LTDRIVEQA: BENCHMARKING SPATIOTEMPORAL INTELLIGENCE IN MULTI-VIEW LONG-TAIL DRIVING SCENARIOS

Anonymous authors
Paper under double-blind review

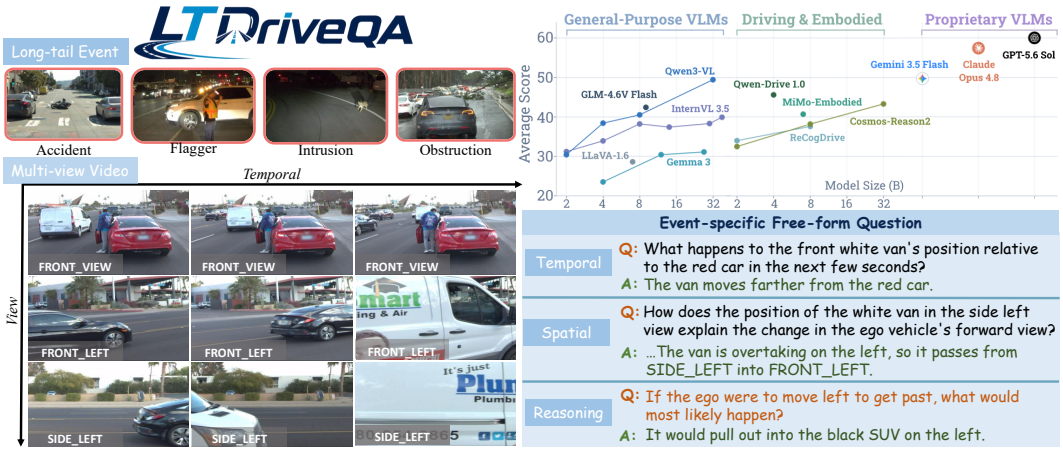


Figure 1: **Overview of LTDriveQA. Left:** Long-tail driving events illustrated with synchronized multi-view video observations. **Top right:** Model performance shows substantial headroom on LTDriveQA across different model families. **Bottom right:** Representative event-specific free-form questions illustrate temporal, cross-view spatial, and driving reasoning. LTDriveQA differs from generic scene-level VQA by grounding questions in specific events and targeting temporal and cross-view evidence.

ABSTRACT

Driving visual question answering (VQA) provides a unified framework for assessing whether vision-language models (VLMs) and Vision-Language-Action (VLA) models can understand dynamic road scenes and reason about driving decisions. However, existing benchmarks often rely on generic templates, weakly probe temporal and cross-view reasoning, and underutilize long-tail driving events. We introduce LTDriveQA, an event-centered benchmark of scene-specific questions designed to probe temporal and cross-view reasoning. It spans Perception, Prediction, Planning, and Reasoning, supports both multiple-choice and open-ended evaluation. A scalable recursive benchmark construction pipeline refines questions through diagnostic feedback, evidence grounding, answer annotation, and restricted-input shortcut checks. Across 25 models, GPT-5.6-Sol leads with an overall score of 60.0; Qwen3-VL-8B and Qwen-Drive-1.0 score 40.5 and 45.6, respectively. Evidence ablations show aggregate visual gains across all open-source models, with temporal gains strongest in Perception and surround-view gains consistent across three models on multi-view questions. By grounding evaluation in challenging real-world driving, LTDriveQA probes the scene understanding and decision reasoning that driving VLMs and VLAs rely on, and provides a testbed for spatiotemporal intelligence in dynamic multi-view environments.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have motivated interest in using language as an interface for understanding and reasoning about complex driving scenes (Sima et al., 2024; Tian et al.,

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.

Table 1: **Design comparison of driving VQA benchmarks.** Data Format summarizes inputs, Question Design compares reasoning and construction properties, and Question Type indicates answer formats. Temporal and Cross-view denote targeted reasoning requirements; Scene-specific and Free-form indicate grounding in individual scenes beyond fixed templates. Audit indicates construction-time quality checks, including validity, non-triviality, evidence grounding, and shortcut resistance.

Benchmark	Data Format			Question Design					Question Type	
	Long-tail	Video	Multi-view	Temporal	Cross-view	Scene-specific	Free-form	Audit	Closed	Open
nuScenes-QA (Qian et al., 2024)	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗
DriveLM (Sima et al., 2024)	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓
LingoQA (Marcu et al., 2024)	✗	✓	✗	✓	✗	✓	✓	✗	✗	✓
DriveMLLM (Guo et al., 2025)	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
DriveVLM (Tian et al., 2024)	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓
DriveBench (Xie et al., 2025)	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓
DVBench (Zeng et al., 2025)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
VRU-Accident (Kim et al., 2025)	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
WaymoQA (Yu et al., 2025)	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
DriveSpatial (Vo et al., 2026)	✗	✓	✓	✓	✓	✗	✗	✗	✓	✗
LTDriveQA (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2024; Wang et al., 2025a). In this context, driving visual question answering (VQA) provides a way to probe whether a model can connect visual observations to language-grounded perception, prediction, and decision making (Marcu et al., 2024; Qian et al., 2024). Existing benchmarks have progressively extended this evaluation from scene-level and graph-structured reasoning to video-based temporal understanding, 3D-aware reasoning, multi-view safety-critical reasoning, and explicit spatiotemporal scene understanding (Zeng et al., 2025; Yu et al., 2025; Vo et al., 2026). However, recent studies show that VLMs can produce plausible driving responses even with corrupted or missing visual inputs, suggesting reliance on language priors rather than faithful visual grounding (Xie et al., 2025). This concern is especially relevant to long-tail driving events (Xu et al., 2025), where critical evidence may span time and camera views, posing a challenge for benchmark design: ensuring that questions genuinely depend on event-specific spatiotemporal evidence.

