EGOPASS-BENCH: IDENTIFYING OBJECTS THAT MAY AFFECT THE WEARER'S PASSAGE IN EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding an egocentric environment requires reasoning about how surrounding entities affect the observer's actions. Existing evaluations of multimodal foundation models (MFMs) examine observer localization, motion trajectories, and action feasibility, but leave the discovery and interpretation of evolving observer–entity interactions insufficiently assessed. We introduce EGOPASS-BENCH, a benchmark for observer-relative interaction understanding in real-world egocentric video. EGOPASS-BENCH comprises 95 manually reviewed egocentric clips recorded using Ray-Ban Meta (Gen 2) smart glasses and 577 distinct questions across five tasks assessing motion perception, interaction judgments about specified entities, and relevant-entity discovery without target cues. Our evaluation of eight MFMs reveals a pronounced discrepancy between motion perception and interaction understanding. On EGOPASS-BENCH, GPT-6 Astra leads on most evaluation metrics and achieves approximately 90% accuracy in recognizing both observer and external-object motion, yet it achieves only 60.00% on relevant-entity discovery without target cues. Further analysis reveals frequent inclusion of entities that neither pose a plausible physical conflict nor constrain passage, showing that accurate motion recognition does not ensure reliable relevant-entity discovery.

1 INTRODUCTION

Visual assistants for smart glasses need to understand how the surrounding environment bears on wearers' ongoing actions. Recent work on continuous egocentric video understanding and proactive assistance has begun to move multimodal models beyond answering explicit questions toward deciding whether to intervene based on the user's context (Gong et al., 2026; Ran et al., 2026; Xu et al., 2026). Safe passage provides a concrete setting for this capability: an assistant must use continuous visual evidence to identify which external entities may require the wearer to adjust their speed or direction. This setting requires more than recognizing objects or describing scene changes; it requires understanding the consequences of those changes for the wearer's actions.

Such implications depend on the relationship between the wearer and external entities, rather than the motion state of either alone. Motion planning research analyzes potential collisions using the positions, velocities, and occupied space of both parties (Fiorini & Shiller, 1998). Research on pedestrian interactions likewise shows that anticipated time to collision can explain avoidance behavior that physical distance alone does not adequately capture (Karamouzas et al., 2014). Accordingly, an object's relevance cannot be determined solely by whether it is visible, approaching, or moving. Instead, relevance depends on the motion trends of both parties, their relative arrival times, and the available passage space. An object on a separate path may not affect the wearer, whereas a stationary structure may block passage. The challenge is to identify visually supported potential physical conflicts and passage constraints, rather than treating all salient objects as requiring caution.

This capability extends beyond wearable assistance. In robotics, systems must plan collision-avoiding velocities and paths based on their motion relative to surrounding obstacles (Fiorini & Shiller, 1998). Similarly, in autonomous driving, detecting vehicles or pedestrians is only a starting point: systems must also determine which road users pose a risk and how that risk affects driving de-

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.

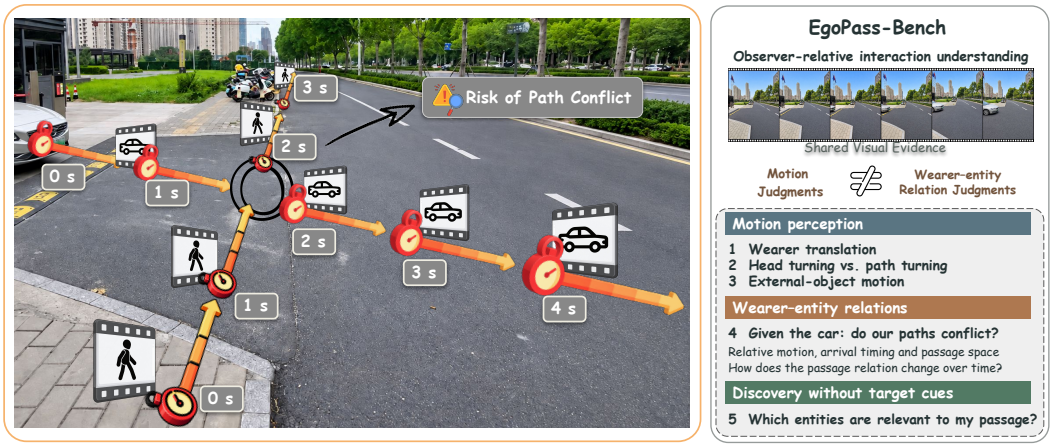


Figure 1: **From motion perception to passage understanding.** EGOPASS-BENCH evaluates how wearer motion and external entities shape passage constraints. Models must separate observed states from their implications for the wearer and identify relevant entities without target cues.

cisions (Malla et al., 2023; Kung et al., 2024). These settings share a common requirement: relating environmental states to an agent's actions.

Despite the importance of this capability across these settings, existing benchmarks assess only parts of it. MotionBench (Hong et al., 2025) evaluates fine-grained motion understanding; SAW-Bench (Li et al., 2026) assesses observer-centric spatial reasoning and action feasibility; and BOSTVG (Yao et al., 2026) grounds entities specified by referring expressions. Together, they leave untested whether a model can identify passage-relevant entities without target cues and determine whether their changing relation to the wearer calls for an adjustment in speed or direction.

To address this gap, we introduce EGOPASS-BENCH, a benchmark for evaluating observer-relative interaction understanding in real-world egocentric video. The benchmark grounds relevance in potential physical conflicts and safe-passage constraints. It comprises 577 manually annotated questions spanning multiple-choice questions (MCQs) and free-form generation (FFG). Five tasks assess wearer motion, turning, external-object motion, interaction judgments for specified objects, and relevant-entity discovery without target cues. The benchmark uses 95 manually reviewed egocentric clips, with each task answered independently so that motion recognition, relational reasoning, and entity discovery can be evaluated separately. We also evaluate models using the current clip alone or with the immediately preceding clip as context. This comparison tests whether recent visual history helps models understand the current situation.

Our evaluation of eight multimodal models reveals a gap between motion perception and relevant-entity discovery. On EGOPASS-BENCH, the highest-performing model, GPT-6 Astra, achieves accuracies of 89.58% and 90.44% on wearer translation and external-object motion recognition, respectively, but an autonomous entity-discovery score of only 60.00%, substantially below the human score of 91.14%. Differences between human and model inputs are described in Appendix B.2.

2 RELATED WORK

General video understanding. General video understanding evaluates whether models can recognize events and reason about temporal structure from continuous visual observations (Li et al., 2024; Fu et al., 2025). Table 1 summarizes representative benchmarks in general video understanding, egocentric video understanding, and safety understanding and proactive assistance. Beyond comprehensive evaluation across content domains and task types, specialized benchmarks examine event order and fine-grained motion (Liu et al., 2024; Hong et al., 2025), as well as motion, viewpoint changes, and the integration of spatial information across observations (Zhou et al., 2025a; Zhang et al., 2026a). Recent work examines consistency across related questions and the temporal evidence supporting answers (Shangguan et al., 2025; Fu et al., 2026; Huang et al., 2026). However,

Table 1: **Evaluation design of representative benchmarks.** Dimensions: [1]wearable video, [2]viewpoint–path distinction, [3]cued reasoning and uncued entity discovery, [4]changes in passage constraints, [5]insufficient evidence versus irrelevance, [6]separate entity-selection and label scoring, [7]motion–relation–discovery evaluation on shared clips, and [8]paired evaluation with and without visual history. ✓, △, and × denote explicit, partial or related, and no explicit coverage. Row colors indicate general video (blue), egocentric video (purple), and safety and proactive assistance (yellow).

Benchmark	Wearable	View–Path	Cued/Uncued	Rel. Δ	Uncertainty	Set–Label	M–R–D	Paired ctx.
MVBench	△	×	×	×	△	×	×	×
Video-MME	×	×	×	×	×	×	×	×
TempCompass	×	△	×	×	×	×	×	×
TOMATO	×	△	×	×	×	△	×	×
MotionBench	×	△	×	×	×	×	×	×
VLM4D	△	△	×	×	×	×	×	×
EgoSchema	✓	×	×	×	×	×	×	×
Minerva-Ego	✓	×	△	×	×	△	△	×
SAW-Bench	✓	✓	△	×	×	×	△	△
DRAMA	×	×	△	×	×	△	△	×
RiskBench	×	×	△	△	×	×	△	×
PaSBench-Video	×	×	△	△	×	△	△	△
EgoPro-Bench	△	×	△	×	×	△	△	△
EGOPASS-BENCH (ours)	✓	✓	✓	✓	✓	✓	✓	✓

recognizing objects and their motion does not establish that a model understands what these changes mean for the observer. We examine whether models distinguish environmental changes that affect the observer's actions from visible dynamics that do not constrain passage.

Egocentric video understanding. Egocentric video understanding studies environments from the wearer's perspective, where image changes reflect both the wearer's own motion and external activity. Egocentric datasets and benchmarks cover daily activities, episodic memory, and anticipation (Grauman et al., 2022; Mangalam et al., 2023). Recent research examines the wearer's perspective and human attention (Jung et al., 2026; Tang et al., 2026), alongside visually grounded reasoning about interactions and object-state changes (Dai et al., 2026). Another line of research concerns retaining and updating continuous experience, covering streaming episodic memory, reasoning about past and ongoing interactions, and the integration of experiences across days (Forte et al., 2026; Lei et al., 2026; Chen et al., 2026). Among the closest benchmarks, Minerva-Ego (Nagrani et al., 2026) emphasizes the visual grounding of reasoning, using spatiotemporal localization and reasoning traces to assess whether models identify and correctly use relevant evidence; SAW-Bench (Li et al., 2026) focuses on the observer's spatial situation, including localization, trajectory reasoning, spatial memory, and action feasibility. However, these evaluations provide limited coverage of discovering passage-relevant entities without target cues and judging how their passage constraints change over time. We examine how an individual entity's passage constraint strengthens or eases as the wearer and entity move. Approaching alone does not imply increasing relevance, and leaving the field of view does not establish that a constraint has eased; unresolved cases retain an insufficient-evidence label.

Safety understanding and proactive assistance. Safety understanding and proactive assistance study how visual observations inform judgments about risks and user needs. Risk localization and explanation ground relevant entities and explain their implications for action (Malla et al., 2023; Zhou et al., 2025b; Gao et al., 2026). Other benchmarks assess hazard anticipation, timely warnings, and avoidance of unnecessary alarms (Kung et al., 2024; Zhao et al., 2026). As continuous video perception develops, research increasingly considers evidence accumulation and response timing (Liu et al., 2026; Gong et al., 2026), while incorporating individual needs, task progress, and ongoing interactions to determine whether, when, and how to provide assistance (Ran et al., 2026; Xu et al., 2026; Zhang et al., 2026b). Rather than primarily evaluating warnings or assistance, we assess the environmental understanding underlying these behaviors. By separately examining state recognition, interaction reasoning, and relevant-entity discovery, our benchmark provides a more specific basis for analyzing models' omissions of relevant entities and inclusion of irrelevant ones.

