaches $\mathrm{tIoU} \geq 0.5$ on L2, and let $\hat{y}_e^I$ and $\hat{y}_e^R$ be its L3 labels on the intact and evidence-removed versions of event e. The EGF rate is

$$\mathrm{EGF}_m = \frac{1}{|\mathcal{S}_m|} \sum_{e \in \mathcal{S}_m} \mathbb{1}\left[\hat{y}_e^R = \hat{y}_e^I\right]. \tag{1}$$

Restricting EGF to $\mathcal{S}_m$ ensures that label retention cannot reflect a missed target or interval.

Six interventions test competing explanations for a model's retained function label. *Evidence removal* deletes each annotated span with a $\pm 0.25\,\mathrm{s}$ margin that never enters G. It also deletes the

aligned ASR and subtitle text, which could otherwise restore the evidence through language. *GT target supplied* states the annotated target in the prompt, and *temporal shuffle* permutes frame order. Three shortcut probes complete the design: *head/eye masking* occludes the gazer's head (Zhang et al., 2026), *sanitized text-only* supplies dialogue stripped of gaze descriptions without video, and *identity shuffle* permutes person identifiers, keeping the frames. Among these conditions, only GT target supplied names the target. Each intervention reuses the events and questions of the intact condition, so any change in an answer traces to the content it alters.

4 EXPERIMENTS

4.1 EVALUATION SETUP

We evaluate fifteen MLLMs on EviGaze, including four proprietary models, GPT-4o (OpenAI, 2024), GPT-5.6 Sol (OpenAI, 2026), Gemini 3.1 Pro (Google DeepMind, 2026), and Claude Opus 5 (Anthropic, 2026). The eleven open-weight models span general-purpose families (Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025b), MiniCPM-V-4.6 (OpenBMB, 2026), and GLM-4.6V (Z.ai, 2025)), video-native models (LLaVA-Video-7B (Zhang et al., 2024) and VideoLLaMA 3 (Zhang et al., 2025)), and the reasoning-trained Video-R1 (Feng et al., 2025). Two scale ladders (Qwen3-VL 2B and 8B; InternVL3.5 2B, 4B, and 8B) and a thinking variant test whether capacity or explicit reasoning helps. Seven open-weight configurations, listed in Table 3, form the diagnostic subset for abstention, hint repair, probing, and prompting. All models receive the same 32 frames at 448 px with aligned ASR and subtitle text, never raw audio, under one prompt template (Appendix D); decoding and execution details are in Appendix C.

Accuracy alone cannot show whether an answer rests on evidence. For each intervention, we therefore report the retention of intact accuracy on eligible events, its change Δ, and the EGF rate of Eq. (1). Evaluation uses four nested sets: the full benchmark ($N = 5{,}120$), the full positive pool without no-uptake events ($N = 4{,}604$), and the removal-applicable set ($N = 4{,}138$). First-failure screening of the last leaves the fully paired subset ($N = 1{,}994$; Table 8), on which all fifteen models are compared. As references, uniform guessing yields 16.7% on the six positive functions, and 35 participants provide judgments on a 560-event subset, with different participants viewing the intact and removed version of each event (Appendix B.2).

4.2 MAIN RESULTS

Table 1 reports accuracy, retention, and EGF rates for all fifteen models and the human reference.

GPT-5.6 Sol is the most evidence-sensitive model, combining the highest Base accuracy with the lowest EGF rate. Open-weight models remain competitive: Video-R1 even attains a lower EGF rate than GPT-4o and Gemini 3.1 Pro. However, the results are far from satisfactory: all fifteen models retain labels on over 60% of eligible events without the deciding evidence. Humans retain only 31.8% of their intact accuracy under the same removal, whereas the average model retains 79.6%.

Temporal disruption also leaves most model accuracy intact. With sampled frames shuffled and aligned transcript text omitted, the fourteen evaluated models retain 77.7% of their intact accuracy, whereas humans retain only 37.5%. Models thus remain accurate on many events even without the original frame order or the aligned ASR/subtitle text.

Scale and explicit reasoning help only modestly. Quadrupling Qwen3-VL from 2B to 8B lowers its EGF rate by 5.4 points, and the thinking variant removes a further 6.3 points. Ground-truth target hints raise intact accuracy by only 2.9 points, so target perception is not their main limitation.

We compare two input baselines. Masking the gazer's head and eyes lowers retention to 68.3%, so models do attend to the look itself. Heavier reliance on the look accompanies weaker reliance on its evidence: across all fifteen models, masking retention is negatively correlated with EGF rate (Pearson $r = -0.87$, 95% CI $[-0.96, -0.65]$; Appendix L.4). Removing the video, however, costs them little: sanitized text-only input retains 79.7% of intact accuracy, practically matching evidence-removed video. Dialogue alone is thus about as informative as the video stripped of its evidence.

Table 1: **Evidence-grounding results across fifteen models and humans.** Models use the 1,994 paired events; †humans use 560 events (a nested 280 for Temp., Mask, and ID), scored by individual responses in preassigned between-subjects slots (Appendix B.2). Base is intact accuracy on all events, Elig. the percentage passing L1 and L2, and Base$_E$ intact accuracy on eligible events. GT tgt gives intact/removed accuracy gains from the ground-truth target. The five retention columns give accuracy as a percentage of Base$_E$, and Δ is the change under evidence removal. EGF is the EGF rate of Eq. (1); bold marks the lowest model EGF rate. Claude's Temp. and ID were not run and are excluded from those macro means ($n = 14$). These metrics are defined in §3.4 and §4.1.

Model	Base	Elig.	Base$_E$	*Evidence manipulation*				*Shortcut probes*			EGF
				GT tgt	Evid. rm	Δ	Temp.	Mask	Text	ID	
GPT-4o OpenAI (2024)	73.3	71.6	81.7	2.7/2.4	77.3	−18.5	73.7	70.0	74.5	71.8	67.9
GPT-5.6 Sol OpenAI (2026)	78.8	78.1	84.9	1.7/1.5	71.1	−24.5	68.0	74.0	84.4	67.6	**61.6**
Gemini 3.1 Pro Google DeepMind (2026)	76.5	75.0	84.3	3.3/3.4	80.7	−16.3	79.1	68.9	80.4	74.5	71.4
Claude Opus 5 Anthropic (2026)	77.1	76.5	83.1	1.3/2.0	74.0	−21.6	-	72.8	88.1	-	64.8
Qwen3-VL-2B Bai et al. (2025)	59.4	50.9	70.7	5.1/5.5	85.4	−10.3	83.2	59.2	77.0	80.7	80.0
Qwen3-VL-8B Bai et al. (2025)	69.4	65.6	76.2	3.2/4.1	81.9	−13.8	79.8	65.9	81.1	77.3	74.6
Qwen3-VL-8B-Thinking Bai et al. (2025)	71.8	68.8	81.0	2.3/2.7	76.2	−19.3	74.8	70.7	85.0	72.9	68.3
InternVL3.5-2B Wang et al. (2025b)	58.7	49.8	67.3	4.0/5.7	86.4	−9.2	85.5	58.0	74.5	82.2	81.4
InternVL3.5-4B Wang et al. (2025b)	64.4	57.6	74.4	4.0/4.7	83.4	−12.4	82.0	60.9	78.0	79.0	76.6
InternVL3.5-8B Wang et al. (2025b)	70.0	66.0	77.2	2.4/3.5	80.5	−15.1	77.8	67.2	81.3	75.9	73.1
MiniCPM-V-4.6 OpenBMB (2026)	66.6	61.1	75.8	3.6/4.5	84.2	−12.0	80.8	63.8	79.4	78.4	77.9
GLM-4.6V Z.ai (2025)	71.3	69.6	78.4	2.7/3.0	79.1	−16.4	76.3	70.0	85.7	74.2	70.7
LLaVA-Video-7B Zhang et al. (2024)	64.1	62.5	71.9	3.1/3.7	82.5	−12.6	80.1	72.7	70.0	76.6	75.4
VideoLLaMA 3 Zhang et al. (2025)	68.3	66.8	77.8	2.8/3.1	77.6	−17.4	75.1	74.7	73.6	71.3	69.1
Video-R1 Feng et al. (2025)	72.6	70.9	79.9	1.9/2.0	73.4	−21.3	70.9	76.3	82.5	69.1	64.2
Macro mean	69.5	66.1	77.6	2.9/3.5	79.6	−16.0	77.7	68.3	79.7	75.1	71.8
Human†	92.0	90.8	94.4	-	31.8	−64.4	37.5	65.0	-	42.8	-

Table 2: **Behavioural controls with human reference points.** (a) Cannot-be-determined (CD) rates after intact (I), neutral-deletion (N), and evidence-removed (R) inputs for seven open-weight models (1,994 events), four proprietary models (800 events), and humans (560 events). (b) Natural pairs with different functions; models answer each clip independently, whereas humans see both events together for direct comparison. (c) Prompting accuracy and stability on fixed baseline-eligible events; S_E and S_N denote evidence- and neutral-deletion stability, and Spec. $= S_N - S_E$. Human rows are excluded from model means (§4.3, §5.3).

(a) Abstention (CD)

Group	I	N	R
Open	4.8	5.9	24.8
Prop.	2.5	3.7	31.6
Human	1.8	3.9	74.6

(b) Natural A/B pairs

	Both correct	Same label	Neutral stab.
Models	44.4	34.7	91.6
Human	86.6	4.9	96.1

(c) Prompting controls

Prompt	Acc.	S_E	S_N	Spec.
Direct	74.3	75.8	93.4	17.6
Placebo	75.0	72.6	92.0	19.4
Evidence-gated	76.5	63.1	90.8	27.7
Oracle	77.6	57.6	89.7	32.1

4.3 Abstention and Natural Pairs

The six-way format forces an answer even after the evidence is gone, so it cannot show whether a model registers that loss. We therefore add an explicit cannot-be-determined (CD) option to the intact, neutral-deletion, and evidence-removed conditions. After evidence removal, humans choose CD in 74.6% of responses, whereas open-weight models on the paired subset and proprietary models on a fixed 800-event subsample do so in only 24.8% and 31.6% (Table 2a; Appendix J). No single model closes this gap: even GPT-5.6 Sol, which abstains most often, stays below half the human rate. Neutral deletion leaves all three groups near their intact rates.

We observe the same behavior without editing any video. Across 246 same-source pairs with matched gaze but different functions, models give the same label to 34.7% of pairs, roughly one in three. Human participants, who viewed each pair together, do so in only 4.9% (Table 2b; Appendix K). Models also answer both members correctly in only 44.4% of pairs, against 86.6% for

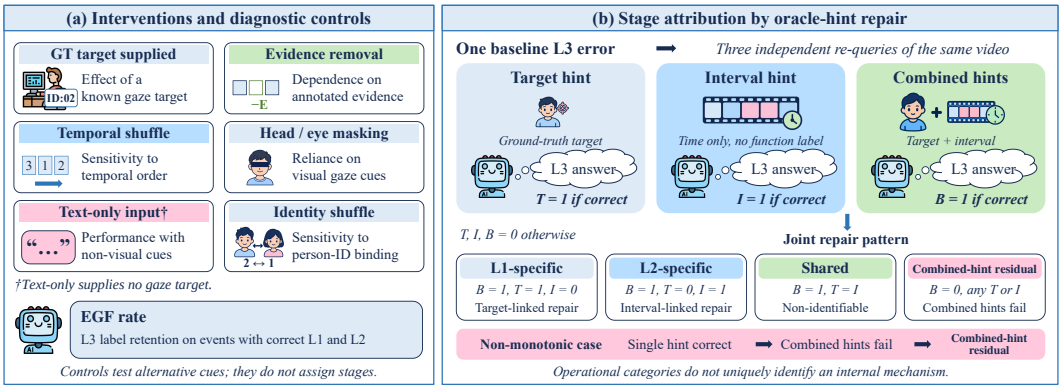

Figure 3: **Interventions and oracle-hint repair.** (a) Six interventions test sensitivity to evidence and alternative cues; text-only input supplies no gaze target. (b) Target, interval, and combined hints yield correctness indicators T, I, and B on each baseline L3 error, whose joint pattern defines L1-specific, L2-specific, shared, and combined-hint residual outcomes. The residual includes cases in which a single hint succeeds but combined hints fail.

humans. Figure 1 shows the same tendency on one event: all three models narrate the removed interaction back into their explanations. Across abstention, natural pairs, and the text-only baseline, models read a look from its setting and surrounding talk rather than from the response it receives.

4.4 Validity Checks

Correct answers are balanced across option positions, and varying the deletion margin from ± 0.10 to ± 0.50 s moves the EGF rate only between 73.4% and 70.9% (Appendix L). Excluding exploratory-scanning shifts Base accuracy and the EGF rate by less than half a point. On the seven diagnostic models, repeated identical queries flip 2.1% of answers under continuous batching (Kwon et al., 2023) but only 0.27% under deterministic execution (Appendix E), so execution controls remove most answer changes when the input remains identical.

Evidence visibility does matter: the EGF rate falls from 85.8% on events with no sampled evidence frame to 66.1% on events with at least three, in the same order for all fifteen models (Appendix C.5). Since this trend is observational, Section 5.1 directly manipulates model inputs.

5 Diagnosing Evidence-Grounding Failure

As analyzed in Section 4.3, models read a look from its setting and surrounding talk rather than from the response it receives. In this section, we locate where this failure arises, including oracle-hint repair, probing of internal states, and evidence-focused prompting.

5.1 Where Does the Failure Occur?

This behavior could stem from three stages, each with a prediction specified in advance. A perceptual bottleneck predicts that supplying the ground-truth target repairs errors, and a localization bottleneck predicts the same for the evidence interval. Neither hint should help much, however, if the model finds both and still does not let the evidence decide, which marks an evidence-use bottleneck. The interventions in Figure 3(a) probe alternative cues but cannot assign a stage, so we test these accounts with oracle target and interval hints.

For each baseline error of the seven diagnostic models, we re-query the same video three times, supplying the ground-truth target, the oracle evidence interval, or both. The interval marks when the evidence occurs without describing its content, so it helps only a model that uses what it sees there. The three outcomes T, I, and B assign each error to one of four categories (Figure 3(b)).