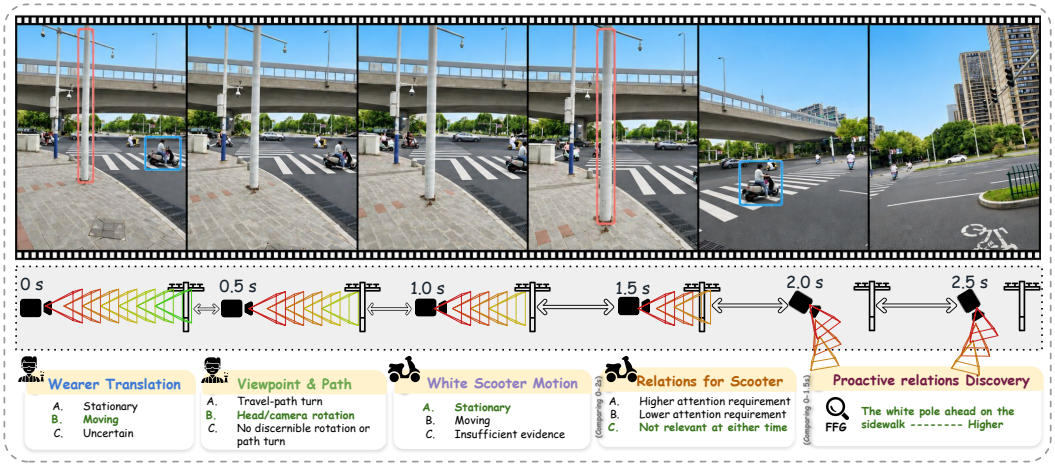


Figure 2: **Five tasks share the same visual evidence.** Timestamped frames support judgments about wearer translation, turning, external object motion, passage relations for a specified entity, and autonomous discovery of all relevant entities.

3 Task Definitions and Benchmark Construction

3.1 Task Definition

EgoPass-Bench evaluates passage understanding in dynamic egocentric environments. Each input consists of a timestamped frame sequence $V = \{(I_j, t_j)\}_{j=1}^{N}$ and a question q, where I_j is the j-th frame and t_j is its timestamp. The question specifies an interval to examine or two time points to compare. Additional visual history is denoted by H: $H = \varnothing$ for No History; History provides the frame sequence from the immediately preceding clip.

Four MCQ tasks assess wearer translation, viewpoint and path turning, external-object motion, and passage relations for specified objects. For each question q, the model selects an answer from the corresponding candidate-answer set $\mathcal{A}_q$:

$$\hat{a} = f(V, H, q), \qquad \hat{a} \in \mathcal{A}_q, \tag{1}$$

where f denotes the evaluated model and $\hat{a}$ is its selected answer.

Autonomous relevant-entity discovery uses FFG and provides no candidate-object cues. An entity should be included if it is passage-relevant at either of the two comparison time points. The passage-relevance criteria are defined in Section 3.2.

The model must identify these entities and return a description and a change label for each:

$$\hat{\mathcal{S}} = f(V, H, q) = \{(\hat{d}_i, \hat{\ell}_i)\}_{i=1}^{m}, \tag{2}$$

where m is the number of reported entities, $\hat{d}_i$ is a description that enables the entity to be located in the scene, and $\hat{\ell}_i$ indicates the change in its required attention from t_1 to t_2. The label is `HIGHER`, `LOWER`, or `INSUFFICIENT_EVIDENCE`, indicating increased attention, decreased attention, or insufficient evidence for a reliable comparison, respectively. When no entity satisfies the passage-relevance criteria, the correct output is the empty set.

All five tasks are answered independently to assess motion perception, passage-relation reasoning, and relevant-entity discovery separately.

3.2 Defining Passage Relevance

Reported driver perception–brake response times to unexpected traffic events are approximately 1.25–1.5 seconds (Green, 2000). Drawing on this timescale, we adopt a two-second window to identify passage-relevant objects that may require prompt adjustments to the wearer's movement.

At either of the two queried time points, an object is considered passage-relevant if the visual evidence and current motion of both the wearer and the object indicate a plausible collision risk or insufficient passage space within the next two seconds. An object also qualifies if avoiding a later conflict requires the wearer to begin adjusting their movement within this window. Required attention denotes monitoring, preparation, or movement adjustment warranted by these passage constraints, rather than visual salience. Full annotation criteria are provided in Appendix D.

3.3 Diagnostic Tasks

Situation awareness research distinguishes the perception of environmental elements, comprehension of their meaning, and projection of their subsequent states (Endsley, 1995). Motion planning and pedestrian interaction research further indicate that passage judgments depend on relative motion and potential conflicts between an agent and surrounding entities, rather than the state of a single entity (Fiorini & Shiller, 1998; Karamouzas et al., 2014). These perspectives motivate five tasks assessing motion perception, passage-relation understanding, and relevant-entity discovery:

- **Wearer Translation.** Determine whether the wearer changes position during the specified interval, distinguishing bodily movement from image changes caused by head or camera rotation.
- **Viewpoint and Path Turning.** Identify the source, left–right direction, and temporal order of turning events, distinguishing head or camera rotation from changes in the actual travel path.
- **External-Object Motion.** Determine the motion state of a specified entity in the physical world, distinguishing stationary and moving states from confirmed starts or stops. The task requires avoiding the misinterpretation of image changes caused by wearer motion as object motion.
- **Passage Relations for Specified Objects.** Given a target object and two comparison time points, determine whether the object meets the passage-relevance criteria and whether the attention it requires increases or decreases. Objects that are irrelevant at both time points and cases with insufficient evidence are handled separately.
- **Autonomous Relevant-Entity Discovery.** Without candidate-object cues, identify all entities that meet the inclusion criteria at either comparison time point. For each entity, provide a description that enables it to be located in the scene and indicate the change in required attention. Return an empty set when no entity qualifies.

3.4 Video Collection

We use second-generation Ray-Ban Meta smart glasses to record everyday passage scenes from the wearer's perspective, covering traffic intersections, sidewalks, residential areas, commercial districts, pedestrian streets, and parks. For each scene category, we record seven videos of varying lengths. Recordings preserve natural movement, pauses, and leftward and rightward head turns. We manually select suitable segments from these recordings to create the benchmark clips, which provide shared visual evidence for the five tasks.

3.5 Clip-Level Evaluation and Visual Evidence

Short clips localize passage judgments in time, avoiding a single judgment over changing interaction states while preserving sequential evidence of wearer motion, external-object motion, and viewpoint changes. EgoPass-Bench comprises 95 manually reviewed egocentric clips and 577 questions across five tasks. Frames are sampled at a nominal 2 fps and presented chronologically with absolute timestamps from the source video. Each question specifies an interval within the current clip.

3.6 Question Construction and Quality Control

Question construction. Reference answers are provided by the video recorders based on the circumstances during filming and the recorded videos. Annotations cover observer and object localization, motion trajectories, and judgments of observer–object relations to minimize ambiguity in the reference answers.

Quality control. Each recording is manually reviewed for video quality. Recordings with overly uniform scenes, poor visibility of key objects, or other issues that may hinder spatial reasoning are

discarded and recorded again under the same filming guidelines. The two participants independently completed the full 577-question No History set. Their responses yield task-wise Cohen's κ of 0.60–0.75 for MCQs, 0.77 for FFG object selection, and 0.76 for conditional change labels. Agreement is above chance across these components, with lower consistency for turning and passage relations; Appendix B.1 reports raw agreement, exact counts, and comparison units.

3.7 Context Conditions

To examine the role of prior observations, we evaluate a subset of 54 clips under two input conditions: **No History** and **History**. No History provides only the current clip. History additionally provides images from the immediately preceding clip of the same video, without earlier footage, previous questions, model answers, or textual summaries. The first clip has identical inputs under both conditions. In both settings, models answer questions about the current interval. The added visual history thus serves as evidence for inferring the current state, rather than as a basis for carrying forward conclusions about earlier moments.

Both conditions use the same relevance criteria and scoring rules. The video recorders provide references using each condition's permitted visual evidence; additional history may change what can be inferred and hence the reference answer. Within each condition, all models use the same references.

4 Evaluation Setup and Results

4.1 Models and Metrics

Evaluation setting. We evaluate eight models from four families: GPT (Singh et al., 2026), Gemini (Comanici et al., 2025), Llama (Grattafiori et al., 2024), and Qwen (Qwen Team, 2026). The evaluated models are GPT-6 Astra, GPT-5.6 Sol, Gemini 3.8 Flash, Gemini 3.1 Pro, Llama 4 Scout, Llama 4 Maverick, Qwen3.8 Max, and Qwen3.8 Flash. Appendix A.3 reports their inference configurations. Humans and models are evaluated on the same 577-question bank across all five task types without history. Table 2 uses every scorable response for each model. Appendix C provides the task counts and model-specific exclusions.

Evaluation metric. MCQ accuracy is the fraction of scorable questions answered correctly. Overall accuracy pools the questions across the four tasks rather than averaging task accuracies. FFG evaluates entity discovery and change labels separately and jointly. We match predicted entities to reference entities one to one. True positives (TP) count correct matches, false positives (FP) count additional reports, and false negatives (FN) count missed reference entities. Each reference entity can be matched once, so duplicate reports do not increase TP. For reference set G, the object score for one question is

$$s(G, P) = \begin{cases} \max\{0, (\mathrm{TP} - \mathrm{FP})/|G|\}, & |G| > 0, \\ \mathbf{1}[P = \varnothing], & |G| = 0, \end{cases} \tag{3}$$

where P is the predicted set. When the reference set is empty, an empty prediction receives a score of one and any nonempty prediction receives zero.

Discovery averages object scores across scorable FFG questions. Labels measures label accuracy among successfully matched entities with an adjudicated label correspondence, excluding omissions and false positives. FFG Overall requires the entity set and all change labels to match the reference exactly. Human scores average the two participants and serve as supplementary references, not a controlled performance ceiling; Appendix B details the comparison units. Only models are ranked.

4.2 Performance without History

Table 2 reports model performance across the five task types. GPT-6 Astra leads in overall MCQ accuracy and all three FFG metrics, whereas Gemini 3.8 Flash leads in turning. A model's overall rank therefore does not describe its strengths across all tasks. Performance is strongest on external-object motion and weaker on passage relations. Even when the question specifies the target, models struggle to determine whether and how it constrains the wearer's passage.

Table 2: Performance without history (%). Bold and underlining mark the best and second-best model results per column. *Human scores average the two participants. †Provider content filtering excludes four responses for Qwen3.8 Max and two for Qwen3.8 Flash from their scores and denominators; see Appendix C. Metrics are defined in Section 4.1.

	MCQ					FFG		
Model	Overall	Wearer translation	Turning	Object motion	Passage relation	Overall	Discovery	Labels
Human reference*	89.83	93.75	86.17	92.28	87.23	86.36	91.14	90.56
GPT-6 Astra	**76.87**	**89.58**	<u>67.02</u>	**90.44**	**61.70**	**54.55**	**60.00**	**83.72**
Gemini 3.8 Flash	<u>71.95</u>	71.88	**70.21**	<u>87.50</u>	<u>58.16</u>	<u>46.36</u>	<u>53.18</u>	66.67
Gemini 3.1 Pro	62.74	66.67	47.87	84.56	48.94	37.27	46.82	69.77
GPT-5.6 Sol	61.67	63.54	59.57	86.76	37.59	29.09	38.18	<u>72.34</u>
Qwen3.8 Max†	62.50	<u>76.04</u>	61.29	83.82	33.09	22.94	30.73	67.50
Qwen3.8 Flash†	59.01	75.00	48.39	82.35	32.62	11.93	18.35	58.97
Llama 4 Scout	48.18	37.50	58.51	66.18	31.21	30.91	42.73	39.13
Llama 4 Maverick	49.89	37.50	53.19	77.94	29.08	17.27	31.36	40.74

4.3 Differences within Model Families

Within-family rankings can change by task. Gemini 3.8 Flash leads Gemini 3.1 Pro on MCQs and discovery, whereas Pro leads on conditional labels. The Llama pair likewise shows task-dependent strengths.

5 Analysis and Discussion

Strong motion recognition does not consistently translate into accurate entity selection and interaction judgments. We examine discovery, change labels, and insufficient evidence to understand how locally correct judgments can still produce incomplete or imprecise descriptions of the environment.

5.1 Finding More Entities Does Not Ensure Better Understanding

Discovery errors are not limited to omissions. Figure 3 (right) averages the two evaluated models in each family and separates correct detections from additional reports. Qwen and Gemini recover the same mean number of reference entities, but Qwen produces nearly twice as many extra reports. GPT recovers the most reference entities on average, while Llama recovers the fewest. Across all eight models, micro-precision ranges from 23.78% to 44.30% on the shared diagnostic subset (Appendix C.4). These results distinguish coverage from precise selection: finding more relevant entities does not imply that a model reliably excludes irrelevant ones.

The empty and nonempty cases further expose this distinction. Qwen has the lowest mean score in both settings, with particularly poor recognition of empty cases, where every reported entity is irrelevant. GPT and Gemini have similar mean scores despite their different detection counts. Recognizing that no entity meets the criteria is therefore part of understanding the situation, not simply a failure to find objects. GPT-6 Astra achieves an FFG discovery score of 60.00%, only slightly above the 59.09% baseline obtained by always returning an empty set. This comparison highlights the influence of empty-case prevalence on the aggregate score: reliable entity discovery must also be assessed on nonempty cases. We therefore report empty and nonempty performance separately. Individual model results remain available in Appendix C.4.

5.2 Approaching Does Not Necessarily Increase Passage Conflict

Approach describes relative motion, whereas conflict also depends on movement direction, the timing of entry into shared space, and available width. Being closer does not imply that the paths are harder to separate. Pedestrian interaction research likewise finds that anticipated time to collision explains observed interactions better than physical distance alone (Karamouzas et al., 2014). This supports interpreting an object's implications through its dynamic relation to the observer.



Figure 3: **(Left) Passage-relation error. (Right) Coverage is not precise selection.**

Physical distance and passage constraints can change in opposite directions. In Figure 3 (left), the pedestrian approaches the observer while yielding toward the left of the image. The reference judgment is therefore decreased attention, yet most models select the opposite change. The target is explicitly given in this question, so failure to discover it cannot alone explain the disagreement. A target cue does not ensure that a model correctly updates the passage relation. The error motivates testing whether models overuse approach cues.

5.3 Correct Local Judgments Can Still Yield Incomplete Answers

A model may judge discovered entities accurately without describing the relevant environment completely. Figure 7 in Appendix C.4 separates reference entities into correct discovery with a correct label, an incorrect label, and omission. GPT-6 Astra correctly labels most matched entities but still misses relevant ones. GPT-5.6 Sol detects more entities, yet the number that it both finds and labels correctly is close to GPT-6 Astra's, alongside more additional reports. Greater coverage does not necessarily add more correct environmental information.

Situation awareness theory distinguishes perceiving environmental elements, understanding their meaning, and anticipating their subsequent states (Endsley, 1995). Our results echo this distinction: correct judgments about some entities can coexist with incorrect entity selection. Better understanding requires both accurate judgments and an appropriate, sufficiently complete entity set.

5.4 Recognizing Insufficient Evidence Remains Difficult

Understanding the environment also requires recognizing what the observations cannot establish. Table 13 in Appendix C.3 shows that models rarely identify insufficient evidence and instead choose definite states or relations. This differs from confusing one motion state with another: the model fails to recognize that the evidence needed for a reliable judgment is missing. Occlusion, viewpoint changes, and an entity leaving view can limit what egocentric observations establish. Reliable understanding must distinguish supported changes from unresolved states. Together, these results expose gaps between entity selection, change interpretation, and evidence assessment. Improving one component alone does not ensure a complete and appropriately qualified answer.

5.5 Discovery Remains Challenging with Visual History

History adds the immediately preceding clip while models continue to answer about the current clip. Figure 4 compares both conditions on the same 257 questions, including 39 FFG questions. MCQ Overall improves for seven of eight models, but Discovery improves for only five. Gemini 3.8 Flash gains 12.82 percentage points in Discovery, whereas Llama 4 Scout loses 16.67 despite gaining 4.13 points in MCQ Overall. Gains also vary within a model: GPT-6 Astra improves in turning and

	MCQ					FFG		
	Overall	Wearer translation	Turning	Object motion	Passage relation	Overall	Discovery	Labels
GPT-6 Astra	+4.13	−3.77	+8.89	+8.06	+3.45	+5.13	+5.13	+2.20
Gemini 3.8 Flash	+3.21	+7.55	+8.89	−6.45	+5.17	+7.69	+12.82	−2.38
Gemini 3.1 Pro	−1.38	−5.66	−8.89	+1.61	+5.17	+5.13	+7.69	+5.77
GPT-5.6 Sol	+3.21	+7.55	+4.44	+3.23	−1.72	+10.26	+10.26	0.00
Qwen3.8 Max	+3.21	+5.66	+2.22	+3.23	+1.72	0.00	−2.56	+9.09
Qwen3.8 Flash	+1.83	+3.77	+2.22	+4.84	−3.45	0.00	+2.56	+14.55
Llama 4 Scout	+4.13	+5.66	+6.67	+3.23	+1.72	−12.82	−16.67	+12.50
Llama 4 Maverick	+5.50	+20.75	−13.33	+8.06	+3.45	+2.56	−1.28	+12.50

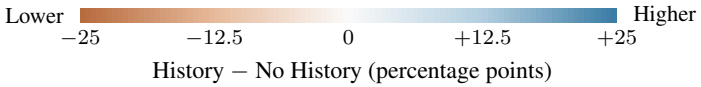


Figure 4: **Changes with visual history on the same questions.** Cells show History minus No History in percentage points on 257 shared questions. Positive values (blue) indicate improvement and negative values (orange) decline; the Overall columns are outlined. Metrics follow Section 4.1; paired absolute scores and reference-answer sources are in Appendix C.1.

object motion while declining in wearer translation. Thus, broader visual context can benefit some judgments without consistently improving which entities are identified as relevant to passage.

This uneven response is consistent with prior evidence that additional frames yield limited gains in fine-grained temporal understanding (Cai et al., 2024), and that temporal perception varies across temporal aspects and task formats (Liu et al., 2024). History benefits motion judgments without consistently improving relevant-entity discovery. The comparison uses condition-specific references; paired absolute scores, full coverage results, and reference-answer sources are provided in Appendix C.1.

6 Conclusion

Understanding dynamic egocentric environments requires identifying which surrounding objects may affect the observer's passage and how this influence changes as both the observer and the objects move. Existing evaluations cover motion recognition, spatial localization, and action feasibility, but provide limited systematic assessment of these dynamic passage relations, particularly the ability to discover relevant entities without target cues. To address this gap, we introduce EgoPass-Bench, comprising 95 manually reviewed real-world egocentric video clips and 577 questions. Its five tasks jointly evaluate motion perception, passage-relation judgments for specified objects, and relevant-entity discovery without target cues.

Our evaluation of eight multimodal foundation models shows that strong motion recognition does not ensure reliable understanding of passage relations. GPT-6 Astra achieves accuracies close to 90% in both wearer translation and external-object motion, yet its relevant-entity discovery score is only 60.00%, below the human reference score of 91.14%. Models miss relevant entities and report many irrelevant ones. Even with correct object identification, they may misjudge passage-constraint changes or give definite answers despite insufficient evidence. Preceding visual context improves some tasks but does not consistently improve relevant-entity discovery. These results highlight the need to integrate object recognition, relevance selection, and dynamic relation reasoning. We hope EgoPass-Bench helps diagnose these gaps and advance understanding of how environmental changes affect the observer's actions.

AI USE STATEMENT

We used GPT to assist with language polishing. We also used generative AI to create some Scalable Vector Graphics (SVG) icons in the main paper, including icons representing the GPT, Qwen, Gemini, and Llama model families. The authors take full responsibility for the final content of this work, including text and graphical materials produced with AI assistance.

REFERENCES

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models, 2024. URL https://arxiv.org/abs/2410.10818.

Weitao Chen, Hu Jiaxin, Xie Tianyidan, Yang Li, Yuyi Qian, Banghao Xu, Ziheng Tang, Shenyi Wang, Mingyue Yu, Duo Li, Jiacheng Shi, Gao Wang, Zhan Xu, Zhicheng Qiu, Xuanfu Li, Jian Yang, Lanjun Wang, and Zili Yi. Egomonth: A month-level egocentric video benchmark for long-term spatiotemporal memory, 2026. URL https://arxiv.org/abs/2608.13113.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Yang Dai, Dian Jiao, Tianwei Lin, and Wenqiao Zhang. Egocot-bench: Benchmarking grounded and verifiable operation-centric chain of thought reasoning for mllms, 2026. URL https://arxiv.org/abs/2605.19559.

Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, March 1995. ISSN 1547-8181. doi: 10.1518/001872095779049543. URL http://dx.doi.org/10.1518/001872095779049543.

Paolo Fiorini and Zvi Shiller. Motion planning in dynamic environments using velocity obstacles. *The International Journal of Robotics Research*, 17(7):760–772, July 1998. ISSN 1741-3176. doi: 10.1177/027836499801700706. URL http://dx.doi.org/10.1177/027836499801700706.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. Egostream: A diagnostic benchmark for streaming episodic memory in egocentric vision, 2026. URL https://arxiv.org/abs/2605.31557.

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

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Xiaowei Gao, Pengxiang Li, Yitai Cheng, Ruihan Xu, James Haworth, Stephen Law, and Yun Ye. Unidrive: A unified vision-language and grounding framework for interpretable risk understanding in autonomous driving, 2026. URL https://arxiv.org/abs/2606.24759.

Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. In *European Conference on Computer Vision (ECCV)*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

Marc Green. "how long does it take to stop?" methodological analysis of driver perception-brake times. *Transportation Human Factors*, 2(3):195–216, Sept 2000. ISSN 1093-9741. doi: 10.1207/sthf0203_1. URL http://dx.doi.org/10.1207/sthf0203_1.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, June 2025.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. Eg-vqa: Benchmarking verifiable video question answering with grounded temporal evidence, 2026. URL https://arxiv.org/abs/2606.24797.

Jihyeok Jung, Jeewu Lee, Sanghyeop Kim, Chanhee Han, and Seong Joon Oh. Egogapbench: Benchmarking egocentric action selection in multi-agent scenes, 2026. URL https://arxiv.org/abs/2607.00547.

Ioannis Karamouzas, Brian Skinner, and Stephen J. Guy. Universal power law governing pedestrian interactions. *Phys. Rev. Lett.*, 113:238701, Dec 2014. doi: 10.1103/PhysRevLett.113.238701. URL https://link.aps.org/doi/10.1103/PhysRevLett.113.238701.

Chi-Hsi Kung, Chieh-Chi Yang, Pang-Yuan Pao, Shu-Wei Lu, Pin-Lun Chen, Hsin-Cheng Lu, and Yi-Ting Chen. Riskbench: A scenario-based benchmark for risk identification. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14800–14807, 2024. doi: 10.1109/ICRA57147.2024.10610270.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. Egosat: A comprehensive benchmark of egocentric streaming interaction understanding, 2026. URL https://arxiv.org/abs/2606.24422.

Chuhan Li, Rilyn R. Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. SAW-bench: Learning situated awareness in the real world. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8lwrYjv6r7.

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

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video, 2026. URL https://arxiv.org/abs/2603.12938.

Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1043–1052, January 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

Arsha Nagrani, Jasper Uijlings, Shyamal Buch, Tobias Weyand, Sudheendra Vijayanarasimhan, Bo Hu, Ramin Mehran, David A Ross, and Cordelia Schmid. Minerva-ego: Spatiotemporal hints for egocentric video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38859–38869, June 2026.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Dongchuan Ran, Linyu Ou, Xueheng Li, Wenwen Tong, Chenxu Guo, Hewei Guo, Kaibing Wang, and Lewei Lu. Egopro-bench: Benchmarking personalized proactive interaction in egocentric video streams, 2026. URL https://arxiv.org/abs/2605.07299.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/16ba99f25a235f1100a4014d71d34ad8-Paper-Conference.pdf.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Qiance Tang, Ziqi Wang, Jieyu Lin, Ziyun Li, Barbara De Salvo, and Sai Qian Zhang. Egoeverything: A benchmark for human behavior inspired long context egocentric video understanding in ar environment, 2026. URL https://arxiv.org/abs/2604.08342.

Lilin Xu, Bufang Yang, Siyang Jiang, Kaiwei Liu, Kaiyuan Hou, Yuang Fan, Hongkai Chen, Zhenyu Yan, and Xiaofan Jiang. Pro^{2}assist: Continuous step-aware proactive assistance with multi-modal egocentric perception for long-horizon procedural tasks, 2026. URL https://arxiv.org/abs/2605.04227.

Jiali Yao, Xin Gu, Xinran Deng, Mengrui Dai, Bing Fan, Zhipeng Zhang, Yan Huang, Heng Fan, and Libo Zhang. Omnistvg: Toward spatio-temporal omni-object video grounding. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 14584–14605, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/18374cc223b6c8141f5d5fd08436499c-Paper-Conference.pdf.

Qifeng Zhang, Kaixiang Huang, Heng Dong, Huang Fang, Junting Chen, Junjie Zhu, Yonghang Chen, Zhiyu Zhang, and Wei Li. Gst-bench: Can vlms develop global spatial awareness from video?, 2026a. URL https://arxiv.org/abs/2608.05747.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding, 2026b. URL https://arxiv.org/abs/2608.05703.

Yusong Zhao, Yuejin Xie, Youliang Yuan, Junjie Hu, Jitian Guo, Yujiu Yang, and Pinjia He. Pasbench-video: A streaming video benchmark for proactive safety warning, 2026. URL https://arxiv.org/abs/2606.02443.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8600–8612, October 2025a.

Yuchen Zhou, Jiayu Tang, Xiaoyan Xiao, Yueyao Lin, Linkai Liu, Zipeng Guo, Hao Fei, Xiaobo Xia, and Chao Gou. Where, what, why: Towards explainable driver attention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2675–2685, October 2025b.

A BENCHMARK DETAILS

A.1 DATA COMPOSITION AND TEMPORAL SAMPLING

Each reviewed clip serves as the unit of visual observation and question construction. The resulting question set spans five tasks, covering wearer motion, external-object motion, passage-relation judgments, and autonomous discovery of relevant entities. Table 3 summarizes the question distribution.

The benchmark contains 577 distinct questions. The current History comparison uses a 54-clip subset containing 293 questions; evaluating this subset under both conditions yields 586 question–condition evaluation instances. These condition-specific instances are not additional distinct questions.

Table 3: Question distribution across the five tasks in EGOPASS-BENCH.

Task	**Questions**
Wearer translation	96
Turning	94
External-object motion	136
Specified-object passage relation	141
Autonomous discovery	110
Total	**577**

The source recordings cover six types of everyday passage scenes: traffic intersections, sidewalks, residential areas, commercial districts, pedestrian streets, and parks. For each type, we record seven videos of varying lengths and manually select suitable segments to construct the benchmark clips. Clip duration defines the visible observation interval, whereas the two-second relevance horizon concerns potential passage constraints at a queried time point.

A.2 ADDITIONAL RELEVANCE AND ATTENTION DEFINITIONS

Passage relevance is determined by the spatiotemporal relationship between the wearer and an external entity. Approaching along separate paths with sufficient clearance, or crossing the same path at clearly different times, does not by itself establish a conflict. Conversely, a stationary structure may be relevant if it restricts the wearer's forward passage. Entities are defined according to the space they physically occupy: a rider and the vehicle being ridden are treated as a single unit, whereas independent objects are assessed separately.

To assess whether models understand how passage constraints evolve over time, we compare the attention required by the same object at two time points within a clip. Required attention refers to the need for observation, preparation, or action adjustment arising from the passage relation, rather than visual salience or subjective anxiety. An object is classified as irrelevant only when it clearly fails to meet the inclusion criteria at both time points. Cases that cannot be judged reliably retain an insufficient-evidence label. An object leaving the field of view is not, by itself, treated as evidence that the required attention has decreased.

A.3 MODEL CONFIGURATIONS

Table 4 summarizes the model identifiers, recorded settings, and documented defaults for our main No History evaluation. Each question is answered independently using the permitted visual evidence and a task-specific output schema. Reference answers and other models' responses are not included in the input.

In Table 4, $--$ denotes an unset or inapplicable parameter. *GPT-6 does not support temperature or top-p adjustment. GPT sampling values and output caps were not recorded. ^{a}Nominal 2 fps; clearer replacement frames retain their actual timestamps. ^{b}Qwen3.8 Flash uses enabled reasoning without a specified effort level. ^{c}Defaults from saved Google model metadata. ^{d}Provider defaults documented on September 22, 2026; effective runtime values were not returned. Top-K=0 disables top-k sampling.

Table 4: **Model configurations for evaluation.** Settings are based on evaluation records and official model or provider documentation. Unset parameters use defaults where applicable. **Temp.**: temperature.

Model	Model identifier	Max output tokens	Temp.	Top-P	Top-K	Frame sampling	Thinking setting
GPT-6 Astra*	`gpt-6-astra`	–	–	–	–	2 fps[a]	Medium
GPT-5.6 Sol	`gpt-5.6-sol`	–	–	–	–	2 fps[a]	Medium
Gemini 3.8 Flash	`gemini-3.8-flash`	16,384	1.0	0.95[c]	64[c]	2 fps[a]	Medium
Gemini 3.1 Pro (Preview)	`gemini-3.1-pro-preview`	16,384	1.0	0.95[c]	64[c]	2 fps[a]	Medium
Llama 4 Maverick	`meta-llama/llama-4-maverick`	4,096	1.0[d]	1.0[d]	0[d]	2 fps[a]	–
Llama 4 Scout	`meta-llama/llama-4-scout`	4,096	1.0[d]	1.0[d]	0[d]	2 fps[a]	–
Qwen3.8 Max (0902)	`qwen/qwen3.8-max-0902`	16,384	0.6[d]	0.95[d]	20[d]	2 fps[a]	Medium
Qwen3.8 Flash	`qwen/qwen3.8-flash`	16,384	0.6[d]	0.95[d]	20[d]	2 fps[a]	On[b]

GPT models use medium reasoning effort. Gemini models use high media resolution and do not return reasoning traces. The temperature, top-p, and top-k parameters for Llama and Qwen use their respective provider defaults.

B Human Reference Scores and Agreement

Both participants completed the full 577-question No History set, comprising 467 multiple-choice questions (MCQs) and 110 free-form generation (FFG) questions. This is the same question bank and the same five task types used for model evaluation; model-specific scoring exclusions are detailed in Appendix C. The results use the two participants' independently completed responses. Percentages and Cohen's κ are reported to two decimal places; the stated numerators and denominators define the exact scores. Human means are arithmetic means of the two participants' individual scores.

B.1 Agreement Between Participants

We compute unweighted Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and $p_e = \sum_c p_{1c} p_{2c}$ is expected agreement from the participants' marginal category proportions. Agreement compares responses with each other, whereas accuracy compares each response with the reference answer. We do not pool different task category spaces into a single κ.

MCQ categories and reference decisions. We compare semantic categories rather than shuffled option letters. Turning labels are normalized by source, direction, and event order, without treating different absolute timestamps as different categories across questions. The task denominators are 96 for wearer translation, 94 for turning, 136 for object motion, and 141 for passage relations. In one turning question, the reference accepts two distinct answer options. The participants select different accepted options, so both responses are correct but disagree. In another turning question, both participants report no turn, whereas the reference specifies no discernible head/camera turn from 9.0 to 12.5 seconds followed by a rightward turn from 12.5 to 13.5 seconds. Both responses are therefore incorrect but agree with each other. Both examples remain included in the accuracy and agreement calculations.

FFG comparison units. Selection agreement uses the fixed set of 325 question–candidate-object instances in the review interface, coded as selected or not selected. There are 43 instances selected by both participants, 261 selected by neither, 8 selected only by User 1, and 13 selected only by

Table 5: Agreement between the two participants on the full 577-question No History set. Each comparison unit receives equal weight.

Task / component	Unit	Same/total	Agree. (%)	κ
Wearer translation (W1)	Question	86/96	89.58	0.74
Turning (W2)	Question	71/94	75.53	0.61
Object motion (O1)	Question	115/136	84.56	0.75
Passage relation (T1)	Question	105/141	74.47	0.60
FFG selection	Candidate object	304/325	93.54	0.77
FFG labels	Joint reference match	37/43	86.05	0.76
FFG joint	Candidate object	298/325	91.69	0.71

Agreement compares responses, not correctness. Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is Same/total and p_e is chance agreement from the two participants' category frequencies. No pooled κ is reported across tasks or comparison units.