Each single hint repairs only about one fifth of baseline errors, and both together repair 44.1% (Appendix F). More than half of the errors thus survive even when the model is told whom the gazer

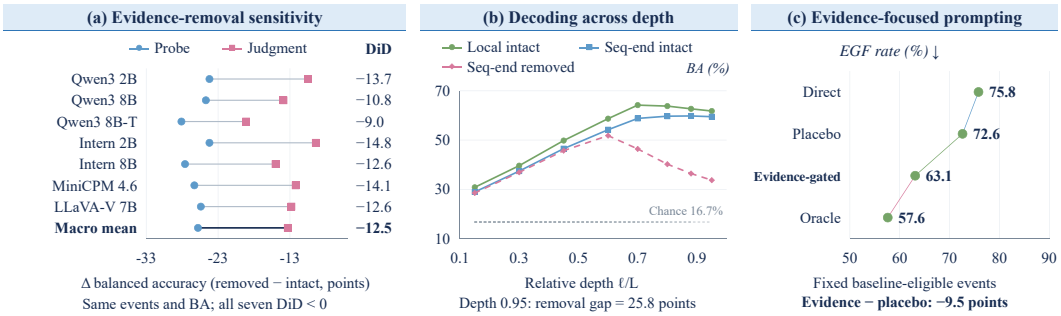

Figure 4: **Removal sensitivity, depth profiles, and evidence-focused prompting.** (a) Probe and judgment changes use the same events and balanced-accuracy metric; all seven probe-minus-judgment contrasts are negative (macro -12.5 points). (b) Local and sequence-end decoding across relative depth; the sequence-end removal gap reaches 25.8 points at depth 0.95. (c) On fixed baseline-eligible sets, evidence-focused prompting lowers the EGF rate by 9.5 points relative to the matched placebo; oracle hints supply the annotated target and interval.

looks at and when the deciding evidence occurs; this residual also contains the 9.9% of errors whose single-hint repair is undone by combined hints. The residual barely varies across configurations, ranging from 55.3% to 56.6% in all seven models. The results therefore favor the evidence-use account, consistent with the small gain from target hints in Section 4.2. Telling models where and when to look leaves most of their baseline errors in place.

5.2 Decodability and Judgment Sensitivity

If evidence use is the bottleneck, the deciding information should reach the model's internal states without fully shaping its answer, which we call the underuse account; the competing encoding account holds that the information is never encoded. Probing these states is not straightforward, since evidence removal also deletes the tokens that would carry the evidence.

To address this, we use two linear readouts (Alain & Bengio, 2016) on the seven diagnostic models. A local probe pools frozen visual tokens over the evidence interval at a preselected relative depth of about 0.75 and tests whether gaze function is encoded at all. A sequence-end probe reads the last input token before generation at depth 0.95; trained once on intact inputs and then frozen, it is applied to intact, evidence-removed, and neutral-deletion inputs alongside model judgments. Both are scored by balanced accuracy (BA; chance 16.7%) with permuted-label controls (Hewitt & Liang, 2019), and all variants of an event share a fold. Since probes and generated answers are different readouts, we compare their changes under removal rather than their levels (Appendix G).

Gaze function is clearly encoded. Local decoding reaches 64.1% BA against 17.1% for permuted labels and exceeds both full-video and gaze-interval pooling in all seven models (Table 3a), placing the information in the evidence interval itself. Yet the answers draw on this information only partly. Evidence removal lowers sequence-end probe BA by 25.8 points but judgment BA by only 13.3, a difference-in-differences (DiD) of -12.5 points (95% CI $[-13.9, -11.1]$) that is negative in all seven models (Table 3b; Figure 4(a)). Neutral deletion moves both readouts by about one point, so the gap reflects the lost evidence rather than a shorter video. Holding out each corpus in turn preserves a sequence-end removal drop of 18.3 to 22.6 points (Appendix G.6). The divergence also emerges late: intact and removed sequence-end states stay within one point through the first half of the network and separate only near relative depth 0.65 and diverge most at 0.95 (Figure 4(b)). Local decoding still reaches 61.8% BA at this depth, so the information remains recoverable late in the network (Table 18). The results therefore favor the underuse account: models encode what the evidence shows, but their answers only partly reflect it.

5.3 Evidence-Focused Prompting

Since evidence is encoded but underused, directing models to it might restore its influence. Any gain, however, could come from generic step-by-step structure. We therefore compare three prompts on

Table 3: **Evidence-specific decoding and sensitivity dissociation.** Seven diagnostic models on the 1,994 paired events; T denotes Thinking. (a) Local probe BA at relative depth 0.75, pooling over the evidence interval (Evid.), the gaze interval (Onset), or the full video (Glob.); Perm. uses permuted training labels, $E - G$ is Evid. minus Glob., and Src. holds out source videos. (b) Sequence-end probe and judgment BA on intact (I) and evidence-removed (R) inputs; Δ_P and Δ_J are removed minus intact, DiD $= \Delta_P - \Delta_J$, and Δ_S is Δ_P under cross-validation folds that keep source videos separate. Changes are in percentage points (§5.2). Bold marks BA column maxima, excluding Perm.

(a) Local decoding controls

Model	Evid.	Perm.	Onset	Glob.	$E - G$	Src.
Qwen3 2B	58.4	17.4	40.1	49.7	+8.7	54.9
Qwen3 8B	64.7	16.8	44.8	52.0	+12.7	60.8
Qwen3 8B-T	68.9	17.1	50.7	**59.1**	+9.8	64.7
Intern 2B	57.6	17.7	37.4	46.1	+11.5	53.8
Intern 8B	66.2	16.5	46.2	52.7	+13.5	62.0
MiniCPM 4.6	62.5	17.9	41.9	53.4	+9.1	58.6
LLaVA-V 7B	**70.3**	16.6	**50.8**	56.0	+14.3	**65.9**
Macro mean	64.1	17.1	44.6	52.7	+11.4	60.1

(b) Paired evidence-removal sensitivity

Model	Probe BA I	Probe BA R	Judgment I	Judgment R	Δ_P	Δ_J	DiD	Δ_S
Qwen3 2B	55.2	31.0	58.7	48.2	−24.2	−10.5	−13.7	−22.7
Qwen3 8B	59.4	34.7	68.9	**55.0**	−24.7	−13.9	−10.8	−23.9
Qwen3 8B-T	63.1	35.0	**71.3**	52.2	−28.1	−19.1	−9.0	−26.2
Intern 2B	55.0	30.8	58.2	48.8	−24.2	−9.4	−14.8	−22.8
Intern 8B	61.2	33.6	69.5	54.5	−27.6	−15.0	−12.6	−25.7
MiniCPM 4.6	57.8	31.5	66.1	53.9	−26.3	−12.2	−14.1	−24.9
LLaVA-V 7B	**64.6**	**39.2**	63.7	50.9	−25.4	−12.8	−12.6	−24.1
Macro mean	59.5	33.7	65.2	51.9	−25.8	−13.3	−12.5	−24.3

each model's fixed baseline-eligible events. The evidence-gated prompt asks the model to identify the target, localize and describe the observable evidence without inferring intent, and then choose a label. A placebo keeps the same three steps and budget but describes a pre-selected neutral interval, and an oracle prompt supplies the ground-truth target and interval (Appendix D).

Evidence focus, not structure, drives the improvement. The placebo lowers the EGF rate by 3.2 points, whereas evidence gating lowers it by a further 9.5 points (95% CI $[-11.8, -7.3]$; Figure 4(c); Table 2c). This is not indiscriminate switching: gating also raises intact accuracy and widens the neutral- versus evidence-deletion stability gap over the placebo in all seven models (Appendix H). Even the oracle leaves the EGF rate at 57.6%, so prompting narrows the failure without closing it.

The remaining errors show what takes the evidence's place. Under gating, 63.6% of residual errors occur even though the model has localized the evidence correctly. When both such an answer and the text-only answer are wrong, they coincide in 72.4% of cases, whereas probe errors coincide with the text-only answer in only 24.6%, close to the 20.0% expected from uniform wrong choices (Appendix I). Models that have found the evidence thus fall back on what dialogue alone suggests, while their internal states do not, confirming the reliance on surrounding talk seen in Section 4.3.

Taken together, the three diagnostic analyses converge. Models find the target, locate the evidence, and encode what happens there, yet their answers rest largely on the look and the talk around it. For current MLLMs, knowing where is not knowing what.

6 Conclusion

In this work, we introduce EviGaze to test whether MLLMs ground social gaze judgments in the ensuing interaction. Its 5,120 gaze events each carry one of seven functions and an evidence interval disjoint from the look. All fifteen MLLMs exhibit evidence-grounding failure: unlike humans, they mostly keep their labels once the deciding evidence is removed. The bottleneck lies in evidence use rather than perception or localization, since evidence stays decodable from model states while answers rely on surrounding dialogue. Prompting models to describe the evidence first reduces but does not remove this failure. Overall, EviGaze provides a rigorous yardstick for both evaluating and improving evidence grounding, the step from knowing where to knowing what.

Limitations. EviGaze covers functions decidable from compact, observable evidence in four public corpora, and excludes interactions whose function depends on longer histories or private context. The labels describe observable interactional roles rather than private intent. Our probes are supervised readouts of internal states, so the dissociation in Section 5.2 concerns recoverable information rather than its causal routing. Appendix N discusses further limitations.

AI USAGE STATEMENT

AI tools assisted with language polishing, model inference code, and arithmetic consistency checks of reported summaries. Their use included sentence-level revisions, terminology alignment, and checks of relationships among counts, percentages, and reported contrasts. The authors conducted the study design, data collection and annotation, experimental execution, and the original statistical analyses underlying the reported results. AI assistance supported manuscript preparation and implementation; it did not replace human annotation or experimental evaluation. The authors take responsibility for the manuscript and all AI-assisted text, code, and interpretations.

REPRODUCIBILITY STATEMENT

The appendix provides a guide to reproducing the benchmark construction, evaluation, and diagnostic analyses. Appendix A specifies source composition, candidate screening, annotation criteria, quality control, and evaluation sets; Appendix B records participant assignment, trial procedures, scoring, and consent. Appendices C and D describe model configurations, frame sampling, aligned transcripts, decoding, answer parsing, prompt templates, and input transformations. Appendix E documents execution controls and repeated-query variability.

For the diagnostic results, Appendix F gives the joint hint-outcome counts and repair calculations, while Appendix G specifies probe features, fitting, folds, seeds, uncertainty estimation, and source-grouped and held-out-corpus validation. Appendices H and I provide prompting controls, eligibility denominators, and residual-error definitions. Appendices J and K detail the behavioural controls; Appendices L and M document robustness checks and no-uptake criteria. Appendix N states the scope and limitations relevant to interpreting and reproducing these results.

ETHICS STATEMENT

EviGaze uses videos from four existing public corpora: VACATION, VideoAttentionTarget, VideoCoAtt, and VSGaze. Its annotations concern observable interactional roles rather than private intent or sensitive personal attributes. Reuse and redistribution of source material must respect the original corpora's access conditions and licensing terms.

Human evaluation involved adult participants. As documented in Appendix B, participants received information about the task, duration, compensation, and research use of anonymized responses, provided affirmative electronic consent, and could stop without penalty. Participants were compensated at an equivalent rate of USD 10 per hour and instructed to judge the supplied interaction rather than identify depicted individuals or infer sensitive attributes. Social-gaze labels remain context-dependent; benchmark performance should not be taken as evidence of suitability for surveillance or other individual-level assessments beyond the evaluated social-gaze tasks.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Anthropic. Introducing Claude Opus 5, 2026. URL https://www.anthropic.com/news/claude-opus-5.

Michael Argyle and Mark Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge, 1976.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. In *Proceedings of Interspeech*, pp. 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5396–5406, 2020.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (eds.), *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, Washington, DC, 1991. doi: 10.1037/10096-006.

Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972. doi: 10.1037/h0033031.

Nathan J. Emery. The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000. doi: 10.1016/S0149-7634(00)00025-7.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6460–6468, 2018.

Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5724–5733, 2019.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776v1*, 2025. URL https://arxiv.org/abs/2503.21776v1.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025. URL https://arxiv.org/abs/2506.08008.

Alexey Gavryushin, Dingxi Zhang, Zhao Huang, Alexandros Delitzas, Jiaqi Chen, Ben Ellis, Cedric Zöllner, Manthan Patel, Manuel Kaufmann, Marc Pollefeys, and Xi Wang. CoMind: Understanding collaborative human activity from multiple minds and views. *arXiv preprint arXiv:2607.06691*, 2026. URL https://arxiv.org/abs/2607.06691.

Google DeepMind. Gemini 3.1 Pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Xiao-Yu Guo, Yuan-Fang Li, and Gholamreza Haffari. DeSIQ: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3169–3180. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.191.

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 15646–15673, 2024a.

Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 615–624, 2024b. URL https://arxiv.org/abs/2406.03907.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2Gc8aj0afg.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. GFIE: A dataset and baseline for gaze-following from 2D to 3D in indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8907–8916, 2023.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Ultralytics software, 2023. URL https://github.com/ultralytics/ultralytics.

Adam Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1): 22–63, 1967. doi: 10.1016/0001-6918(67)90005-4.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 3587–3602. ELRA and ICCL, 2024.

Junho Kim, Xu Cao, Houze Yang, Bikram Boote, Ana Jojic, Fiona Ryan, Bolin Lai, Sangmin Lee, and James M. Rehg. GRASP: Learning to ground social reasoning in multi-person non-verbal interactions. *arXiv preprint arXiv:2605.15764*, 2026. URL https://arxiv.org/abs/2605.15764.

Chris L. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100(1):78–100, 1986. doi: 10.1037/0033-2909.100.1.78.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. URL https://arxiv.org/abs/2309.06180.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, and James M. Rehg. Towards social AI: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024. URL https://arxiv.org/abs/2409.15316.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.730.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Surbhi Madan, Shreya Ghosh, Ramanathan Subramanian, Abhinav Dhall, and Tom Gedeon. CSGaze: Context-aware social gaze prediction. *arXiv preprint arXiv:2511.05955*, 2025. URL https://arxiv.org/abs/2511.05955.

Manuel J. Marín-Jiménez, Vicky Kalogeiton, Pablo Medina-Suárez, and Andrew Zisserman. LAEO-Net: Revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3477–3485, 2019.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025. URL https://arxiv.org/abs/2502.15109.

Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. R^3-VQA: "read the room" by video social reasoning. *arXiv preprint arXiv:2505.04147*, 2025. URL https://arxiv.org/abs/2505.04147.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010. URL https://www.jmlr.org/papers/v11/ojala10a.html.

OpenAI. GPT-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/. Model release documentation; accessed 2026-09-26.

OpenBMB. MiniCPM-V 4.6 model card, 2026. URL https://huggingface.co/openbmb/MiniCPM-V-4.6.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Chiara Plizzari, Alessio Tonioni, Yongqin Xian, Achin Kulshrestha, and Federico Tombari. Omnia de EgoTempo: Benchmarking temporal understanding of multi-modal LLMs in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24129–24138, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Plizzari_Omnia_de_EgoTempo_Benchmarking_Temporal_Understanding_of_Multi-Modal_LLMs_in_CVPR_2025_paper.html.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28, pp. 199–207, 2015.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. doi: 10.1353/lan.1974.0010.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977. doi: 10.2307/413107.

scikit-learn developers. StratifiedGroupKFold, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html. Online documentation; accessed 2026-09-26.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. MediConfusion: Can you trust your AI radiologist? probing the reliability of multimodal medical foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=H9UnNgdq0g.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. ChildPlay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20935–20946, 2023.

Michael Tomasello. Joint attention as social cognition. In Chris Moore and Philip J. Dunham (eds.), *Joint Attention: Its Origins and Role in Development*, pp. 103–130. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

Hengfei Wang, Anshul Gupta, Pierre Vuillecard, and Jean-Marc Odobez. Eyes on VLM: Benchmarking gaze following and social gaze prediction in vision language models. *arXiv preprint arXiv:2605.19859*, 2026. URL https://arxiv.org/abs/2605.19859.

Shijing Wang, Chaoqun Cui, Yaping Huang, Hyung Jin Chang, and Yihua Cheng. VL4Gaze: Unleashing vision-language models for gaze following. *arXiv preprint arXiv:2512.20735*, 2025a. URL https://arxiv.org/abs/2512.20735.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265v1*, 2025b. URL https://arxiv.org/abs/2508.18265v1.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017. doi: 10.1109/ICIP.2017.8296962.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. URL https://arxiv.org/abs/2504.15279.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Z.ai. GLM-4.6V model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V.

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

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. doi: 10.1109/TPAMI.2017.2778103.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713v1*, 2024. URL https://arxiv.org/abs/2410.02713v1.

Zory Zhang, Pinyuan Feng, Bingyang Wang, Tianwei Zhao, Suyang Yu, Qingying Gao, Hokin Deng, Ziqiao Ma, Yijiang Li, and Dezhi Luo. Vision-language models mistake head orientation for gaze direction: Nonverbal conversation cues. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10370–10389, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.504. URL https://aclanthology.org/2026.findings-acl.504/.

A BENCHMARK CONSTRUCTION AND ANNOTATION

The benchmark contains 15,360 core questions, one L1, L2, and L3 question per event, plus 24,700 intervention variants. Its annotation unit is a gaze event; multiple events can share a source video. In tables and quantitative comparisons below, EGF denotes the EGF rate. Figure 5 summarizes event construction, independent diagnostic tasks, and evidence-grounding analyses.

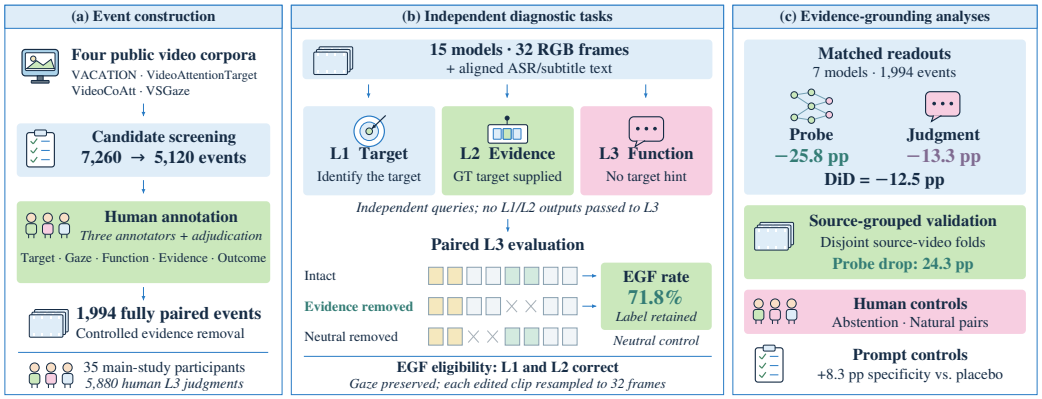

Figure 5: Experimental and analytical workflow. **(a)** Candidate screening, human annotation, and paired-set construction. **(b)** Independent diagnostic queries and paired L3 evaluation. Gold and green blocks mark the gaze and evidence intervals; crosses indicate removed spans schematically. Aligned ASR/subtitle text is removed with each span; no raw audio is supplied. The EGF rate is the fifteen-model macro mean on model-specific eligible events. **(c)** Sequence-end probe and judgment comparisons, source-grouped validation, human controls, and prompting controls. Changes in balanced accuracy are evidence-removed minus intact, in percentage points.

A.1 SOURCE CORPORA AND DUPLICATE CONTROL

Videos come from VACATION, VideoAttentionTarget, VideoCoAtt, and VSGaze (Fan et al., 2019; Chong et al., 2020; Fan et al., 2018; Gupta et al., 2024a). VSGaze contributes only ChildPlay and UCO-LAEO material to avoid overlap with the other sources (Tafasca et al., 2023; Marín-Jiménez et al., 2019). Duplicate and near-duplicate clips also undergo pre-annotation quality control. Table 4 separates the retained core events from those in the paired positive subset.

Table 4: Source composition of the core benchmark and paired positive subset. VSGaze contributes only its ChildPlay and UCO-LAEO portions. The paired subset supports the six positive functions and the main intervention comparisons; source contributions are counts of events, not independent recordings. Multiple events can originate from the same source video.

Source corpus	Core events	Paired events
VACATION	2,650	1,130
VideoAttentionTarget	1,430	488
VideoCoAtt	610	215
VSGaze	430	161
Total	5,120	1,994

A.2 SCREENING AND EXCLUSION ACCOUNTING

Candidate proposals use an 8 s sliding window with a 2 s stride. Person detection and tracking, gaze-trajectory estimation, and speaker diarization run within each window. A candidate gaze episode must last at least 0.4 s, with at least 60% of valid frames pointing to the same predicted person or object target. Episodes with the same gazer and target are merged if their gap is at most 0.5 s, and overlapping duplicates with temporal IoU ≥ 0.7 are removed. Both the gazer and target must be trackable in at least 70% of frames within the candidate gaze interval before manual screening. At

least 2 s of observable context is required after the gaze; video endings, shot cuts, or severe occlusion that prevent observation of subsequent interaction are marked as insufficient temporal context. These heuristics only propose events; human annotation and review determine all final gaze boundaries, functions, evidence intervals, and outcomes for each event.

After automatic proposal, merging, and deduplication, the candidate pool contains 7,260 events, of which 5,120 are retained (70.5%). Table 5 accounts for all 2,140 exclusions. Manual screening removes 1,662 candidates, principally for ambiguous function, poor visibility, or insufficient context. The remaining 478 fail pre-annotation quality control: duplicate media, unstable identity tracking, or media and timing failures. These failures concern whether the stimulus can support a reliable annotation before its functional interpretation is scored.

Table 5: Exclusions from the candidate pool. The five manual-screening categories account for 1,662 exclusions, and pre-annotation quality control accounts for another 478. Together they reconcile the 7,260 candidates with 5,120 retained core events.

Recorded exclusion reason	Events
Manual screening	
Ambiguous interactional function	548
Poor visibility or severe occlusion	372
Insufficient temporal context	314
Evidence overlaps gaze or cannot be cleanly removed	254
Requires three or more disjoint evidence spans	174
Pre-annotation quality control	
Cross-corpus duplicate or near-duplicate clip	171
Unstable gazer or recipient identity track	159
Media decoding, audiovisual synchronization, or timestamp failure	148
Total excluded	2,140

A.3 Function taxonomy and executable criteria

The labels describe observable interactional roles rather than private intent. Recipient response and action sequence motivate the taxonomy (Kendon, 1967; Sacks et al., 1974; Schegloff et al., 1977; Clark & Brennan, 1991). The criteria below identify the evidence that distinguishes the labels.

Referential-communicative. The gaze participates in a directed communicative act, such as requesting, answering, confirming, cueing, or referring. The deciding observation is uptake that completes or advances that act, rather than orientation toward another person alone. This label takes precedence when uptake completes a specific request or reference.

Social monitoring. The gaze checks another person's state, availability, attention, or reaction without itself advancing a specific referential request or an object or action exchange. A person-directed look alone is insufficient. A completed request instead supports referential-communicative.

Object access. The gaze supports obtaining, receiving, locating, or coordinating access to a physical object. Subsequent access or transfer supplies the disambiguating evidence. The target being an object does not alone establish this function: the later interaction must connect the look to access.

Joint-attention building. The gaze establishes or confirms that multiple people attend to the same external referent. Recipient gaze-following or coordinated orientation toward that referent is central to the label. The deciding relation is shared attention to an external referent.

Exploratory-scanning. The gaze forms part of active visual search rather than a recipient-directed social act. The executable criterion requires at least two gaze-target transitions involving at least three distinguishable targets. The threshold distinguishes visual search from an isolated gaze event.

Action preparation. The gaze prospectively supports an imminent action by the gazer, such as reaching, moving, or manipulating. The gazer's subsequent action is the disambiguating observation. The observed action must explain the preceding look; plausibility alone is insufficient.

No observable uptake. The recipient remains continuously visible and behaviourally available for at least 2.0 seconds after the gaze, but no observable response or uptake occurs. A shot cut or severe occlusion invalidates the opportunity to observe a response. This is a label for an observed opportunity without uptake, not for a clip in which the response happens to be missing. Its different evidential structure is the reason for separating it from positive-uptake deletion (Appendix M).

A.4 Class composition and annotation agreement

Table 6 reports the core and paired counts together with class-specific agreement. The positive classes sum to $4{,}604$ core events; adding 516 no-uptake events gives the full $5{,}120$. The paired set contains all six positive functions, with 259 to 431 events per class. Class-balanced scoring gives each function equal weight in the probe-judgment comparison despite these unequal event counts.

Table 6: Function composition and class-specific agreement before adjudication. The six positive functions contribute 4,604 core events and all 1,994 paired events. No observable uptake remains in the full benchmark. The final column reports class-specific agreement, distinct from overall function agreement of $\kappa = 0.76$. The agreement values precede the final adjudication of labels.

Function	Core	Paired	κ
Referential-communicative	987	431	0.79
Social monitoring	901	389	0.73
Object access	781	335	0.82
Joint-attention building	713	314	0.76
Exploratory-scanning	628	266	0.74
Action preparation	594	259	0.74
No observable uptake	516	—	0.79
Total	5,120	1,994	—

Overall function agreement is Fleiss' $\kappa = 0.76$, and outcome agreement is 0.80. Evidence intervals have mean pairwise temporal intersection over union (tIoU) of 0.64, with 78.9% of comparisons reaching 0.5. L1 target annotations have 94.6% pairwise exact agreement, while all three annotators agree in 91.2% of cases. These quantities measure different annotation decisions and do not constitute a single interchangeable reliability score.

A.5 Evidence interval annotation

Evidence starts at the first externally observable signal that distinguishes plausible functions and ends at the earliest point at which the function is decidable. The interval therefore need not extend to the end of an action or the end of the clip. It must lie outside the gaze interval, so that the look and its target remain visible when evidence is removed. At most two non-contiguous spans are retained; cases needing three or more disjoint spans are excluded under the screening rule.

A single span ordinarily lasts 0.3 to 5.0 seconds, with total evidence duration no greater than 8 seconds as a soft constraint. Temporal compactness is useful for the intervention, but the deciding criterion remains the information needed to distinguish functions. For model inputs, aligned ASR/subtitle text inside the interval is removed with the visual segment; otherwise the text could preserve the withdrawn evidence. Models receive no raw audio waveform.

A.6 Annotator qualification and quality control

Annotators complete 120 training items followed by a 60-event qualification set. Qualification requires at least 90% L1 accuracy, 80% function accuracy, 85% outcome accuracy, and evidence tIoU of at least 0.55. An annotator who fails may retrain and retake the qualification once. Each component must meet its threshold independently of performance on the other annotation tasks.

Three of the five qualified annotators label each event independently. A senior adjudicator resolves cases without a majority. Hidden gold items constitute 5% of the annotation stream; repeated events provide an additional check on within-annotator consistency. Agreement statistics use the original annotations before adjudication. Adjudication produces the evaluation target, while agreement on the original independent labels measures annotation reliability before disagreements are resolved.

A.7 Evaluation sets and distribution shift

Table 7 separates annotation scope from intervention applicability. Model-specific L1/L2 eligibility is an additional condition within each set, not a fixed split shared by all models.

Table 7: Nested evaluation sets and their model-macro summaries. Base is function accuracy on the named set; the full benchmark has seven labels, whereas the positive sets have six. EGF additionally conditions on each model passing L1 and L2. Dashes denote quantities not reported for that set.

Evaluation set	Events	Base (%)	EGF (%)
Full benchmark	5,120	68.4	—
Full positive pool	4,604	69.0	—
Removal-applicable positive set	4,138	—	71.0
Fully paired positive subset	1,994	69.5	71.8

The 4,138 removal-applicable events undergo the ordered first-failure checks in Table 8. Their mutually exclusive counts remove 2,144 events, leaving 1,994. Although 401 excluded events carry at least two failure flags, each is counted only at its first failed check. These criteria define pairing feasibility, not model-specific eligibility.

Table 8: Ordered first-failure screening from the removal-applicable pool to the fully paired subset. Counts are mutually exclusive despite overlapping flags.