User 2. These units receive equal weight, so the statistic depends on the candidate universe and questions with more candidates contribute more units. Conditional label agreement uses only the 43 assessable reference objects successfully matched by both participants: 37 labels agree and 6 differ. The nominal labels are higher attention, lower attention, and insufficient evidence; missed objects and false positives are excluded. The joint measure uses all 325 units and adds not selected as a fourth state. Not selected denotes absence from a response, rather than an adjudicated irrelevance judgment.

Question-level agreement. MCQ semantic responses agree on $377/467 = 80.73\%$ of questions. FFG object sets agree on $89/110 = 80.91\%$, and object sets with every label agree on $84/110 = 76.36\%$. Strict agreement over the full set is $(377 + 84)/577 = 461/577 = 79.90\%$. Both participants return an empty set on 54 FFG questions; these count as question-level agreements but create no conditional label units.

B.2 Accuracy Against Reference Answers

Table 6: Individual human reference scores on all 577 questions without history (%). Each mean averages the two participants' scores. MCQ and FFG metrics follow Section 4.1; exact counts are reported in this appendix.

	MCQ					FFG		
Participant	Overall	Wearer translation	Turning	Object motion	Passage relation	Overall	Discovery	Labels
User 1	89.08	95.83	87.23	92.65	82.27	85.45	90.91	87.23
User 2	90.58	91.67	85.11	91.91	92.20	87.27	91.36	93.88
Human mean	89.83	93.75	86.17	92.28	87.23	86.36	91.14	90.56

Overall MCQ accuracy uses the total correct answers across all 467 questions, not an unweighted average of task accuracies. User 1 answers 416 questions correctly and 51 incorrectly; User 2 answers 423 correctly and 44 incorrectly. Table 7 gives the exact task counts underlying Table 6.

Table 7: Exact human MCQ counts on the full No History question set. U1 and U2 denote User 1 and User 2, respectively. W1: wearer translation; W2: turning; O1: object motion; T1: passage relations.

Task	Questions	U1 correct	U1 wrong	U2 correct	U2 wrong
W1	96	92	4	88	8
W2	94	82	12	80	14
O1	136	126	10	125	11
T1	141	116	25	130	11
All MCQ	467	416	51	423	44

The 110 FFG questions contain 53 reference-object instances: 45 questions have nonempty reference sets and 65 have empty sets. Objects are matched one to one; a rider and vehicle form one occupied unit. For a nonempty reference set G, the per-question object score is $\max\{0, (\mathrm{TP}-\mathrm{FP})/|G|\}$. For an empty reference set, the score is one only when the response is also empty. The mean is taken over all 110 questions. Conditional label accuracy includes only successfully matched reference objects with assessable labels and excludes omissions and false positives.

Table 8: Human FFG counts and scores on the full No History set. The two participants are scored separately; human references average their individual ratios.

Metric	User 1	User 2
Scorable questions	110	110
Empty / nonempty reference sets	65/45	65/45
Reference-object instances	53	53
Reported objects	51	56
True positives (TP)	47	49
False positives (FP)	4	7
False negatives (FN)	6	4
Sum of per-question Discovery scores	100.00	100.50
Discovery numerator / denominator	100.00/110	100.50/110
Discovery score (%)	90.91	91.36
Correct / assessable matched labels	41/47	46/49
Incorrect matched labels	6	3
Conditional label accuracy (%)	87.23	93.88
Exact set-and-label answers	94/110	96/110
FFG Overall (%)	85.45	87.27

The mean object score is $\frac{1}{2}(100/110 + 100.5/110) = 91.14\%$. The mean conditional label accuracy is $\frac{1}{2}(41/47 + 46/49) = 90.56\%$, rather than the pooled fraction $87/96$. The label-accuracy denominators (47 and 49) differ from the agreement denominator (43), which requires a successful match by both participants.

Human evaluation procedure. Each annotator answers all questions independently, with unlimited time. For FFG, the interface (Figure 5) provides a list of scene objects, including distractors, to simplify response recording and aggregation. Annotators select relevant objects and assign change labels, or return an empty set. Because the list may omit relevant objects, annotators can add unlisted objects through the custom-object field or notes. These additions are submitted for individual review to the annotators who provided the initial reference answers.

Human–model comparison. Humans and models answer the same question bank, with model-specific exclusions reported separately. Humans receive candidate-object descriptions, whereas models generate object descriptions and change labels without a candidate list. Human scores therefore serve as a reference under candidate-assisted conditions.

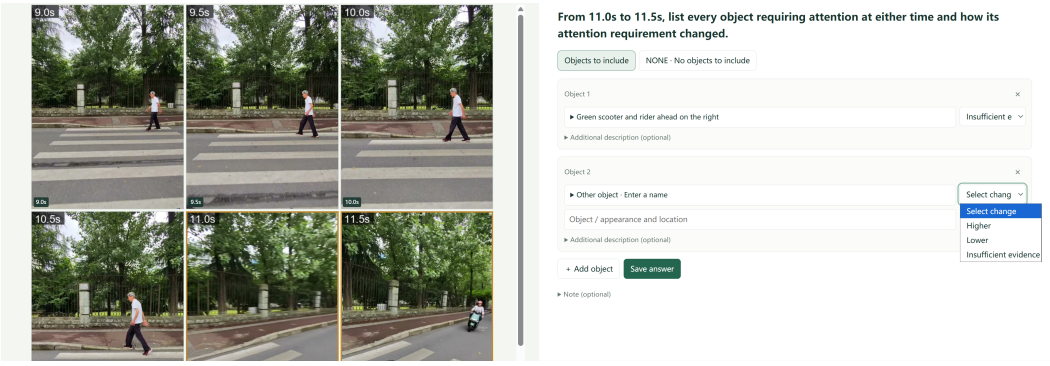


Figure 5: **Human evaluation interface for FFG.** Annotators select objects and change labels, add unlisted objects, or return an empty set.

C Evaluation Coverage and Additional Results

The benchmark contains 577 distinct questions from 95 reviewed clips. Wearer translation, turning, external object motion, passage relations, and entity discovery contain 96, 94, 136, 141, and 110 questions, respectively. Human and model evaluation uses this same question bank and all five task types. Each model is scored on every available, adjudicated response rather than on a global intersection of model coverage. Table 9 lists the denominators used in Table 2.

Table 9: Exact No History counts underlying Table 2. W1/W2/O1/T1 denote wearer translation, turning, object motion, and passage relations. MCQ entries are correct/scorable questions. FFG entries give scorable questions (N), empty/nonempty reference sets, summed Discovery scores ($\sum s$), exact set-and-label answers, correct/additional/missed objects (TP/FP/FN), and correct/assessable labels.

(a) MCQ counts and coverage

Model	W1	W2	O1	T1	MCQ total	FFG	All	Excl.
GPT-6 Astra	86/96	63/94	123/136	87/141	359/467	110	577	0
Gemini 3.8 Flash	69/96	66/94	119/136	82/141	336/467	110	577	0
Gemini 3.1 Pro	64/96	45/94	115/136	69/141	293/467	110	577	0
GPT-5.6 Sol	61/96	56/94	118/136	53/141	288/467	110	577	0
Qwen3.8 Max	73/96	57/93	114/136	46/139	290/464	109	573	4[†]
Qwen3.8 Flash	72/96	45/93	112/136	46/141	275/466	109	575	2[†]
Llama 4 Scout	36/96	55/94	90/136	44/141	225/467	110	577	0
Llama 4 Maverick	36/96	50/94	106/136	41/141	233/467	110	577	0

(b) FFG counts

Model	N	Empty/nonempty	$\sum s$	Exact	TP/FP/FN	Labels
GPT-6 Astra	110	65/45	66.00	60	43/69/10	36/43
Gemini 3.8 Flash	110	65/45	58.50	51	36/44/17	24/36
Gemini 3.1 Pro	110	65/45	51.50	41	43/71/10	30/43
GPT-5.6 Sol	110	65/45	42.00	32	47/110/6	34/47
Qwen3.8 Max	109	64/45	33.50	25	40/95/13	27/40
Qwen3.8 Flash	109	65/44	20.00	13	39/128/12	23/39
Llama 4 Scout	110	65/45	47.00	34	23/49/30	9/23
Llama 4 Maverick	110	65/45	34.50	19	27/66/26	11/27

Discovery $= 100 \sum s/N$; FFG Overall $= 100\,\mathrm{Exact}/N$; Labels $=$ 100 times the stated label fraction. [†]Exclusions are provider-content-filtered responses, omitted from the affected model's numerator and denominator.

Provider content filtering prevented valid No History responses on four questions for Qwen3.8 Max (two passage-relation questions, one turning question, and one entity-discovery question) and two for Qwen3.8 Flash (one turning question and one entity-discovery question). The requests were attempted but did not yield valid answers. Each missing response is excluded from that model's numerator and denominator, rather than counted as incorrect; valid responses from other models to the same question remain included. Overall MCQ accuracy therefore uses 464 questions for Qwen3.8 Max, 466 for Qwen3.8 Flash, and 467 for each other model. FFG uses 109 questions for each Qwen model and 110 for each other model.

Two examples illustrate the scoring rules. In the turning example described in Appendix B.1, the reference includes a rightward head/camera turn after an initial period without a discernible turn; all eight model responses are incorrect. In an entity-discovery question, GPT-5.6 Sol reports two correctly matched entities, including a pedestrian, and one additional object against a reference set of two entities. Its object score is therefore $\max(0, (2-1)/2) = 0.5$, and one of the two matched change labels is correct. Both questions are included in the reported scores.

An auxiliary aggregate score sums correct MCQ answers and FFG object scores, then divides by the model's scorable question count; it excludes conditional label accuracy. The values are 73.66% for GPT-6 Astra, 68.37% for Gemini 3.8 Flash, 59.71% for Gemini 3.1 Pro, 57.19% for GPT-5.6 Sol, 56.46% for Qwen3.8 Max, 51.30% for Qwen3.8 Flash, 47.14% for Llama 4 Scout, and 46.36% for Llama 4 Maverick. This aggregate determines the model row order but does not replace task-specific reporting.

C.1 Full History Evaluation

The History evaluation covers 293 questions across 54 current clips with an immediately preceding clip. These are the same questions evaluated without history, not additional benchmark questions. In task order, the counts are 55, 53, 64, 67, and 54. Table 10 reports all scorable answers for each model. Missing responses and one Llama 4 Scout response whose entity identity could not be established are excluded; the shared comparison in the main text instead uses the same 257 questions for every model and both conditions. The shared comparison contains 53 wearer translation, 45

Table 10: History results over each model's scorable set. W1/W2/O1/T1 denote wearer translation, turning, object motion, and passage relations. Score formats follow Table 11; label counts appear in Table 12. The lower panel gives the auxiliary aggregate numerator (MCQ correct plus summed Discovery scores) and denominator.