First failed check	Excluded events
No legal neutral interval	572
No usable sanitized transcript	418
Insufficient tracking	386
Failed 32-frame quality control	311
Failed source-video/audio-text synchronization	249
Inconsistent metadata	208
Total excluded	2,144
Retained from 4,138	1,994

Paired-set Base accuracy is 69.5%, compared with 69.0% on all positive events. EGF is 71.8% on the paired set and 71.0% on the removal-applicable set. These aggregate checks do not imply unchanged class or source composition. Full-benchmark accuracy additionally uses a seventh label.

B Human Evaluation

B.1 Participants and assignment

The main study includes 35 adults who are fluent in English and have normal or corrected-to-normal vision. Recruitment uses a university participant mailing list and local community advertising. The study produces 5,880 L3 function judgments, equivalent to 168 per participant on average. L1 target and L2 evidence-localization diagnostic responses are recorded separately and are not included in this total. Versions are assigned between subjects: a participant never sees multiple versions of the same event. This prevents the intact interaction from revealing what was removed in a later presentation of that event.

The 560 primary events are sampled from the 1,994 paired events by source-corpus $\times$ function strata, using largest-remainder allocation and seed 42. The 280 auxiliary events form a nested subset of those 560, using the same stratification and seed 43. In source order VACATION, VideoAttentionTarget, VideoCoAtt, and VSGaze, counts are 317/137/61/45 for 560 and 159/68/30/23 for 280.

The primary comparison uses 560 events with three preassigned response slots per event, yielding 1,680 slots. Each $(\text{event_id}, \text{replicate_id})$ binds one intact participant and a different removal participant before responses are collected; no post-hoc rematching is allowed. Auxiliary interventions use three L3 judgments for 280 events per condition. Table 9 gives the allocation. The auxiliary human columns in Table 1 therefore use a smaller event sample.

Table 9: Allocation of human L3 function judgments. The two main-study blocks sum to 5,880 L3 judgments from 35 participants; separately recorded L1/L2 diagnostic responses are excluded. The abstention session uses the same participant pool with disjoint event assignments for each person. The natural-pair pilot uses a separate 24-person cohort and is described in Appendix K. No individual participant sees multiple versions of the same event.

Study block	Events	Versions	Ratings per cell	Judgments
Main: intact / removal	560	2	3	3,360
Main: auxiliary interventions	280	3	3	2,520
Main-study total	—	—	—	5,880
Abstention session	560	3	2	3,360

B.2 Response-level scoring and eligibility

Table 1 scores individual L3 responses, without majority voting. Let $s = (e, r)$ index an event and its preassigned replicate slot, and let $\hat{y}_s^I$ and $\hat{y}_s^R$ be its intact and removal L3 responses from different participants. The intact-side participant's L1 answer must identify the correct target and their L2 interval must reach tIoU ≥ 0.5 for the slot to be eligible. Thus $\mathcal{S}_E$ is selected solely from intact-side diagnostics; its removal response is the one already assigned to that slot. For condition $c \in \{I, R\}$ and slot set $\mathcal{S}$,

$$\text{Acc}_c(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1}[\hat{y}_s^c = y_{e(s)}]. \tag{2}$$

Of 1,680 intact responses, 1,546 are correct (Base 92.0%) and 1,525 slots pass both diagnostics (Elig. 90.8%). Within these eligible slots, 1,440 intact responses are correct (Base$_E$ = 94.4%), versus 458 of their preassigned removal responses (30.0%). The accuracy change is $100(458 - 1{,}440)/1{,}525 = -64.4$ percentage points after rounding; retention is $458/1{,}440 = 31.8\%$.

The comparison retains the same slots across conditions but uses different participants. It therefore does not measure same-person answer changes and does not define a human EGF. Eligibility describes the intact-side responses, not demonstrated localization ability of the removal-side participants. The uniform six-way guessing reference for an individual response is $1/6 = 16.7\%$, as for a single model prediction.

B.3 Instructions and trial procedure

Participants receive the following task instructions:

> You will watch short clips of social interaction and answer questions about one person's gaze. A highlighted interval identifies the look being evaluated. Person identifiers are arbitrary labels. Use only what is observable in the clip. You may replay each clip up to two times after the initial presentation. In the main task there is no uncertain option; choose the best available answer. Do not record the screen or discuss individual clips with other participants.

For intact-side eligibility diagnostics, the task order is L3 function judgment, then L1 target identification, then L2 evidence localization. The L3 answer is collected before either diagnostic, so L1/L2 do not reveal the target or evidence interval in advance of the function judgment. Eligibility is computed from these later diagnostic responses and applied to the fixed slots defined in Appendix B.2.

Each trial begins with one automatic presentation and permits up to two optional replays. The response screen has a 20-second limit with a warning at 15 seconds. An unanswered trial is marked missing at the limit and repeated once at the end of the session. A three-minute break follows every 56 trials. The main session lasts approximately 60 minutes and is compensated at USD 10 per hour.

B.4 Abstention session and natural-pair cohort

The same 35 participants complete the abstention session 7 to 14 days later. It contains 560 events in intact, neutral-deletion, and evidence-removal conditions, with two independent judgments per event-condition and 3,360 judgments overall. Each participant sees at most one version of an event. Events previously shown to that person in the main study are excluded from their assignment. The session lasts approximately 40 minutes and is compensated at USD 10 per hour.

The natural A/B pilot uses a separate cohort of 24 adults. Each of the 246 intact pairs receives three independent judgments. A separate counterbalanced assignment evaluates matched neutral deletion; no participant sees both versions of a pair. Presentation order of A and B is balanced equally. Compensation is at an equivalent hourly rate. The response instructions for both auxiliary studies appear with their corresponding analyses in Appendices J and K.

B.5 Consent and response interpretation

Before beginning, participants view an information sheet describing the task, duration, compensation, research use of anonymized responses, voluntary participation, and the right to stop without penalty. They provide affirmative electronic consent. The instructions concern observable interactional content; participants are asked to judge gaze function using the supplied clip rather than identify depicted individuals or infer sensitive personal attributes.

On the 1,525 eligible matched response slots, human accuracy decreases from 94.4% to 30.0%, corresponding to 31.8% retention after rounding. These between-subjects accuracies do not define same-person label stability, so no human EGF is inferred from them.

C Model Evaluation Protocol

C.1 Model coverage and analysis scope

The main pool contains the fifteen models in Table 1. The seven diagnostic configurations are Qwen3-VL-2B, Qwen3-VL-8B, Qwen3-VL-8B-Thinking, InternVL3.5-2B, InternVL3.5-8B, MiniCPM-V-4.6, and LLaVA-Video-7B. Their feature-extraction configuration uses bfloat16 precision. Proprietary models use all 1,994 paired events in the main evaluation and a fixed 800-event subset for auxiliary analyses that require repeated queries to the same model.

All model evaluations, including API calls, took place from 29 July to 2 August 2026, inclusive. The dates identify the evaluation window for the reported results; model names follow the configurations listed in Table 1 throughout the main and auxiliary analyses. Claude Opus 5 Temporal Shuffle and Identity Shuffle were not run. Their dashes are missing values, not zeros, and are excluded from those column means ($n = 14$); other main model means use $n = 15$.

L1, L2, and baseline L3 are separately prompted diagnostic tasks. L2 receives the annotated GT target directly through `Known gaze target`; it does not receive L1's prediction. L2 therefore measures function-defining evidence localization conditional on a correct target, without propagation of L1 errors. Baseline L3 receives neither the GT target nor L1/L2 outputs. Eligibility is the intersection of correct L1 and correct target-conditioned L2; it does not represent an end-to-end chain in which L1's answer is passed to L2 and L3. Among L3 function-judgment conditions, only GT-target and oracle prompts supply the target; baseline L3, text-only, and CD do not.

C.2 Frame sampling and preprocessing

All models receive 32 RGB frames plus temporally aligned ASR/subtitle text; no raw audio waveform is supplied. The intact input samples frames uniformly by timestamp, including the first and last frames. Each frame is resized to a 448-pixel long side with aspect-preserving padding and no centre crop. Models share the sampled frames for each input version. Modality-changing ablations are specified in Appendix D.

Evidence removal first deletes the visual and aligned text segments for each annotated evidence span with a ± 0.25-second buffer, clipped to preserve the gaze interval. The remaining segments retain their order and form an edited timeline, on which 32 frames are sampled uniformly anew. Original frame indices are not retained. Thus removal changes both evidence content and the positions of

sampled frames. Neutral deletion supplies a matched-duration control outside the evidence and gaze intervals; Appendix D gives its interval-selection rules.

C.3 Decoding and answer parsing

Decoding uses temperature 0, top-p 1.0, seed 42, and at most 512 new tokens. The requested answer marker is `FINAL: <label>`. If the marker is absent, the parser accepts a unique legal label in the final 50 tokens. Conflicting labels, no valid label, refusals, empty responses, and truncated outputs are counted as errors. All such outputs remain in the sample and count as errors, keeping the evaluation denominator independent of output validity and parsing success.

C.4 The fixed auxiliary sample

The 800-event sample is stratified by source corpus and positive function. Largest-remainder allocation is used with at least five events in every non-empty cell and seed 42. All four proprietary models use the same sample. This design retains source and function coverage while limiting repeated-call cost. In the abstention results, open-weight and proprietary blocks are reported separately because the former uses $1{,}994$ events and the latter uses this auxiliary sample.

C.5 Evidence-Frame Coverage

Table 10 summarizes sampled-frame coverage before evidence removal. Of $1{,}994$ events, 164 (8.2%) have zero evidence frames and $1{,}044$ (52.4%) have at least three. This measures visual coverage only: synchronous ASR/subtitle text is removed and frames are resampled on the edited timeline, so zero evidence-frame coverage does not imply identical model inputs. Excluding zero-coverage events leaves $1{,}830$ events, with reported model-macro EGF 70.6% and an accuracy change of -16.9 points. The gradient is descriptive: neither event difficulty nor evidence duration is held constant across events in the three evidence-frame coverage strata.

Table 10: Evidence-frame coverage under the shared 32-frame protocol. Events counts the paired sample; Base_E, Δ, and EGF are model-macro quantities defined as in Table 1. EGF is conditional on model-specific eligibility, so event counts alone are not its denominators.

Evidence frames	Events	Base_E	Δ	EGF
0	164	75.6	-3.9	85.8
1–2	786	77.1	-11.8	76.5
≥ 3	1,044	78.3	-20.7	66.1
All	1,994	77.6	-16.0	71.8

Table 11 and Figure 6a show lower EGF with higher evidence-frame coverage in all fifteen models. Excluding zero-coverage events checks persistence of the aggregate pattern when visual evidence is sampled.

Table 11: Per-model EGF by evidence-frame coverage. All fifteen models have lower EGF in higher-coverage strata. All reproduces Table 1; the stratum-specific eligibility sets determine aggregation within each model. Rates are percentages within the corresponding eligible event set.

Model	0 frames	1–2 frames	≥ 3 frames	All
GPT-4o	84.1	73.2	61.4	67.9
GPT-5.6 Sol	79.1	67.8	54.2	61.6
Gemini 3.1 Pro	86.2	76.1	65.5	71.4
Claude Opus 5	81.7	70.6	57.8	64.8
Qwen3-VL-2B	90.8	83.5	75.7	80.0
Qwen3-VL-8B	87.8	79.0	69.2	74.6
Qwen3-VL-8B-Thinking	83.9	73.4	62.0	68.3
InternVL3.5-2B	91.6	84.6	77.4	81.4
InternVL3.5-4B	88.7	80.7	71.6	76.6
InternVL3.5-8B	86.5	77.7	67.5	73.1
MiniCPM-V-4.6	89.4	81.8	73.2	77.9
GLM-4.6V	84.8	75.5	64.9	70.7
LLaVA-Video-7B	87.7	79.6	70.3	75.4
VideoLLaMA 3	84.1	74.1	63.0	69.1
Video-R1	80.9	69.9	57.3	64.2
Macro mean	85.8	76.5	66.1	71.8

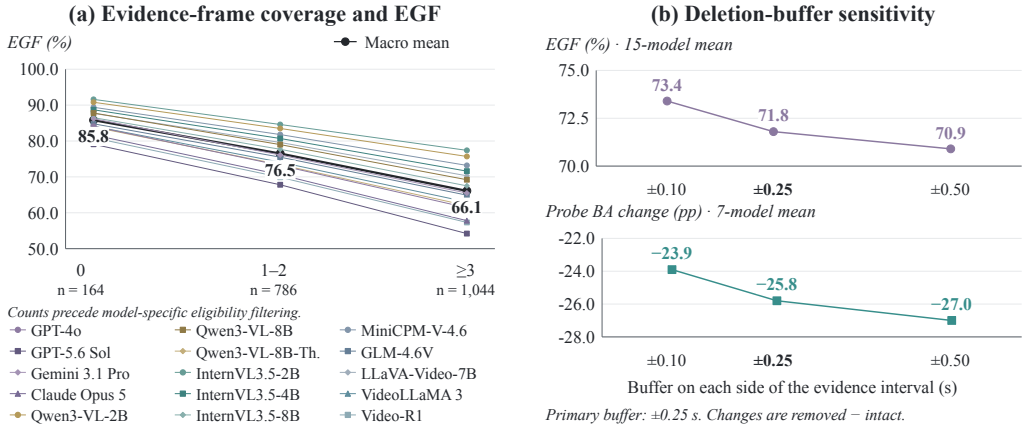

Figure 6: Coverage and deletion-buffer sensitivity. **(a)** EGF across evidence-frame strata for fifteen models. Counts precede model-specific eligibility filtering. **(b)** Buffer sensitivity of EGF (fifteen models) and sequence-end probe BA change (seven models), with the primary setting at ± 0.25 s. Buffers preserve the gaze interval.

D Prompts and Input Transformations

The blocks below reproduce the prompt library. Braced fields are event-specific substitutions for identifiers, timestamps, candidate lists, or transcript text. They are populated from the corresponding input record. Input transformations specify changes to the stimulus; they are not additional narrative descriptions of the removed evidence. Line wrapping is adjusted for the page width while preserving the wording and response formats of the supplied prompt text.

D.1 Common instructions and diagnostic questions

Aligned ASR/subtitle text is concatenated with each video prompt in the same user message, before the question or instruction block. The `[TRANSCRIPT]` field contains the text supplied for that input condition, with aligned text removed when the corresponding span is deleted. Combined hints share one transcript block; temporal shuffle omits it.

Common system prompt.

```
You are evaluating a short social-interaction video.
Use only information that is directly observable in the provided input. Do
not invent actions, dialogue, intentions, or events that are not visible or
present in the supplied transcript. Person identifiers such as ID:01 and
ID:02 are arbitrary identifiers and do not imply role, status, or speaker
order.
When the task asks for a social function, choose exactly one label from the
allowed label set. Follow the requested output format exactly. The final
answer must appear on a line beginning with:
FINAL:
```

The six positive labels are referential-communicative, social-monitoring, object-access, joint-attention-building, exploratory-scanning, and action-preparation. The prompts reproduce these names to specify the exact response vocabulary used for the positive-function task.

The full benchmark adds the seventh label `no-observable-uptake`.

L1: gaze-target perception.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

TASK: Gaze target identification
Highlighted gaze interval: {GAZE_START} to {GAZE_END}
Question: Who or what is ID:{GAZER_ID} attending to during this gaze
interval?
Visible candidate targets:
{TARGET_OPTION_LIST}
Output format:
FINAL: <candidate identifier>
```

For an object target, the output uses the object label in place of a person identifier:

```
FINAL: OBJ:<object label>
```

L2: target-conditioned evidence localization. The known target below is the annotation, not an L1 prediction. This diagnostic asks where the function-defining evidence is, conditional on that target.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

TASK: Evidence interval localization
Gaze interval: {GAZE_START} to {GAZE_END}
Known gaze target: {TARGET_ID}
Question: What is the minimal time interval AFTER the gaze-establishing
interval that contains the externally observable evidence needed to determine
what this gaze was for?
Choose the earliest start and earliest end at which the function becomes
decidable. Do not extend the interval to the remainder of the clip. The
interval must not overlap the gaze-establishing interval.
If two separated intervals are required, return at most two.
One interval:
FINAL: {START_SEC}-{END_SEC}
Two intervals:
FINAL: {START1_SEC}-{END1_SEC}; {START2_SEC}-{END2_SEC}
```

L2 correctness requires temporal IoU of at least 0.50; invalid intervals receive zero.

L3: social gaze function, paired positive subset.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

TASK: Social gaze function

Consider ID:{GAZER_ID}'s highlighted gaze event.

Which function best describes the role of this gaze in the interaction?

Allowed labels:
1. referential-communicative
2. social-monitoring
3. object-access
4. joint-attention-building
5. exploratory-scanning
6. action-preparation

FINAL: <one allowed label>
```

L3: full benchmark.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

TASK: Social gaze function
Consider ID:{GAZER_ID}'s highlighted gaze event.
Allowed labels:
1. referential-communicative
2. social-monitoring
3. object-access
4. joint-attention-building
5. exploratory-scanning
6. action-preparation
7. no-observable-uptake
Use 'no-observable-uptake' only when the recipient remains visibly available
for at least 2.0 seconds after the gaze and no observable response or uptake
occurs in that opportunity window.
FINAL: <one allowed label>
```

D.2 Intervention blocks and input transformations

GT target supplied.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Ground-truth gaze target:
ID:{TARGET_ID}

Treat this target as known. Determine the function of the gaze toward this
target.

FINAL: <one allowed label>
```

Oracle evidence interval supplied.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Oracle evidence interval:
{EVIDENCE_START} to {EVIDENCE_END}
This interval marks when the function-defining evidence occurs. It does not
state what the evidence means. Use only observable content in this interval.
```

Target + interval supplied. The combined condition supplies both the GT-target and oracle-interval blocks, preserving their respective roles as target and temporal hints.

Evidence removal. Delete each adjudicated evidence span with a ± 0.25 s buffer. Do not delete frames from the gaze-establishing interval. Delete aligned ASR/subtitle text from the same span; no raw audio is input. Resample 32 RGB frames uniformly on the edited timeline, rather than retaining original frame indices. Preserve temporal order and person identifiers.

Neutral deletion. Delete an interval outside both the gaze-establishing interval and every annotated evidence span. Match the evidence-deletion duration to within one sampled frame, and avoid a shot boundary whenever a valid alternative exists.

If several intervals satisfy the criteria, use the temporally nearest interval after the evidence span; otherwise use the nearest valid preceding interval.

Temporal shuffle. Use exactly the same sampled frames as in the intact input, but randomly permute their temporal order using a fixed event-level seed shared across all models. This condition omits aligned ASR/subtitle text; no raw audio is input in any model condition.

Head/eye masking. For frames with a valid gazer track, mask the upper 55% of the tracked head box, with 5% horizontal padding on each side. The evaluation log flags frames without a reliable track, and these frames are retained without masking or additional alteration.

Sanitized text-only. Sanitized dialogue is supplied without video frames. As in baseline L3 and CD, the prompt gives only the gazer identifier and gaze interval, without a GT target or an L1 target prediction. Among L3 protocols, target information is supplied only by the GT-target and oracle conditions.

```
Based only on the sanitized dialogue below, choose the most likely social
gaze function for ID:{GAZER_ID} during {GAZE_START}-{GAZE_END} seconds.
[TRANSCRIPT]
{SANITIZED_TRANSCRIPT}
Allowed labels: referential-communicative social-monitoring object-access
joint-attention-building exploratory-scanning action-preparation
FINAL: <one allowed label>
```

Sanitization removes explicit gaze verbs, gaze-direction descriptions, target-revealing identity substitutions, and nonverbal annotations that reveal evidence from the removed interval.

Identity shuffle. Apply one random bijection to all person identifiers in an event. Use the same mapping in visual overlays, transcript references, target hints, and question text.

D.3 Four prompting conditions

Direct baseline. The direct condition uses the six-option L3 baseline reproduced above, with no target hint and no intermediate reasoning fields. L1 and L2 outputs do not enter this prompt. The gated variants below introduce their own requested steps within a function query; they do not reuse answers from the independent diagnostic evaluations.

Placebo-gated.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Complete the following three fields.
STEP 1 : TARGET CHECK
TARGET: <identifier>
STEP 2 : CONTROL INTERVAL
Inspect only the pre-selected control interval:
{NEUTRAL_START}-{NEUTRAL_END}
Describe only directly observable events in this interval.
Do not infer intent, purpose, or mental state.
CONTROL_OBSERVATION: <one or two concise factual sentences>
STEP 3 : FUNCTION
FINAL: <one allowed label>
```

Evidence-gated.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Complete the following three fields.
STEP 1 : TARGET CHECK
TARGET: <identifier>
STEP 2 : EVIDENCE LOCALIZATION
Find the minimal interval after the gaze-establishing interval that
contains the observable evidence needed to determine what the gaze was for.
Do not infer intent in this step.
INTERVAL: <start>-<end>
OBSERVATION: <one or two concise sentences describing only directly
observable events inside that interval>
STEP 3 : FUNCTION
FINAL: <one allowed label>
```

Oracle target + interval.

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Ground-truth target:
{TARGET_ID}
Ground-truth evidence interval:
{EVIDENCE_START}-{EVIDENCE_END}
Inspect that interval and describe only directly observable content.
OBSERVATION: <one or two concise factual sentences>
FINAL: <one allowed label>
```

D.4 Explicit abstention variant

The system instruction defines insufficient evidence for the additional abstention option:

```
The option 'cannot-be-determined-from-the-video' is a valid answer. Use it
only when the supplied input lacks enough observable information to
distinguish the six function labels. Do not use it merely because the scene
is difficult.
```

The user-facing question retains the function vocabulary and adds CD as an explicit seventh answer:

```
[TRANSCRIPT]
{ALIGNED_ASR_OR_SUBTITLE_TEXT}

Which function best describes ID:{GAZER_ID}'s gaze during
{GAZE_START}-{GAZE_END} seconds?
Allowed labels:
1. referential-communicative
2. social-monitoring
3. object-access
4. joint-attention-building
5. exploratory-scanning
6. action-preparation
7. cannot-be-determined-from-the-video
FINAL: <one allowed label>
```

The same seven-option prompt is used for intact, neutral-deletion, and evidence-removal inputs.

E Determinism Audit

E.1 Execution configurations

An answer change under an intervention is interpretable only in relation to the frequency of changes under identical input. The audit repeats unchanged queries for the seven diagnostic models while progressively controlling the execution environment. Production serving with continuous batching gives a macro flip rate of 2.1%. Fixing batch composition and order reduces it to 0.9%, a batch size of one to 0.47%, and deterministic kernels with fixed video decoding to 0.27% (Table 12). The configuration sequence reduces flip rates in each of the seven models.

Table 12: Repeated-query answer changes in the seven-model execution audit. The upper block reports macro flip rates as execution controls are introduced. The lower block gives the model-specific floors in the final configuration. All entries are percentages. Flip rates characterize execution stability and are not used to correct repair rates. All seven models use the same audit sequence.

Execution configuration	Flip rate
Production serving with continuous batching	2.10
Fixed batch composition and order	0.90
Batch size of one	0.47
Deterministic kernels and fixed video decoding	0.27
Model-specific floors under the final configuration	
Qwen3-VL-2B	0.2
Qwen3-VL-8B	0.2
Qwen3-VL-8B-Thinking	0.5
InternVL3.5-2B	0.1
InternVL3.5-8B	0.2
MiniCPM-V-4.6	0.4
LLaVA-Video-7B	0.3

E.2 Relation to repair rates

The main diagnostic experiments use the final controlled configuration. An unchanged-query label-flip rate includes transitions between wrong labels as well as transitions into or out of the correct label. It therefore does not estimate the error-to-correct probability among baseline errors. The main text reports raw oracle-hint repair. A separate repeat-query count on baseline errors estimates error-to-correct transitions; these conditional floors, not the overall label-flip rates above, enter the sensitivity adjustment in Appendix F.

F Oracle-Hint Repair

F.1 Repair conditions and operational categories

The unit of analysis is a baseline error for a particular model and event. Three subsequent queries supply the ground-truth target, the oracle evidence interval, or both. The interval identifies when evidence occurs without naming the correct function or describing the evidence's meaning. Let T, I, and B denote correct answers with target, interval, and combined hints.

Table 13 reports raw repair rates. For the conceptual outcome diagram in Figure 3b, the operational categories are:

$$\text{L1-specific}: \quad B \wedge T \wedge \neg I, \tag{3}$$

$$\text{L2-specific}: \quad B \wedge \neg T \wedge I, \tag{4}$$

$$\text{Shared / non-identifiable}: \quad B \wedge \neg(T \oplus I), \tag{5}$$

$$\text{Combined-hint residual}: \quad \neg B. \tag{6}$$

Here $\oplus$ denotes exclusive OR. The shared category includes both single hints succeeding and neither single hint succeeding when their combination does succeed. The latter is a combined-hint rescue; assigning it to a single stage would depend on hint order.

Table 13: Oracle-hint repair on baseline errors, averaged equally over seven models. Raw rates are primary; adjusted rates use each model's conditional error-to-correct floor for sensitivity analysis.

Repair condition	Raw (%)	Adjusted (%)
Ground-truth target only	19.0	18.8
Oracle evidence interval only	21.1	20.9
Both hints	44.1	43.9

F.2 Integer records and sensitivity adjustment

Table 14 gives the joint outcomes on 4,786 baseline model-event errors. Bits are ordered (T, I, B), with 1 denoting a correct hinted answer. The repeat-query column counts baseline errors becoming correct under an unchanged query; its denominator is the same model's baseline-error count n_m. The sum $13/4{,}786$ is a pooled floor, not the value used for every model.

Table 14: Joint hint outcomes and unchanged-query error-to-correct counts. Each row's eight cells sum to n_m; T denotes the corresponding thinking variant.

Model	n_m	000	001	010	011	100	101	110	111	Repeat E→C
Qwen3-VL-2B	810	369	131	65	95	13	123	9	5	2/810
Qwen3-VL-8B	610	294	84	9	100	41	58	1	23	1/610
Qwen3-VL-8B-T	562	239	99	17	96	48	53	7	3	3/562
InternVL3.5-2B	824	385	108	59	110	13	143	1	5	1/824
InternVL3.5-8B	598	279	109	41	65	9	82	4	9	1/598
MiniCPM-V-4.6	666	301	126	10	101	57	44	2	25	3/666
LLaVA-Video-7B	716	340	131	11	103	44	53	8	26	2/716
Pooled	4,786	2,207	788	212	670	225	556	32	96	13/4,786

For model m, let r_m be a raw repair proportion and $f_m = k_m/n_m$ its unchanged-query error-to-correct proportion. The adjustment is

$$r_m^{\mathrm{corr}} = \frac{r_m - f_m}{1 - f_m}, \qquad \overline{r}^{\mathrm{corr}} = \frac{1}{7}\sum_m r_m^{\mathrm{corr}}. \tag{7}$$

It assumes $r_m = f_m + (1 - f_m)r_m^{\mathrm{corr}}$: the background false-repair probability transfers from unchanged to hinted queries. This is a sensitivity model, not an identified causal correction. Applying it before macro averaging yields 18.8%, 20.9%, and 43.9%, versus raw 19.0%, 21.1%, and 44.1%. The main text retains the raw rates.

F.3 Operational partition and non-monotonic trajectories

The joint records determine the raw partition in Table 15. They do not identify the causal stage responsible for an error. We do not apply a separate floor subtraction to each category, which would not preserve the partition; no confidence intervals are inferred from these aggregate counts.

Table 15: Raw operational outcomes on baseline errors. Counts pool model-event cells; macro percentages weight the seven models equally. Rounding can affect the sum of displayed percentages.

Category	Count	Pooled (%)	Macro (%)
L1-specific (101)	556	11.6	11.3
L2-specific (011)	670	14.0	14.1
Shared / non-identifiable (001, 111)	884	18.5	18.6
Combined-hint residual ($B = 0$)	2,676	55.9	55.9

Non-monotonic trajectories satisfy $B = 0$ and $(T = 1 \vee I = 1)$, namely cells 010, 100, and 110. Their count is $212 + 225 + 32 = 469$, giving $469/4{,}786 = 9.8\%$ pooled and 9.9% by seven-model macro averaging. They belong to the combined-hint residual along with the 2,207 all-fail (000) cells. Thus combined-hint failure does not imply failure under both single hints. Applying Equation 7 mechanically to the per-model non-monotonic rates before macro averaging gives 9.6%, but error-to-correct floors do not estimate false non-monotonic transitions. We therefore report the unadjusted non-monotonic rates only and make no model exclusions based on adjusted thresholds.

G Probe Methodology and Extended Results

G.1 Features, classifier, and fitting procedure

Probes predict one of the six positive functions from frozen hidden states. For a feature $x \in \mathbb{R}^d$, the linear classifier is $p(y \mid x) = \mathrm{softmax}(Wx + b)$. No backbone weights are updated. The classifier measures recoverability under supervised readout, rather than whether the backbone's own generation procedure uses the same information (Alain & Bengio, 2016; Belinkov, 2022).

Local visual features are mean-pooled over the specified token set; sequence-end features use one token without pooling, as defined below. Features are standardized using training-fold statistics only. Training minimizes cross-entropy with inverse-frequency class weights computed within the training fold. AdamW (Loshchilov & Hutter, 2019) uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and gradient clipping at norm 1.0. Training runs for at most 100 epochs, with early-stopping patience 10, minimum improvement 0.0005, and balanced accuracy as the selection metric. Training holds out 10% for validation and uses seeds $13, 29, 47, 71, 101$.

Balanced accuracy averages the recall of each of the C evaluated function classes:

$$\mathrm{BA} = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_i \mathbf{1}[y_i = c, \hat{y}_i = c]}{\sum_i \mathbf{1}[y_i = c]}. \tag{8}$$

For the paired positive task, $C = 6$, giving 16.7% for uniform random prediction. Matching probe and judgment BA removes a difference in class weighting from their comparison. It does not equate their training information, readout capacity, or conditional sample selection in other tables.

G.2 Readout locations and controls

The local evidence probe averages visual tokens within the annotated evidence interval at relative depth 0.75. The onset control pools tokens from the gaze interval at the same depth, and the global control pools all sampled visual tokens. Thus the three readouts share a video and network depth while varying the temporal region represented in the feature.

At relative depth 0.95, the sequence-end feature is the hidden state of the last non-padding input token immediately before generation begins. No token pooling or generated-answer tokens enter this readout. For each model, fold, and seed, a single probe is trained on the intact training split only and then frozen to evaluate held-out intact, evidence-removal, and neutral-deletion inputs. The same trained classifier is used for all three conditions; condition-specific retraining is not performed. Thus the paired comparison changes the input while holding the supervised readout fixed.

Evidence pooling exceeds both global and onset pooling in all seven models (Table 3a; Figure 7a). The comparison localizes the decoding advantage to evidence-bearing intervals within the same video.

For the permutation control, training labels are permuted across training events, features and fold membership are held fixed, and the classifier is refitted before evaluating on unchanged test labels. This is a training-label control, informed by work on probe selectivity and classifier permutation tests (Hewitt & Liang, 2019; Ojala & Garriga, 2010). Selectivity is defined here as true-label BA minus permuted-label BA. This subtraction measures the decoding advantage over randomized training labels; it does not estimate mutual information or yield a permutation-test significance level.

G.3 Paired changes and neutral deletion

Table 16: Extended measurements on the seven-model diagnostic subset. F1 is local-probe macro-F1; Sel. is true-label minus permuted-label BA. Neutral deletes a matched-duration interval outside both the evidence and gaze windows. Changes are evidence-removed minus intact; DiD subtracts judgment change from probe change. Both readouts share the evaluated events.

Model	Local probe		Neutral BA		Δ_{probe}	Δ_{judg}	DiD
	F1	Sel.	Probe	Judg.			
Qwen3-VL-2B	52.1	41.0	54.4	58.3	−24.2	−10.5	−13.7
Qwen3-VL-8B	59.9	47.9	58.1	68.2	−24.7	−13.9	−10.8
Qwen3-VL-8B-Thinking	63.2	51.8	62.5	70.7	−28.1	−19.1	−9.0
InternVL3.5-2B	51.9	39.9	53.4	57.7	−24.2	−9.4	−14.8
InternVL3.5-8B	60.7	49.7	60.0	68.7	−27.6	−15.0	−12.6
MiniCPM-V-4.6	58.7	44.6	57.1	65.6	−26.3	−12.2	−14.1
LLaVA-Video-7B	64.4	53.7	63.2	63.1	−25.4	−12.8	−12.6
Macro mean	58.7	46.9	58.4	64.6	−25.8	−13.3	−12.5

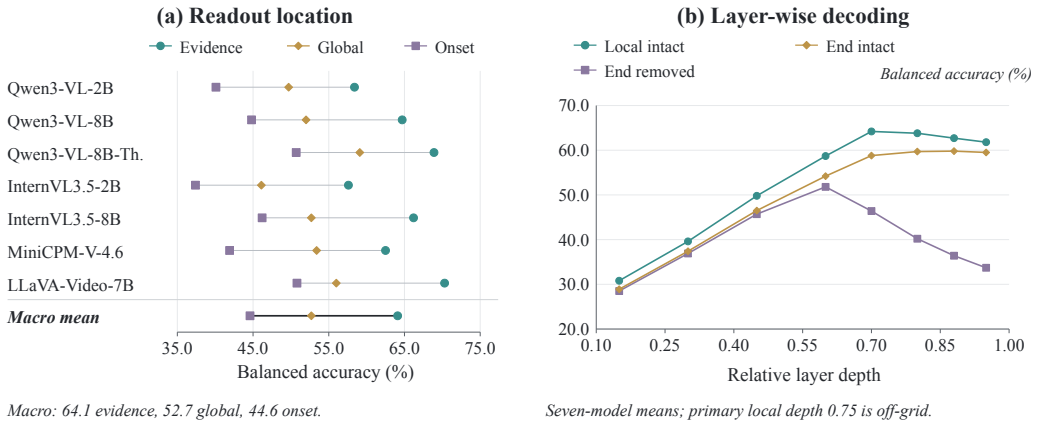

Figure 7: Readout controls and layer-wise decoding. **(a)** Evidence, global, and onset pooling at relative depth 0.75 for seven models and their macro mean. Lines connect readout conditions within each model. **(b)** Seven-model mean BA over eight measured depths for local intact and sequence-end intact/removed states. The primary local depth is absent from this grid; these profiles measure representation recoverability, not causal feature routing.

In Table 16, the sequence-end probe falls from 59.5% to 33.7% BA after evidence removal; judgments fall from 65.2% to 51.9% on the same 1,994 events. For each model, define

$$\mathrm{DiD}_m = (\mathrm{BA}_{P,R,m} - \mathrm{BA}_{P,I,m}) - (\mathrm{BA}_{J,R,m} - \mathrm{BA}_{J,I,m}). \tag{9}$$

Here P and J denote probe and judgment readouts, and I and R denote intact and removed inputs. Figure 8a compares the paired changes. The mean contrast is -12.5 points, with all seven values between -14.8 and -9.0. The local intact score of 64.1% is below judgment BA of 65.2%; the result concerns differences in response to removal, not superior intact performance of the supervised probe.

To reduce dependence on the number of configurations per family, we also average the displayed contrasts within Qwen3-VL (three configurations), InternVL3.5 (two), MiniCPM-V (one), and LLaVA-Video (one). The family means are -11.2, -13.7, -14.1, and -12.6 points, respectively; weighting these four families equally gives -12.9 points. This descriptive reaggregation shows that the negative contrast is not driven by oversampling one family. Using the reported model-macro neutral changes, the contrast for evidence versus neutral deletion is $-12.5 - (-1.1 - (-0.6)) = -12.0$ points. Sample-level uncertainty is assessed below with paired source-video-clustered bootstrap resampling.

Seven negative model contrasts give a nominal two-sided exact sign-test value $2/2^7 = 0.015625$ under independent configurations with equally likely signs. Shared model families limit that independence assumption. The test is a directional configuration-level check; it does not supply an event-level confidence interval for the mean contrast. Interpretation rests on the paired cluster-bootstrap intervals, per-model contrasts, and neutral-deletion and source-grouped controls; the nominal sign test is supplementary.

Source-video-clustered paired bootstrap. The 1,994 evaluated events form 1,276 source-video clusters. Within each corpus, source videos are sampled with replacement. All events, input variants, and models belonging to a sampled video are resampled together, preserving the pairing of probe and judgment predictions and of intact, removal, and neutral conditions. Each replicate recomputes six-class BA, forms each model's change and DiD, and then averages the seven models equally. We use 10,000 replicates and percentile 95% intervals (Table 17). The neutral-adjusted DiD subtracts the probe-minus-judgment change under neutral deletion from that under evidence removal. Resampling uses the existing predictions; these intervals are conditional on the fitted probes and evaluated model configurations and do not include probe-refitting variability.

Table 17: Source-video-clustered paired bootstrap for the seven-model macro contrasts. All entries are percentage points; intervals use $10{,}000$ within-corpus cluster resamples and percentile limits.

Statistic	Point estimate	95% CI