Model	MCQ	W1	W2	O1	T1	FFG Overall	Discovery	Labels (%)
History: full coverage								
GPT-6 Astra	**195/239**	**45/55**	42/53	**61/64**	**47/67**	**38/54**	**41.50/54**	**78.95**
Gemini 3.8 Flash	179/239	42/55	**44/53**	54/64	39/67	32/54	36.50/54	64.71
Gemini 3.1 Pro	147/239	34/55	23/53	56/64	34/67	25/54	28.50/54	76.47
GPT-5.6 Sol	159/239	38/55	36/53	60/64	25/67	23/54	26.00/54	70.00
Qwen3.8 Max	148/232	42/55	30/51	57/64	19/62	13/51	15.00/51	60.00
Qwen3.8 Flash	134/226	38/53	24/47	55/62	17/64	5/44	9.00/44	57.14
Llama 4 Scout	112/239	26/55	30/53	42/64	14/67	14/51	17.50/51	60.00
Llama 4 Maverick	129/239	33/55	19/53	56/64	21/67	10/54	14.00/54	63.64

Model	Scorable/all	Excluded	Auxiliary aggregate
GPT-6 Astra	293/293	0	236.50/293
Gemini 3.8 Flash	293/293	0	215.50/293
Gemini 3.1 Pro	293/293	0	175.50/293
GPT-5.6 Sol	293/293	0	185.00/293
Qwen3.8 Max	283/293	10	163.00/283
Qwen3.8 Flash	270/293	23	143.00/270
Llama 4 Scout	290/293	3	129.50/290
Llama 4 Maverick	293/293	0	143.00/293

turning, 62 external object motion, 58 passage relation, and 39 discovery questions. FFG label accuracy uses successfully matched reference entities separately in each condition. Humans did not complete the History condition. Table 11 provides absolute scores for both conditions on this shared set; Figure 4 visualizes their differences. For each condition, the video recorders provide reference answers based on the filming circumstances and the permitted visual evidence. The same condition-specific references apply to every model.

Table 11: Paired results on 257 questions. W1/W2/O1/T1 denote wearer translation, turning, object motion, and passage relations. MCQ and FFG Overall give correct/scorable counts; Discovery gives summed scores/scorable counts; Labels gives conditional accuracy (%; counts in Table 12). Bold/underlining mark best/second-best scores per condition.

Model	MCQ	W1	W2	O1	T1	FFG Overall	Discovery	Labels (%)
No History								
GPT-6 Astra	**174/218**	**46/53**	34/45	54/62	**40/58**	**24/39**	**27.50/39**	**69.23**
Gemini 3.8 Flash	163/218	37/53	**36/45**	**57/62**	33/58	21/39	22.00/39	66.67
Gemini 3.1 Pro	146/218	36/53	26/45	55/62	29/58	16/39	18.50/39	**69.23**
GPT-5.6 Sol	143/218	33/53	30/45	**57/62**	23/58	12/39	14.00/39	66.67
Qwen3.8 Max	136/218	38/53	27/45	54/62	17/58	11/39	14.00/39	45.45
Qwen3.8 Flash	130/218	36/53	23/45	52/62	19/58	4/39	6.00/39	40.00
Llama 4 Scout	93/218	21/53	23/45	39/62	10/58	16/39	20.00/39	50.00
Llama 4 Maverick	109/218	21/53	24/45	51/62	13/58	7/39	12.00/39	37.50
History								
GPT-6 Astra	**183/218**	**44/53**	38/45	**59/62**	**42/58**	**26/39**	**29.50/39**	71.43
Gemini 3.8 Flash	170/218	41/53	**40/45**	53/62	36/58	24/39	27.00/39	64.29
Gemini 3.1 Pro	143/218	33/53	22/45	56/62	32/58	18/39	21.50/39	**75.00**
GPT-5.6 Sol	150/218	37/53	32/45	**59/62**	22/58	16/39	18.00/39	66.67
Qwen3.8 Max	143/218	41/53	28/45	56/62	18/58	11/39	13.00/39	54.55
Qwen3.8 Flash	134/218	38/53	24/45	55/62	17/58	4/39	7.00/39	54.55
Llama 4 Scout	102/218	24/53	26/45	41/62	11/58	11/39	13.50/39	62.50
Llama 4 Maverick	121/218	32/53	18/45	56/62	15/58	8/39	11.50/39	50.00

C.2 Labels and Evidence Boundaries

External-object motion has four labels: stationary, moving, motion state changed, and uncertain. A state change requires a confirmed start or stop, not a direction change alone. Passage-relation answers distinguish higher or lower attention requirements, irrelevance, and insufficient evidence. FFG returns qualifying entities and their change labels, or an empty set when none qualifies. A change label cannot justify including an irrelevant object.

Qualification requires concrete visual evidence of physical conflict, insufficient passage space, or an avoidance adjustment that must begin within two seconds. Visibility, approach, motion, and salience are not sufficient. Nor is an inability to rule out hypothetical danger. Riders and their vehicles form one occupied unit. Traffic rules, signal colors, paths separated by adequate space, and crossings at clearly different times do not substitute for visual evidence of a conflict. In this benchmark, no relevant object has an unchanged attention requirement across the queried time pair; insufficient evidence means the visual evidence cannot resolve the comparison. Both conditions permit all allowed current frames, rather than only endpoint images. History adds only the immediately preceding clip, never earlier answers.

C.3 Recognizing Insufficient Evidence

Table 13 reports the diagnostic subset discussed in Section 5.4. The reference answer is insufficient evidence in all 16 questions: six wearer-translation and ten passage-relation questions. Counts are separated by task and use each model's available responses; Qwen3.8 Max has eight scorable passage-relation responses because its two content-filtered cases are excluded.

Table 12: FFG counts for the paired and full History evaluations. N is the number of scorable questions; empty/nonempty refer to reference sets. Discovery $= 100 \sum s/N$, FFG Overall $= 100\,\mathrm{Exact}/N$, and Labels is 100 times the stated correct/assessable fraction.

Model	N	Empty/nonempty	$\sum s$	Exact	TP/FP/FN	Labels
Paired No History (257 questions)						
GPT-6 Astra	39	24/15	27.50	24	13/9/4	9/13
Gemini 3.8 Flash	39	24/15	22.00	21	12/12/5	8/12
Gemini 3.1 Pro	39	24/15	18.50	16	13/20/4	9/13
GPT-5.6 Sol	39	24/15	14.00	12	15/27/2	10/15
Qwen3.8 Max	39	24/15	14.00	11	11/28/6	5/11
Qwen3.8 Flash	39	24/15	6.00	4	10/38/7	4/10
Llama 4 Scout	39	24/15	20.00	16	8/14/9	4/8
Llama 4 Maverick	39	24/15	12.00	7	8/23/9	3/8
Paired History (257 questions)						
GPT-6 Astra	39	23/16	29.50	26	14/7/4	10/14
Gemini 3.8 Flash	39	23/16	27.00	24	14/8/4	9/14
Gemini 3.1 Pro	39	23/16	21.50	18	12/15/6	9/12
GPT-5.6 Sol	39	23/16	18.00	16	15/21/3	10/15
Qwen3.8 Max	39	23/16	13.00	11	11/29/7	6/11
Qwen3.8 Flash	39	23/16	7.00	4	11/40/7	6/11
Llama 4 Scout	39	23/16	13.50	11	8/20/10	5/8
Llama 4 Maverick	39	23/16	11.50	8	8/24/10	4/8
Full History coverage (293 questions)						
GPT-6 Astra	54	33/21	41.50	38	19/9/5	15/19
Gemini 3.8 Flash	54	33/21	36.50	32	17/11/7	11/17
Gemini 3.1 Pro	54	33/21	28.50	25	17/23/7	13/17
GPT-5.6 Sol	54	33/21	26.00	23	20/30/4	14/20
Qwen3.8 Max	51	31/20	15.00	13	15/46/8	9/15
Qwen3.8 Flash	44	26/18	9.00	5	14/46/7	8/14
Llama 4 Scout	51	32/19	17.50	14	10/29/11	6/10
Llama 4 Maverick	54	33/21	14.00	10	11/36/13	7/11

Table 13: **Recognizing insufficient evidence.** Entries show correct answers over questions with an insufficient evidence reference. This diagnostic subset contains six wearer translation and ten passage relation questions. Two passage relation questions have no valid Qwen3.8 Max response because of provider content filtering and are excluded from its denominator.

Model	Wearer translation	Passage relation
GPT-6 Astra	0 / 6	**2 / 10**
Gemini 3.8 Flash	0 / 6	0 / 10
Gemini 3.1 Pro	0 / 6	0 / 10
GPT-5.6 Sol	0 / 6	0 / 10
Qwen3.8 Max	0 / 6	0 / 8
Qwen3.8 Flash	0 / 6	0 / 10
Llama 4 Scout	0 / 6	0 / 10
Llama 4 Maverick	0 / 6	0 / 10

C.4 Complete Entity Selection Comparison

Figure 6 reports all eight models on the same 108 entity discovery questions (64 empty and 44 nonempty, with 51 reference entities). The two FFG questions without a valid Qwen response are excluded only for these shared diagnostics; the headline table retains every valid response for each model. Figure 3 averages the two evaluated models in each family. The pairs are GPT-6 Astra and GPT-5.6 Sol; Gemini 3.8 Flash and Gemini 3.1 Pro; Qwen3.8 Max and Qwen3.8 Flash; and Llama 4 Scout and Llama 4 Maverick. Each metric is averaged separately, without pooling reports across models.

Table 14: **Entity discovery precision and recall.** Counts are pooled within each model over the same 108 No History questions (64 empty and 44 nonempty), containing 51 reference entities. TP counts matched entities, FP additional reports including duplicates, and FN omitted reference entities under one-to-one matching. Micro-precision is $100\,\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and micro-recall is $100\,\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Bold and underlining mark the best and second-best model results for each rate.

Model	TP	FP	FN	Precision (%)	Recall (%)
GPT-6 Astra	41	69	10	37.27	80.39
Gemini 3.8 Flash	35	44	16	**44.30**	68.63
Gemini 3.1 Pro	42	69	9	<u>37.84</u>	<u>82.35</u>
GPT-5.6 Sol	45	109	6	29.22	**88.24**
Qwen3.8 Max	38	95	13	28.57	74.51
Qwen3.8 Flash	39	125	12	23.78	76.47
Llama 4 Scout	22	49	29	30.99	43.14
Llama 4 Maverick	26	65	25	28.57	50.98

Figure 7 further separates the same reference entities into correctly labeled matches, incorrectly labeled matches, and omissions.