Probe ΔBA	-25.8	$[-27.2, -24.4]$
Judgment ΔBA	-13.3	$[-14.5, -12.1]$
DiD	-12.5	$[-13.9, -11.1]$
Neutral-adjusted DiD	-12.0	$[-13.5, -10.6]$

Table 18: Layer-wise balanced accuracy, averaged across the seven diagnostic models. A reads local evidence tokens; B and C read sequence-end states before and after evidence removal. The primary local readout at depth 0.75 is a separate measurement, absent from this grid. B minus C is the removal-associated loss in balanced accuracy (percentage points).

Relative depth	A: local intact	B: end intact	C: end removed	B $-$ C
0.15	30.8	28.9	28.5	0.4
0.30	39.6	37.4	36.9	0.5
0.45	49.8	46.5	45.7	0.8
0.60	58.7	54.2	51.8	2.4
0.70	64.2	58.8	46.4	12.4
0.80	63.8	59.7	40.2	19.5
0.88	62.7	59.8	36.4	23.4
0.95	61.8	59.5	33.7	25.8

G.4 Layer-wise decoding

Figure 7b shows the depth profiles. The sequence-end removal gaps are 0.4, 0.5, and 0.8 points through depth 0.45, 2.4 at 0.60, and 25.8 at 0.95. Local decoding peaks near 0.70 and remains 61.8% at 0.95 (Table 18). These profiles describe recoverability, not causal routing of a feature.

G.5 SOURCE-GROUPED VALIDATION

Event-grouped folds use `event_id`, keeping all versions of an event together. Source-grouped folds instead use `source_video_id`; the paired set contains 1,276 such groups. Both use five-fold `StratifiedGroupKFold` with shuffling and the same five seeds. Class balancing is constrained by indivisible source-video groups (scikit-learn developers, 2026).

Table 19: Source-video-grouped validation, averaged over five seeds. Local probes use depth 0.75 and sequence-end probes use 0.95. Event and Source denote the grouping schemes; Perm. and Sel. use source grouping. Changes are removed minus intact BA in points. All 1,276 source-video groups stay disjoint across every training and test fold.

Model	Local evidence probe (depth 0.75)				Sequence-end probe (depth 0.95)			
	BA		Perm.	Sel.	Source		Δ_{probe}	
	Event	Source			Intact	Evid. rm	Event	Source
Qwen3-VL-2B	58.4	54.9	17.2	37.7	51.8	29.1	−24.2	−22.7
Qwen3-VL-8B	64.7	60.8	16.8	44.0	56.3	32.4	−24.7	−23.9
Qwen3-VL-8B-Thinking	68.9	64.7	17.0	47.7	60.2	34.0	−28.1	−26.2
InternVL3.5-2B	57.6	53.8	17.6	36.2	51.0	28.2	−24.2	−22.8
InternVL3.5-8B	66.2	62.0	16.4	45.6	58.0	32.3	−27.6	−25.7
MiniCPM-V-4.6	62.5	58.6	17.5	41.1	54.7	29.8	−26.3	−24.9
LLaVA-Video-7B	70.3	65.9	16.7	49.2	60.9	36.8	−25.4	−24.1
Macro mean	64.1	60.1	17.0	43.1	56.1	31.8	−25.8	−24.3

Table 19 shows that source grouping lowers local BA from 64.1% to 60.1% but preserves a 24.3-point sequence-end removal drop, versus 25.8 with event grouping. All seven models retain this response (Figure 8b). The local-score decrease reflects both source separation and altered fold composition, so it cannot quantify information attributable to shared recordings.

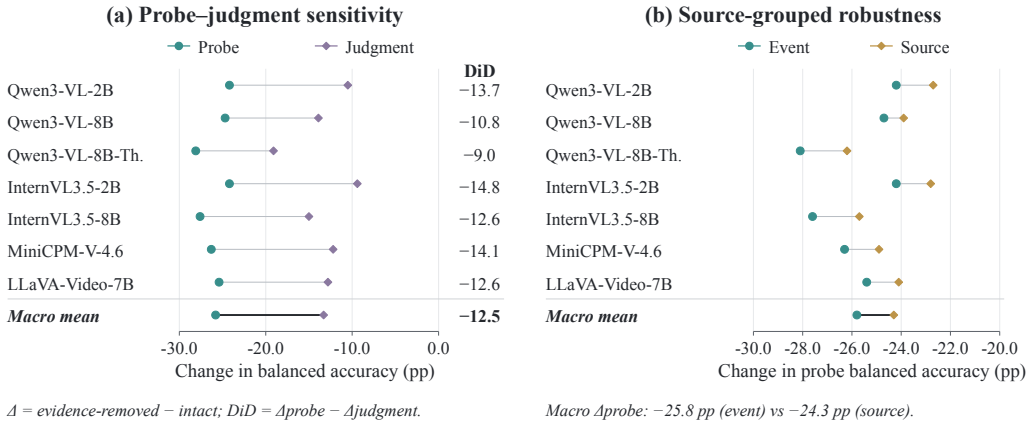

Figure 8: Paired sensitivity and source-grouped validation. **(a)** Probe and judgment changes use matched events and balanced accuracy; DiD is probe minus judgment change. All seven contrasts are negative. **(b)** Sequence-end probe changes under event- and source-grouped folds, with source-video groups disjoint across training and test. Changes are removed minus intact in percentage points; connecting segments show paired measurements, not confidence intervals.

G.6 LEAVE-ONE-CORPUS-OUT VALIDATION

We additionally train on three corpora and test on the fourth (Table 20). Local BA exceeds the permutation control on all four held-out corpora, while evidence removal lowers sequence-end BA by 18.3 to 22.6 points. The pattern persists under the resulting shifts in label and video distributions.

Table 20: Leave-one-corpus-out validation, with model-macro means over the seven diagnostic models. Classes is the number of positive functions present in the test corpus; BA averages recall over those classes. Perm. is the empirical label-permutation control, reported directly without inferring a random-guess baseline from test-label support. The final row weights corpora equally; weighting the displayed local BA values by event count instead gives approximately 55.9%.

Held-out corpus	Events	Classes	Local probe		Sequence-end probe		
			BA	Perm.	Intact	Rm	Δ
VACATION	1,130	6	58.1	16.9	54.2	31.6	-22.6
VideoAttentionTarget	488	6	54.7	16.5	51.3	30.4	-20.9
VideoCoAtt	215	5	49.8	20.4	47.1	28.8	-18.3
VSGaze	161	5	52.6	19.7	49.4	27.1	-22.3
Mean over corpora	1,994	-	53.8	18.4	50.5	29.5	-21.0

H Evidence-Focused Prompting

H.1 Structural controls and response budgets

The evidence-gated condition asks for a target check, a minimal evidence interval with a factual observation, and a function judgment. The observation step prohibits intent inference. The placebo retains a three-step structure but replaces evidence localization with a description of a pre-selected neutral interval. The oracle supplies the ground-truth target and interval, leaving the model to describe the content and infer its function. Exact prompt blocks appear in Appendix D. These are controlled behavioural interventions: improvements do not by themselves establish an internal causal mechanism or effectiveness outside this evaluation protocol.

The mean completion length is 127 tokens for placebo and 133 for evidence gating, with 2.95 and 2.97 of the three requested steps completed on average. Their step non-compliance rates are 3.6% and 4.9%. These are descriptive checks of comparable structure and output length, rather than a formal equivalence test. Non-compliant responses remain part of the evaluation, and the observed differences in compliance are retained when comparing the conditions.

H.2 Evidence specificity and paired contrasts

All conditions retain each model's fixed baseline-eligible event set. Write S_E for label stability after evidence removal and S_N for stability after neutral deletion. Specificity is $S_N - S_E$: it is large when answers change more under evidence deletion than under a duration-matched control. The measure helps distinguish evidence sensitivity from a generic tendency to switch labels, although neither kind of change alone establishes correctness.

Table 21: Prompting controls on fixed baseline-eligible sets, macro-averaged over seven models. S_E/S_N denote evidence/neutral-deletion stability; Spec. is their difference. Tokens and Non-compl. describe completion length and step compliance. The lower block gives paired contrasts with 95% event-bootstrap confidence intervals ($10{,}000$ resamples).

Condition	Intact acc.	S_E	S_N	Spec.	Non-compl.	Tokens
Direct baseline	74.3	75.8	93.4	17.6	-	84
Placebo-gated	75.0	72.6	92.0	19.4	3.6	127
Evidence-gated	76.5	63.1	90.8	27.7	4.9	133
Oracle	77.6	57.6	89.7	32.1	2.1	119
Paired contrasts (second line: 95% CI)						
Placebo vs. Direct	$+0.7$	-3.2	-1.4	$+1.8$	—	—
	$[+0.2, +1.2]$	$[-4.6, -1.8]$	$[-2.5, -0.4]$	$[+0.4, +3.2]$	—	—
Gated vs. Placebo	$+1.5$	-9.5	-1.2	$+8.3$	—	—
	$[+0.9, +2.1]$	$[-11.8, -7.3]$	$[-2.4, +0.1]$	$[+6.2, +10.5]$	—	—
Oracle vs. Gated	$+1.1$	-5.5	-1.1	$+4.4$	—	—
	$[+0.4, +1.8]$	$[-7.0, -4.0]$	$[-2.3, +0.1]$	$[+2.7, +6.2]$	—	—

In Table 21, evidence gating improves intact accuracy relative to placebo by 1.5 points and specificity by 8.3 points. The 95% paired event-bootstrap intervals are $[0.9, 2.1]$ and $[6.2, 10.5]$, respectively, using 10,000 resamples. Neutral-deletion stability changes by -1.2 points, with interval $[-2.4, +0.1]$, while evidence-deletion stability falls by 9.5 points. The larger change under evidence deletion indicates increased specificity, beyond a general tendency to switch answers.

H.3 Per-model outcomes and denominators

Table 22: Per-model accuracy and specificity on fixed eligible events. D, P, G, and O denote direct, placebo, evidence-gated, and oracle conditions. Specificity is neutral minus evidence-deletion stability in points. The final row averages the seven models equally; human responses are excluded.

Model	Intact acc.				Spec.			
	D	P	G	O	D	P	G	O
Qwen3-VL-2B	70.7	71.1	72.9	74.2	14.6	15.8	24.6	29.4
Qwen3-VL-8B	76.2	77.1	78.4	79.3	17.9	20.2	28.1	32.2
Qwen3-VL-8B-Thinking	81.0	81.5	83.1	84.1	20.7	22.2	30.7	35.3
InternVL3.5-2B	67.3	68.3	70.0	71.4	12.8	14.9	24.1	29.1
InternVL3.5-8B	77.2	77.8	79.2	80.2	18.9	20.6	28.3	32.3
MiniCPM-V-4.6	75.8	76.6	78.1	79.3	17.1	19.1	27.2	31.5
LLaVA-Video-7B	71.9	72.6	73.8	74.7	21.2	23.0	30.9	34.9
Macro mean	74.3	75.0	76.5	77.6	17.6	19.4	27.7	32.1

Table 22 and Figure 9 report per-model outcomes: relative to placebo, gating improves accuracy by 1.2 to 1.8 points and specificity by 7.7 to 9.2 points in all seven models. The oracle retains a further 5.5-point decrease in evidence stability beyond the evidence-gated condition.

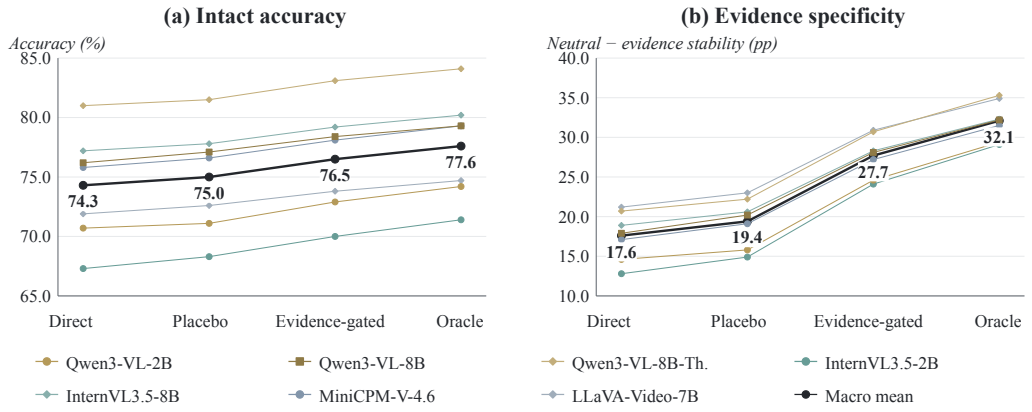

Figure 9: Per-model prompting outcomes on fixed baseline-eligible events. **(a)** Intact accuracy and **(b)** evidence specificity across direct, placebo, evidence-gated, and oracle conditions. Specificity subtracts evidence-deletion stability from neutral-deletion stability. Thin curves show seven models; black curves and labels show their unweighted means. Lines connect discrete prompting conditions and do not represent a continuous intervention dose.

I Residual Error Structure

I.1 Localization quality

Predicted evidence intervals reach macro mean tIoU 0.57 against the annotation. Invalid and non-compliant interval predictions receive zero. Mean human inter-annotator tIoU is 0.64, which is a reference for annotation agreement rather than a model-performance ceiling. Interval overlap evaluates whether the model found the right temporal support; it does not evaluate its explanation of that support.

Table 23 reports the four-way counts on the same fixed baseline-eligible model-event cells used in the prompting comparison. Pooling the seven models gives $8{,}468$ cells: localization succeeds (tIoU ≥ 0.5) in $5{,}808/8{,}468 = 68.6\%$. Function accuracy is $4{,}565/5{,}808 = 78.6\%$ when localization succeeds and $1{,}948/2{,}660 = 73.2\%$ when it fails. There are $1{,}955$ function errors, of which $1{,}243/1{,}955 = 63.6\%$ occur despite correct localization. The remaining 36.4% occur after localization failure.

These localization and error-group statistics pool model-event cells. Pooled overall function accuracy is $6{,}513/8{,}468 = 76.9\%$; the 76.5% in the prompting tables instead averages the seven per-model accuracies equally. Different model-specific eligibility counts explain the difference. An error share conditions on an incorrect function prediction, whereas each conditional accuracy uses all cells in its localization group.

Table 23: Evidence-gated residual counts on fixed baseline-eligible cells. L+/L− denote successful/failed localization at tIoU 0.5; F+/F− denote correct/incorrect function predictions. The pooled row sums model-event cells; prompting macro accuracy gives each model equal weight.

Model	n	L+ F+	L+ F−	L− F+	L− F−
Qwen3-VL-2B	1,015	465	165	275	110
Qwen3-VL-8B	1,308	730	190	295	93
Qwen3-VL-8B-Thinking	1,372	875	135	265	97
InternVL3.5-2B	993	405	185	290	113
InternVL3.5-8B	1,316	749	181	293	93
MiniCPM-V-4.6	1,218	659	171	292	96
LLaVA-Video-7B	1,246	682	216	238	110
Pooled	8,468	4,565	1,243	1,948	712

I.2 Prior-consistency rates

Let $\hat{y}_G$ be the gated function prediction, $\hat{y}_T$ the sanitized text-only prediction, and y the true label. Within localization group L, full-group PCR conditions on an incorrect gated answer:

$$\mathrm{PCR}_L = P(\hat{y}_G = \hat{y}_T \mid L, \hat{y}_G \neq y), \tag{10}$$

while the conditional rate adds $\hat{y}_T \neq y$ to the conditioning set. Table 24 compares these rates; the added condition isolates agreement between two wrong predictions. For probe errors, the probe prediction replaces $\hat{y}_G$ in the corresponding comparison.

Table 24: Prior consistency among incorrect predictions. Full-group PCR uses all errors in the named group; conditional PCR additionally requires an incorrect text-only prediction. Values are percentages. The probe row uses held-out probe errors and reports only the conditional statistic. Human responses do not enter these groups. The two PCR columns use different denominators.

Error group	Full-group PCR	Conditional PCR
Correct evidence localization	61.7	72.4
Failed evidence localization	49.8	58.6
Probe errors	—	24.6
Uniform wrong-label reference	—	20.0

I.3 Wrong-label entropy

For true class c in error group g, let $p_{g,c,j}$ be the empirical frequency of wrong label j. With six positive labels, there are five possible wrong labels, and normalized entropy is

$$h_{g,c} = \frac{-\sum_{j=1}^{5} p_{g,c,j} \log p_{g,c,j}}{\log 5}. \tag{11}$$

The convention is $0 \log 0 = 0$. Scores are macro-averaged over true classes with at least ten errors. A value near one indicates a diffuse distribution among the wrong labels; a lower value indicates

concentration on fewer alternatives. Because only errors enter this calculation, entropy describes their distribution across wrong labels rather than overall prediction accuracy.

Table 25: Concentration of wrong-label predictions. Entropy is normalized by $\log 5$ and macro-averaged over true classes with at least ten errors. The final row contrasts localization failure with correct localization; its interval comes from 10,000 event-level bootstrap resamples stratified by corpus and true label. Only incorrect function predictions contribute to each entropy estimate.

Error group / contrast	Normalized entropy	95% CI
Failed evidence localization	0.86	—
Correct evidence localization	0.74	—
Probe errors	0.99	—
Failed minus correct localization	0.12	[0.08, 0.16]

Table 25 reports entropy estimates and their contrast; Figure 10 pairs them with PCR. Entropy is 0.86 after localization failure, 0.74 after correct localization, and 0.99 for probe errors. The failed-minus-correct difference is 0.12, with 95% interval $[0.08, 0.16]$ from event-level bootstrap resampling stratified by corpus and true label. Together with PCR, this shows that correctly localized residual errors are more concentrated and more aligned with text-only predictions. These are complementary distributional findings; entropy alone does not identify which prior or contextual feature produced the concentration. The 20% reference assumes uniform choice among five incorrect labels. It is not adjusted to observed error frequencies. Text-only and baseline L3 both omit target hints; PCR remains an agreement statistic, not a causal estimate of language-prior dependence.

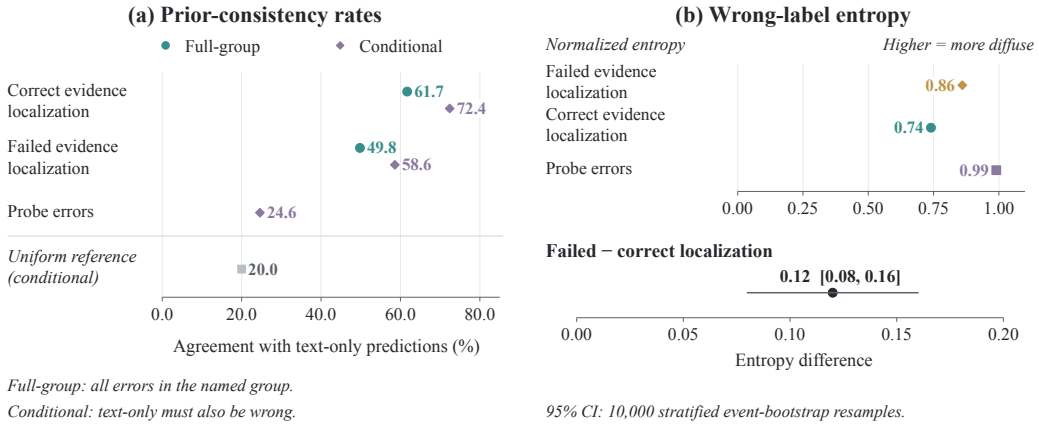