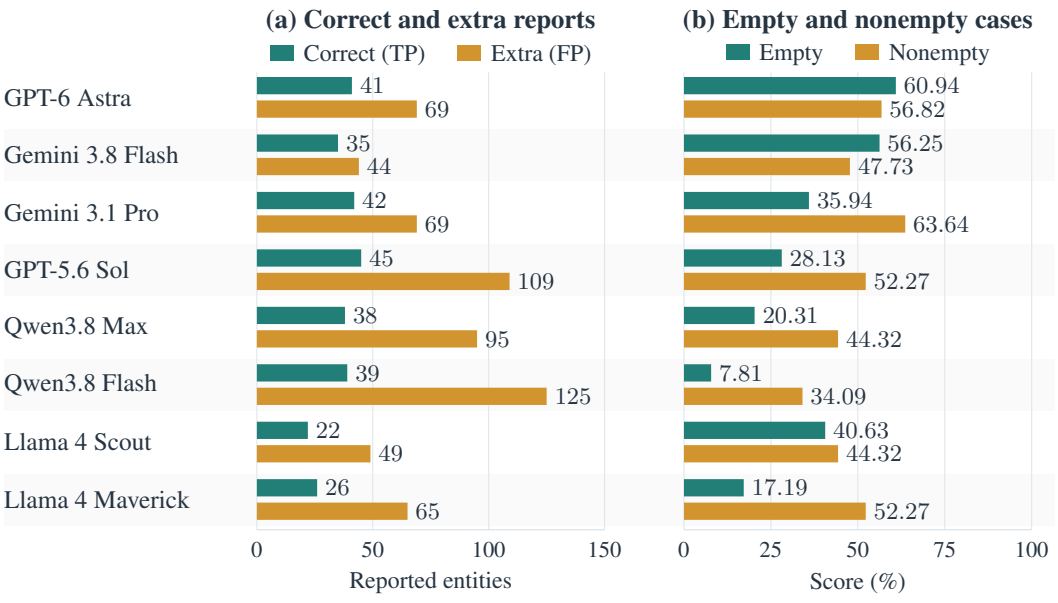


Figure 6: **Entity selection across all eight models.** (a) Correct detections and additional reports, including duplicates. (b) Accuracy on 64 empty reference questions and mean object score on 44 nonempty questions. These 108 questions have valid responses from every model and contain 51 reference entities. Always returning an empty set scores 59.26%; Astra matches this baseline and the other models fall below it.

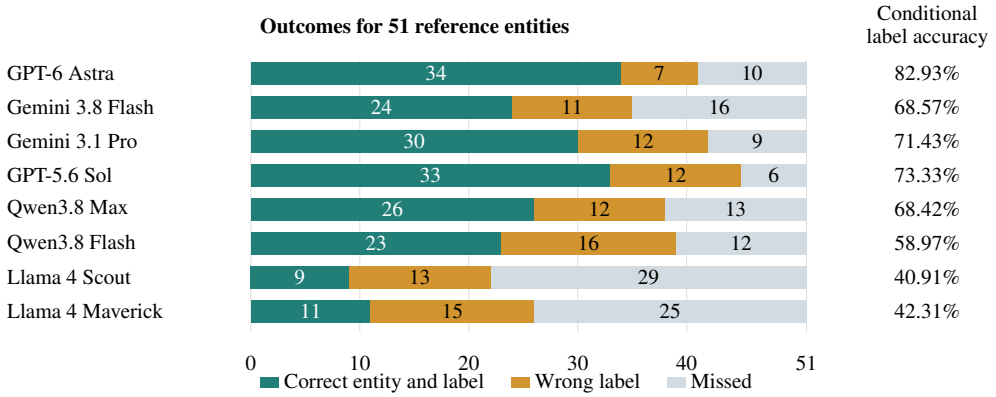


Figure 7: **Local correctness and incomplete coverage.** Every bar contains the same 51 reference entities from the 108 FFG questions with valid responses from every model. Conditional label accuracy measures correct change judgments among matched entities and excludes omissions and false reports. Extra reports are shown in Figure 6(a).

D Evaluation Prompts

This appendix provides the English prompts used for the five evaluation tasks. Each question is submitted independently with timestamped frame grids. Exactly one context condition from Section D.1 is inserted at {history_instruction} in the corresponding task template. Other braced fields are replaced with the video and clip identifiers, frame timestamps, queried times, question text, and answer options. Object descriptions are supplied only for tasks with a specified target; entity discovery instead uses a neutral identifier for the queried time pair. The instructions and response formats below are shared across models. Responses are required in JavaScript Object Notation (JSON).

D.1 Context Conditions

The condition instructions are listed once here and inserted into every independent request. No History uses only the current clip. History also includes the immediately preceding clip from the same video. Both conditions require an answer about the current question.

No History

[NO-HISTORY CONDITION]
Use only the current slice images and the current question. Do not assume access to earlier slices, earlier answers, or events outside the supplied evidence.

History

[PREVIOUS-SLICE VISUAL-HISTORY CONDITION]
For current slice S_k, the input contains only the frames of the immediately preceding slice S_(k-1) of the same video and the current slice S_k. No frames from S_(k-2) or earlier slices are provided. For the first slice, use only its current-slice frames.
No previous questions, model answers, human answers, or textual summaries of earlier events are provided. Treat this question as an independent request, not a continuing conversation.
Use the frames' absolute timestamps to determine event order and associate the same object across these two slices. Visual evidence from the immediately preceding slice may help infer current identity, motion state, and trajectory, and distinguish object motion from wearer motion.
Do not assume an earlier state persists automatically. Integrate the previous and current frames and update the inference when current evidence supports a state change. If occlusion, a

temporal gap, or uncertain identity prevents a reliable inference, retain uncertainty rather than inventing continuity.
Answer only the current question for its specified interval or timestamp pair. Relevance in the previous slice alone does not establish relevance at a queried endpoint. Use only the supplied frames; do not assume access to older slices or unprovided later images.

D.2 WEARER TRANSLATION

[LABEL DEFINITIONS]
STATIONARY: There is sufficient visual evidence that the wearer undergoes no discernible positional translation in the physical world during the specified interval.
MOVING: The wearer undergoes confirmable positional translation at any time in the specified interval, including starting, stopping, or changing speed.
UNCERTAIN: The available frames are insufficient to reliably distinguish STATIONARY from MOVING, or the visual evidence is contradictory.

[CURRENT INPUT]
Video: {video_id}
Current slice: {slice_id}
You will see contact sheets of first-person video frames. Each frame carries its absolute timestamp in the original video.
Current-slice frame timestamps: [{allowed_times_s}].
Available current-slice visual evidence spans {first_frame_time_s}s to {last_frame_time_s}s in the original video.
Within each contact sheet, frames are ordered left to right, then top to bottom. Contact sheets are ordered from earlier slices to later slices; the final sheet is the current slice.
{history_instruction}

[TASK]
Determine whether the wearer undergoes real positional translation from {t1_s}s to {t2_s}s.
1. Global image translation, rotation, or blur may be caused by head or camera rotation and does not by itself prove wearer translation.
2. Use ground texture, nearby fixed structures, perspective-scale change, parallax, and cross-frame relative positions to determine real translation.
3. Do not mistake the motion of other people or vehicles for wearer motion.
4. Judge only whether real translation occurs. Do not judge direction, speed trend, turn type, or risk.
5. If evidence is insufficient, select UNCERTAIN. Do not guess that the wearer starts moving based on scene conventions.

[QUESTION]
{question_text}
{options_block}
Return only JSON: {"answer":"one of {option_letters}"}

[EVALUATION BOUNDARY]
Answer only from the attached images, permitted history, and this prompt. Do not use tools, access local files, browse the web, or request external information. Do not infer reference answers from question wording, option order, or question identifiers. Return only the specified JSON, without explanations, Markdown, or additional fields.

D.3 VIEWPOINT AND PATH TURNING

[EVENT COMPONENT DEFINITIONS]
NONE: Sufficient visual evidence supports no discernible gaze-direction change and no actual travel-path turn during the specified interval.
HEAD_TURN_LEFT / **HEAD_TURN_RIGHT:** Sufficient visual evidence supports a leftward or rightward head/camera orientation change while the wearer remains stationary or continues along a translational path without a corresponding turn.
PATH_TURN_LEFT / **PATH_TURN_RIGHT:** Sufficient visual evidence supports real wearer translation and a travel path that curves left or right in the physical world.
UNCERTAIN: The turning source, direction, or temporal order cannot be reliably determined, or visual evidence is contradictory.

[CURRENT INPUT]
Video: {video_id}
Current slice: {slice_id}
You will see contact sheets of first-person video frames. Each frame carries its absolute timestamp in the original video.
Current-slice frame timestamps: [{allowed_times_s}].
Available current-slice visual evidence spans {first_frame_time_s}s to {last_frame_time_s}s in the original video.
Within each contact sheet, frames are ordered left to right, then top to bottom. Contact sheets are ordered from earlier slices to later slices; the final sheet is the current slice.
{history_instruction}

[OPTION INTERPRETATION]
Each event option lists all discernible turning events within the queried interval. Unlisted periods in that option indicate no discernible head/camera rotation or travel-path turn, not missing information.
This is the hypothesis asserted by the option, not visual evidence: verify it against the permitted images. Do not infer an absence of turning merely because an option omits it. Select UNCERTAIN when the visual evidence cannot reliably determine the turning source, direction, or temporal order.
No discernible turning does not mean the wearer is stationary; the wearer may continue moving along an unchanged path.

[TASK]
Select the event hypothesis that most accurately describes the source, direction, and temporal order of turning from {t1_s}s to {t2_s}s.
1. Interpret left and right from the wearer's perspective.
2. Image rotation alone does not establish PATH_TURN. A path turn requires evidence of translation and changing travel direction, such as ground trajectory, nearby-structure parallax, and sustained motion.
3. HEAD_TURN also requires positive evidence. Failure to establish PATH_TURN does not automatically establish HEAD_TURN.
4. Select UNCERTAIN when head/camera rotation cannot be reliably distinguished from path turning.
5. Use the supplied timestamps to judge direction and temporal order. Do not invent unprovided timestamps. If several events are identifiable, choose the option matching their composition and order.
6. NONE means evidence supports no turn; UNCERTAIN means evidence is insufficient for a reliable determination. Do not interchange them.
7. Judge only turning, not speed trend, external-object motion, or risk.

[QUESTION]
{question_text}
{options_block}
Return only JSON: `{"answer":"one of {option_letters}"}`

[EVALUATION BOUNDARY]
Answer only from the attached images, permitted history, and this prompt. Do not use tools, access local files, browse the web, or request external information. Do not infer reference answers from question wording, option order, or question identifiers. Return only the specified JSON, without explanations, Markdown, or additional fields.

D.4 External Object Motion

[LABEL DEFINITIONS]
STATIONARY: There is sufficient visual evidence that the specified object undergoes no discernible positional translation in the physical world during the interval. A stationary person, parked vehicle, or fixed structure may be STATIONARY.
MOVING: Sufficient evidence supports that the object remains in translational motion over the interval, without a confirmable transition between stationary and moving. Turning, acceleration, or deceleration without such a transition remains MOVING.
MOTION_STATE_CHANGED: The same object shows a confirmable transition from stationary to moving or from moving to stationary within the interval, including one or more starts or stops. A change of direction or speed alone does not qualify.
UNCERTAIN: Evidence is insufficient to reliably determine the object's real motion state or maintain its identity.