Figure 10: Residual error structure. **(a)** Full-group PCR uses all errors in each group; conditional PCR additionally requires an incorrect text-only answer. The 20% reference applies only to the conditional statistic. **(b)** Wrong-label entropy normalized by $\log 5$. Only the failed-minus-correct localization contrast carries a 95% interval, from $10{,}000$ stratified event-bootstrap resamples; individual groups have no reported intervals in the available results.

J An Explicit Abstention Option

The main six-way task requires a function even after its evidence is removed. The abstention variant adds `cannot-be-determined-from-the-video` (CD) to all three input conditions: intact, neutral deletion, and evidence removal. The option denotes insufficient evidence to distinguish functions; it is different from the benchmark label `no-observable-uptake`, which describes an observed interaction. Like baseline L3, the CD prompt supplies the gazer identifier and gaze interval without a target hint, in all three conditions (Appendix D). The answer vocabulary adds CD while retaining the six function labels.

Human participants receive this additional instruction:

> This version includes a seventh option: cannot-be-determined-from-the-video. Choose this option only when the supplied video genuinely lacks enough observable information to distinguish the six functions. Do not choose it merely because the scene is difficult.

Response-level aggregation. Human CD rates pool individual responses rather than event-level majority labels. If one of the two responses selects CD and the other selects a function, the cell contributes one CD response out of two. Two CD responses contribute two, and two function responses contribute none. The rate for condition c is computed across all valid individual responses:

$$\mathrm{CDRate}_c = \frac{\#\{\text{valid individual responses selecting CD in } c\}}{\#\{\text{all valid individual responses in } c\}}. \tag{12}$$

This statistic estimates how often participants express insufficient evidence. It does not require a tie-breaking rule for the two responses, and no human abstention accuracy or same-person stability is derived from it. Model rates use one prediction per model-event cell, without a voting stage. CD measures explicit option selection under this prompt and decoding protocol; it is not a calibrated or complete measure of uncertainty.

Table 26: Abstention rates with CD available in every condition. Open-weight models use $1{,}994$ events; proprietary models use the fixed 800-event sample. Δ_{CD} is evidence minus neutral CD. Intact accuracy counts CD as incorrect. Human CD rates pool two responses per event-condition over 560 events and are excluded from model means.

Model	Intact CD	Neutral CD	Evid. rm CD	Δ_{CD}	Acc.
Open-weight models, 1,994 events					
Qwen3-VL-2B	5.8	7.1	20.8	13.7	57.9
Qwen3-VL-8B	4.3	5.6	26.4	20.8	68.0
Qwen3-VL-8B-Thinking	3.1	4.0	31.7	27.7	70.5
InternVL3.5-2B	6.4	7.5	18.9	11.4	57.2
InternVL3.5-8B	4.0	5.2	27.8	22.6	68.5
MiniCPM-V-4.6	4.7	5.9	24.6	18.7	65.1
LLaVA-Video-7B	5.0	6.2	23.7	17.5	62.7
Macro mean	4.8	5.9	24.8	18.9	64.3
Proprietary models, 800 events					
GPT-4o	2.9	4.2	29.8	25.6	71.8
GPT-5.6 Sol	2.1	3.3	36.1	32.8	77.4
Gemini 3.1 Pro	2.6	3.8	27.4	23.6	74.9
Claude Opus 5	2.4	3.5	33.2	29.7	75.6
Macro mean	2.5	3.7	31.6	27.9	74.9
Human reference	1.8	3.9	74.6	70.7	-

In Table 26, mean CD rate for the seven open-weight models is 4.8% on intact inputs, 5.9% after neutral deletion, and 24.8% after evidence removal. Thus the evidence-minus-neutral increase is 18.9 points. The proprietary auxiliary sample gives 2.5%, 3.7%, and 31.6%, respectively, with an increase of 27.9 points. Different event pools require separate group summaries.

Figure 11a compares the model and human CD rates. The human reference gives 1.8% CD on intact clips, 3.9% after neutral deletion, and 74.6% after evidence removal. This reference makes the scale of the model response interpretable: models do increase expressed uncertainty when evidence is withdrawn, but do so substantially less often than the human participants. These CD results describe a separate task with a different response vocabulary and sample.

K Natural A/B Evidence Pairs

K.1 Matching and verification

Natural pairs compare two observed gaze events from the same source video, with the same gazer and recipient or target identities but different functional labels. Both gaze intervals pass target-visibility quality control. Their gaze durations differ by no more than 20% relative to the longer interval, and their mean head orientations differ by no more than 15 degrees. Target visibility category also matches: fully visible or partially visible.

Two annotators independently verify that subsequent interactional evidence, rather than a visibly different look, distinguishes the functions. Borderline adjudicated pairs are excluded. Pairs consist of naturally observed events without frame splicing. Table 27 tracks pair retention through matching and verification; other contextual cues can still vary.

Table 27: Cumulative construction of natural A/B pairs. Events counts unique events participating in the retained pairs; an event can occur in multiple pairs. The final set contains 246 pairs involving 421 distinct events. Final verification checks gaze matching and the evidential distinction.

Filter	Pairs	Events
Same source video	8,742	4,306
+ same dyad	2,136	2,487
+ gaze matched	846	1,208
+ labels differ	312	519
+ independently verified	246	421

K.2 Human presentation and outcome definitions

Model predictions are obtained from independent per-clip queries. Human participants instead see paired events, so the presentation protocols differ. The separate 24-person cohort receives the following instructions:

> You will see two naturally occurring gaze events from the same source video. They involve the same gazer and recipient and have been matched for gaze configuration. The two events may serve different functions. Judge each event independently. Do not assume the two answers should be the same or different.

Each intact pair receives three independent judgments, and A/B order is balanced equally. Neutral-deletion presentation uses a separate counterbalanced assignment, avoiding exposure to both intact and altered versions of a pair for the same participant. The instructions explicitly avoid prescribing an answer difference, even though the retained analysis pairs have distinct ground-truth functions.

Both correct requires a correct function for each member. Same label counts an identical prediction for the two members despite different true functions. Neutral stability records label retention after a matched-duration deletion outside the relevant gaze and evidence intervals. These quantities describe complementary aspects of behaviour: same-label answers necessarily include an error on different-function pairs, whereas different labels alone do not establish that either answer is correct. Same-label rates also depend on marginal label preferences, so we interpret them jointly with both-correct rates rather than as a standalone measure of evidence use. We do not report a collision baseline estimated from class marginals, and these rates are not adjusted for label-frequency effects.

K.3 Results and scope of the comparison

Table 28 and Figure 11b report the natural-pair comparison. Across the fifteen models, both answers are correct in 44.4% of pairs and the same label is returned in 34.7%. The corresponding human rates are 86.6% and 4.9%. Neutral-deletion stability is 91.6% for models and 96.1% for humans. The human reference is excluded from the model macro mean. The human reference describes performance with paired presentation.

Table 28: Descriptive behaviour on the 246 natural A/B pairs. Both correct requires correct labels for both members. Same label counts identical predicted labels despite different ground truth. Neutral stab. measures label stability after deleting a matched-duration neutral interval from each member. Models receive independent clips; humans see paired events. Rates are unadjusted for label marginals. The macro row averages the fifteen models, excluding the human reference.

Model	Both correct	Same label	Neutral stab.
GPT-4o	49.6	28.7	93.1
GPT-5.6 Sol	57.3	21.4	94.0
Gemini 3.1 Pro	53.1	24.8	92.6
Claude Opus 5	54.8	23.1	93.8
Qwen3-VL-2B	31.2	46.9	89.5
Qwen3-VL-8B	42.9	37.2	91.3
Qwen3-VL-8B-Thinking	47.8	31.8	92.1
InternVL3.5-2B	29.4	49.1	88.8
InternVL3.5-4B	36.8	42.6	90.0
InternVL3.5-8B	43.7	36.0	91.5
MiniCPM-V-4.6	39.5	39.8	90.8
GLM-4.6V	45.2	33.7	91.9
LLaVA-Video-7B	40.8	40.6	90.7
VideoLLaMA 3	43.1	35.4	91.2
Video-R1	50.6	29.9	92.8
Macro mean	44.4	34.7	91.6
Human reference	86.6	4.9	96.1

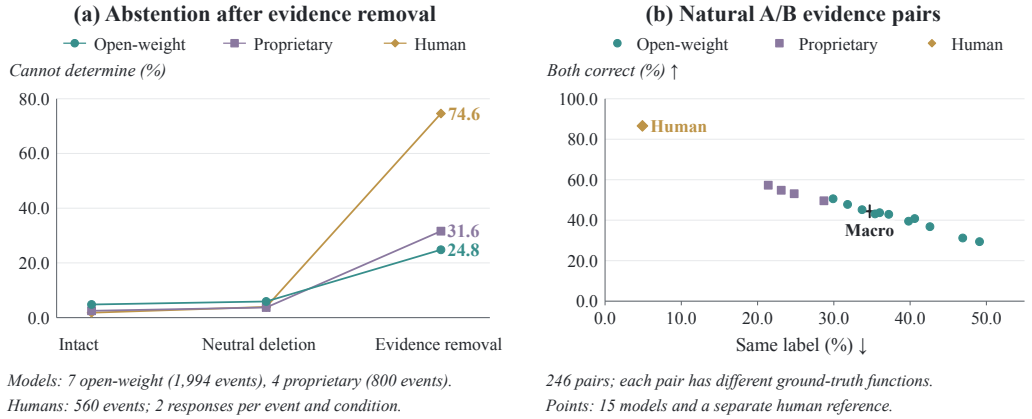

Figure 11: Human references for two behavioural controls. **(a)** Cannot-determine rates for seven open-weight models, four proprietary models, and humans; model groups use different event pools, and human rates pool individual responses. **(b)** Both-correct versus same-label rates on 246 natural pairs with different true functions. Points denote fifteen models and a separate human reference; the black cross is the unweighted model mean and excludes the human reference. Models answer clips independently; humans see paired events, so this is a descriptive comparison across protocols.

L Additional Robustness Checks

L.1 Deletion-buffer sensitivity

The primary intervention expands annotated evidence by 0.25 seconds on each side, subject to preserving the gaze interval. Smaller and larger buffers test whether a narrow boundary choice determines the result. Table 29 distinguishes the fifteen-model EGF summary from the sequence-end probe comparison across seven diagnostic configurations.

Table 29: Sensitivity to the evidence-deletion buffer. EGF is the fifteen-model macro rate, whereas the sequence-end probe change averages the seven diagnostic models. The central row is the primary setting. Buffers are clipped to preserve the gaze interval, without erasing the evaluated look.

Buffer around evidence	EGF (%)	Probe BA change (points)
± 0.10 s	73.4	-23.9
± 0.25 s	71.8	-25.8
± 0.50 s	70.9	-27.0

Figure 6b summarizes the buffer check: EGF changes from 73.4% to 71.8% to 70.9% as the buffer grows, while the probe drop increases from 23.9 to 25.8 to 27.0 points.

L.2 Taxonomy robustness

Removing exploratory-scanning from the $4{,}604$-event positive pool leaves $3{,}976$ events and changes Base accuracy from 69.0% to 69.2%. The corresponding restriction of the $4{,}138$-event removal-applicable pool changes EGF from 71.0% to 70.7%. Both starting pools already exclude no uptake. The comparison therefore tests the contribution of the exploratory category rather than simultaneously changing two categories or substituting the $1{,}994$-event paired subset.

L.3 Option positions and complementary controls

Correct answers are balanced across option positions. This limits positional shortcuts but does not remove preferences for particular semantic labels. Text-only, temporal-order, head/eye, and identity controls each change a different source of information with interpretations specific to each control.

L.4 Cross-model associations

Using the displayed model-level values in Table 1, masking retention and EGF rate have Pearson $r = -0.8723$ across fifteen models (95% CI $[-0.9569, -0.6508]$). Temporal-shuffle and evidence-removal retention have $r = 0.9886$ across the fourteen models with both measurements (95% CI $[0.9632, 0.9965]$); Claude is excluded because its temporal-shuffle condition was not run. Intervals use the Fisher z transform with standard error $1/\sqrt{n-3}$. These nominal intervals assume independent observations; shared model families limit their interpretation beyond the evaluated configurations.

M No-Uptake Events

The full benchmark includes 516 no-uptake events. Their evidence consists of a response opportunity lasting at least 2.0 seconds during which the recipient is continuously visible and behaviourally available, with no observed response. Visibility is part of the definition: an occluded recipient or a cut to another shot cannot establish an absence of uptake.

The six positive functions are supported by an observed action or response that distinguishes their role in the interaction. No uptake instead requires an observable opportunity in which a response does not occur. We therefore retain these events in the full seven-label benchmark and exclude them from the paired positive-deletion comparison and its six-way probe analysis. An event qualifies as no uptake only when the recipient remains visible and available throughout the required response opportunity, rather than merely lacking an annotated positive response.

N Further Limitations

Annotation and stimulus scope. The benchmark requires compact observable evidence and excludes events needing three or more disjoint spans. It does not cover functions that require a longer interaction history, private information, or unobserved cultural context. Labels characterize observable interactional roles, without establishing private intent.

Intervention scope. Evidence removal changes aligned ASR/subtitle text as well as visual content; raw audio is never supplied. Frames are resampled on the edited timeline, so the measured effect includes changes to sampling positions. Nonzero visual coverage does not remove this confound, and natural pairs retain uncontrolled contextual differences. Source-video grouping prevents a recording from entering both training and test folds, but cannot guarantee that recurring identities or settings are absent across distinct source videos.

Readout and inference scope. Matching BA and events does not equate a supervised probe with generated answers. A larger probe drop can partly reflect readout-specific sensitivity to the changed input distribution. Recoverability alone does not identify causal use of a feature (Belinkov, 2022). The diagnostic configurations share model families, so agreement across their contrast signs does not establish the prevalence of the pattern among independent model populations.

Behavioural scope. L2 success is conditional on the annotated target. Joint L1/L2 eligibility therefore establishes success on two separate diagnostics, not autonomous localization from a model-predicted target. Label stability can conceal falling confidence, and label switching can occur without a correct evidential reason. Abstention and neutral deletion provide additional references but do not resolve every ambiguity in an individual answer. Human comparisons use different event allocations across studies and must retain their stated denominators. Evidence-focused prompting reduces the observed failure within these tasks; generalization beyond social gaze and the sampled social-video corpora remains outside the present evaluation.