[CURRENT INPUT]
Video: {video_id}
Current slice: {slice_id}
You will see contact sheets of first-person video frames. Each frame carries its absolute timestamp in the original video.
Current-slice frame timestamps: [{allowed_times_s}].
Available current-slice visual evidence spans {first_frame_time_s}s to {last_frame_time_s}s in the original video.
Within each contact sheet, frames are ordered left to right, then top to bottom. Contact sheets are ordered from earlier slices to later slices; the final sheet is the current slice.
{history_instruction}
Specified object: {target_object}
The description identifies the target only; it is not a motion-state or risk conclusion.

[TASK]
Determine the observed motion state of this object itself from {t1_s}s to {t2_s}s.
1. Distinguish real object motion from wearer translation and head or camera rotation. Use fixed backgrounds, ground references, and cross-frame parallax.
2. Image-plane motion alone does not establish physical motion. A lack of visible motion alone is not sufficient evidence of STATIONARY.
3. Treat a rider and the ridden vehicle as one motion unit while they move together.
4. Judge one specified entity or rider-vehicle unit, not an aggregate of independent entities. Inspect the full queried interval: a confirmed start or stop takes precedence over a single endpoint state. If the interval contains a confirmed stationary-to-moving or moving-to-stationary transition, select MOTION_STATE_CHANGED even when both endpoints have the same state.
5. A fixed structure being STATIONARY does not imply it is safe. Do not judge attention priority or future risk in this task.
6. If identity or motion cannot be reliably determined, select UNCERTAIN instead of guessing.

[QUESTION]
{question_text}
{options_block}
Return only JSON: `{"answer":"one of {option_letters}"}`

[EVALUATION BOUNDARY]
Answer only from the attached images, permitted history, and this prompt. Do not use tools, access local files, browse the web, or request external information. Do not infer reference answers from question wording, option order, or question identifiers. Return only the specified JSON, without explanations, Markdown, or additional fields.

D.5 Passage Relations for Specified Objects

[TASK AND INPUT]
Compare the attention required by external objects at {t2_s}s with {t1_s}s. Attention means monitoring, preparation, or action adjustment warranted by a supported physical conflict or safe-passage constraint, not visibility or general interest.
Video: {video_id}. Current slice: {slice_id}.
Current-slice timestamps: [{allowed_times_s}]. Visual evidence spans {first_frame_time_s}s to {last_frame_time_s}s.
Frames within each contact sheet run left to right, then top to bottom; sheets run from earlier to later slices, with the current slice last.
{history_instruction}

[EVIDENCE AND MOTION]
Use all condition-permitted frames to associate identities and estimate each endpoint state. This is a retrospective endpoint comparison, not future-hidden online forecasting.
At each endpoint, separately project the next 2 seconds using the wearer's and object's motion trends supported at that endpoint. Later stopping, turning, successful avoidance, or absence of collision must not replace the earlier trend assessment.
Infer wearer translation, direction, speed trend, and actual path turning before assessing relative motion. Head/camera rotation alone does not establish a path turn. When evidence supports no discernible wearer translation, keep the occupied position fixed; do not assume a later start. Do not carry earlier speed forward after stopping.
Do not invent sudden movement, cooperative avoidance, unseen trajectories, or precise distances, speeds, or probabilities. Occlusion or leaving view alone proves neither safety nor reduced attention. If unresolved motion or identity affects the comparison, retain uncertainty.

[STEP 1: DECIDE WHETHER AN OBJECT QUALIFIES]
Include an object only when concrete visual evidence at at least one queried endpoint supports a reasonable possibility of physical conflict or insufficient safe-passage space under the wearer's and object's motion trends at that endpoint. The conflict or constraint must be relevant within the next 2 seconds, or an adjustment to avoid it must need to begin within those 2 seconds even if contact would occur later.
Assess relative position, trajectory, occupied width, timing, and available passage space. A collision need not be certain, and avoidance need not already have begun.
Determine qualification before assigning a change label. Do not select objects merely because they are visible, moving, approaching, nearby, or visually prominent. If the specified object clearly fails to qualify at both endpoints, select NOT_RELEVANT. Do not report a change label to justify including an object that fails the gate.
Normal passage along already separated paths with adequate space is not an object-induced need to adjust. Paths crossing at clearly different times do not alone establish conflict. Parallel motion alone proves neither conflict nor safety. Inability to rule out hypothetical danger is not positive inclusion evidence.
Apply this gate equally to moving objects and static structures. Treat each rider and their ridden vehicle as one occupied unit with nonzero width. Do not infer safety or relevance from traffic rules, right of way, or signal colors; do not list traffic signals.

[STEP 2: COMPARE QUALIFYING OBJECTS]
Compare the overall evidence for conflict likelihood, urgency, passage constraints, and necessary action adjustment. Do not compare distance or image position alone, and do not use the slice maximum instead of either endpoint.

HIGHER: Supported attention requirement is greater at the second endpoint.
LOWER: Supported attention requirement is smaller at the second endpoint.
INSUFFICIENT_EVIDENCE: Relevant endpoint states or their comparison cannot be resolved reliably, including conflicting cues without a supported overall direction.
When both endpoints are assessable, clearly nonqualifying to qualifying means HIGHER; the reverse means LOWER. Uncertainty is not proof of nonqualification.

[JUDGE THE SPECIFIED OBJECT]
Object: {target_object}
This name identifies the object; it does not imply relevance.
1. If evidence supports nonqualification at both endpoints, select NOT_RELEVANT.
2. If relevance cannot be determined reliably, or the required comparison is unresolved, select INSUFFICIENT_EVIDENCE.
3. Otherwise, at least one endpoint qualifies; select HIGHER or LOWER using Step 2.
Question: {question_text}
{options_block}
Return only JSON: {"answer":"one of {option_letters}"}

[RESPONSE BOUNDARY]
Use only the attached images, permitted history, and this prompt. Do not use tools or external information, or infer reference answers from wording, identifiers, or option order. Return only the specified JSON, using English for any free-text field, without explanations, absolute attention levels, rankings, timelines, or separate predictions.

D.6 AUTONOMOUS RELEVANT ENTITY DISCOVERY

[TASK AND INPUT]
Compare the attention required by external objects at {t2_s}s with {t1_s}s. Attention means monitoring, preparation, or action adjustment warranted by a supported physical conflict or safe-passage constraint, not visibility or general interest.
Video: {video_id}. Current slice: {slice_id}.
Current-slice timestamps: [{allowed_times_s}]. Visual evidence spans {first_frame_time_s}s to {last_frame_time_s}s.
Frames within each contact sheet run left to right, then top to bottom; sheets run from earlier to later slices, with the current slice last.
{history_instruction}

[EVIDENCE AND MOTION]
Use all condition-permitted frames to associate identities and estimate each endpoint state. This is a retrospective endpoint comparison, not future-hidden online forecasting.
At each endpoint, separately project the next 2 seconds using the wearer's and object's motion trends supported at that endpoint. Later stopping, turning, successful avoidance, or absence of collision must not replace the earlier trend assessment.
Infer wearer translation, direction, speed trend, and actual path turning before assessing relative motion. Head/camera rotation alone does not establish a path turn. When evidence supports no discernible wearer translation, keep the occupied position fixed; do not assume a later start. Do not carry earlier speed forward after stopping.
Do not invent sudden movement, cooperative avoidance, unseen trajectories, or precise distances, speeds, or probabilities. Occlusion or leaving view alone proves neither safety nor reduced attention. If unresolved motion or identity affects the comparison, retain uncertainty.

[STEP 1: DECIDE WHETHER AN OBJECT QUALIFIES]
Include an object only when concrete visual evidence at at least one queried endpoint supports a reasonable possibility of physical conflict or insufficient safe-passage space under the wearer's and object's motion trends at that endpoint. The conflict or constraint must be relevant within the next 2 seconds, or an adjustment to avoid it must need to begin within those 2 seconds even if contact would occur later.

Assess relative position, trajectory, occupied width, timing, and available passage space. A collision need not be certain, and avoidance need not already have begun.
Determine qualification before assigning a change label. Do not select objects merely because they are visible, moving, approaching, nearby, or visually prominent. Do not include objects that fail to qualify. If no object qualifies, return an empty objects array. Do not report a change label to justify including an object that fails the gate.
Normal passage along already separated paths with adequate space is not an object-induced need to adjust. Paths crossing at clearly different times do not alone establish conflict. Parallel motion alone proves neither conflict nor safety. Inability to rule out hypothetical danger is not positive inclusion evidence.
Apply this gate equally to moving objects and static structures. Treat each rider and their ridden vehicle as one occupied unit with nonzero width. Do not infer safety or relevance from traffic rules, right of way, or signal colors; do not list traffic signals.

[STEP 2: COMPARE QUALIFYING OBJECTS]
Compare the overall evidence for conflict likelihood, urgency, passage constraints, and necessary action adjustment. Do not compare distance or image position alone, and do not use the slice maximum instead of either endpoint.
HIGHER: Supported attention requirement is greater at the second endpoint.
LOWER: Supported attention requirement is smaller at the second endpoint.
INSUFFICIENT_EVIDENCE: Relevant endpoint states or their comparison cannot be resolved reliably, including conflicting cues without a supported overall direction.
When both endpoints are assessable, clearly nonqualifying to qualifying means HIGHER; the reverse means LOWER. Uncertainty is not proof of nonqualification.

[DISCOVER QUALIFYING OBJECTS WITHOUT TARGET HINTS]
Independently identify every object with positive qualification evidence at at least one endpoint. Do not report an object solely because relevance is uncertain at both endpoints.
For each included object, apply Step 2. If qualification is established at one endpoint but the other endpoint or cross-time identity is unresolved, use INSUFFICIENT_EVIDENCE. Omit objects nonqualifying at both endpoints; do not emit NOT_RELEVANT.
Use neutral, localizable descriptions. Report each physical object once; do not merge unrelated people or vehicles. There is no fixed or hidden target count.
If no object has positive qualification evidence, return objects: []. This means no qualifying object was established, not proof of absolute safety.

[OUTPUT]
Return one comparison with these exact fields. Replace illustrative values; select one attention_change label per object:

```
{"comparisons": [{"pair_id": "{pair_id}",
 "t1": "{t1_s}s", "t2": "{t2_s}s", "objects": [{
 "object": "neutral, localizable description",
 "attention_change": "HIGHER | LOWER | INSUFFICIENT_EVIDENCE"}]}]}
```

For an empty result, retain the comparison fields and set only objects to [].

[RESPONSE BOUNDARY]
Use only the attached images, permitted history, and this prompt. Do not use tools or external information, or infer reference answers from wording, identifiers, or option order. Return only the specified JSON, using English for any free-text field, without explanations, absolute attention levels, rankings, timelines, or separate predictions.