Current driving benchmarks fall short of this goal in two important respects. First, as summarized in Tab. 1, existing benchmarks cover long-tail, temporal, and multi-view reasoning to varying degrees: LingoQA and DVBench emphasize video-based temporal understanding, while WaymoQA and DriveSpatial additionally target multi-view and cross-view reasoning. However, none jointly centers on long-tail events, probes temporal evolution and cross-view evidence through scene-specific questions, and audits evidence dependence during construction. Second, many benchmarks rely on predefined question structures or capability-driven construction, which supports scalable coverage but can limit scene-specific question discovery (Qian et al., 2024; Sima et al., 2024). Even benchmarks with free-form questions do not explicitly verify whether each question truly depends on the intended temporal or cross-view evidence (Marcu et al., 2024). As a result, seemingly complex questions may still be answerable from a salient frame, a single camera, or language priors (Liu et al., 2024b). A more diagnostic benchmark should therefore be event-centered, construct scene-specific questions that probe temporal and cross-view reasoning, and explicitly assess their dependence on the intended evidence during both construction and evaluation.

We introduce LTDriveQA, an event-centered benchmark of 2,889 questions over 719 clips selected from 192 long-tail Waymo E2E segments, with eight synchronized camera views and four driving competencies: Perception, Prediction, Planning, and Reasoning. The benchmark supports both multiple-choice questions (MCQ) and open-ended questions (OpenVQA). Building scene-specific questions with temporal and cross-view dependencies at scale is difficult through manual, one-shot authoring, which is labor-intensive and costly. We therefore develop an automatic and scalable pipeline that recursively leverages VLMs to localize long-tail events and actors, generate competency-aware questions, diagnose their quality and evidence dependence, and refine subsequent generations through feedback. Grounded templates further provide controlled coverage of motion patterns and cross-camera transitions. Human reviewers inspect generated questions and verify answer annotations, while uncertain or low-quality cases are revised or discarded. Answers are grounded in visual, ego-state, or calibrated camera evidence, while text-only, single-frame, and front-only probes expose trivial or shortcut-prone questions for revision; future evidence is withheld for Prediction. The resulting benchmark links each question to its event, reference answer, and diagnostic evidence, enabling targeted evaluation of temporal and cross-view reasoning.

Evaluating 25 recent VLMs reveals substantial headroom on LTDriveQA. On a matched 100-MCQ subset, humans achieve 91.0% accuracy versus 65.0% for the best model. The proprietary GPT-5.6-Sol achieves the highest average score of 60.0, while Qwen3-VL-8B and the driving-specialized

Qwen-Drive-1.0 reach 40.5 and 45.6, respectively. Input ablations further show that all open-source models achieve higher aggregate MCQ accuracy with visual input, and that temporal and surround-view gains are strongest on Perception and multi-view questions, respectively, whereas on Prediction the final frame alone outperforms the full video on average, revealing weak use of video history. Together, these results show that LTDriveQA remains challenging while enabling targeted analysis of how temporal and cross-view evidence affects driving reasoning across model families, competencies, and input conditions. Our main contributions are summarized as follows:

- We propose LTDriveQA, an event-centered driving VQA benchmark that jointly evaluates long-tail, temporal, and multi-view scenarios while separating driving competencies, evidence requirements, and answer formats, comprising 2,889 human-reviewed questions over 719 event clips.
- We introduce an automatic and scalable recursive construction pipeline that generates scene-specific questions, audits their quality and evidence dependence, and iteratively refines the generation guidance from each round's keep/discard feedback before human review.
- A comprehensive evaluation of 25 recent general-purpose and driving-specialized VLMs under textual, temporal, and view-controlled settings, revealing substantial headroom on LTDriveQA for both groups: proprietary models retain a pronounced advantage on free-form questions, while Planning and Prediction remain particularly challenging across model scales.

2 Related Work

Vision-Language Models for Autonomous Driving. VLMs increasingly connect driving scene understanding to explanations and decisions (Xu et al., 2024; Nie et al., 2024), evolving from decision making over structured driving context (Fu et al., 2024; Wen et al., 2024) to multimodal reasoning for perception, prediction, planning, and control (Tian et al., 2024; Ma et al., 2024). Work on richer visual representations (Ding et al., 2025; Han et al., 2025; Wang et al., 2025a), language-guided action generation (Hwang et al., 2025; Zhou et al., 2025b;a), and generative driving models (Hu et al., 2023; Wang et al., 2024b; 2023b) further expands their inputs and outputs. However, plausible explanations or actions do not necessarily establish correct scene understanding or actor-level reasoning (Li et al., 2024; Xie et al., 2024; Kong et al., 2023b). LTDriveQA addresses this gap through questions about observed motion, future outcomes, ego decisions, and visible explanations, paired with controls that restrict the available evidence.

Driving VQA and Spatiotemporal Benchmarks. Driving VQA benchmarks cover scene description and spatial relations (Kim et al., 2018; Deruyttere et al., 2019; Qian et al., 2024; Cao et al., 2024; Guo et al., 2025; Inoue et al., 2023), perception-to-planning reasoning (Sima et al., 2024; Ding et al., 2024), and temporal, spatial, or safety-critical understanding (Marcu et al., 2024; Wang et al., 2024a; Zeng et al., 2025; Kim et al., 2025; Liu et al., 2023; Shiri et al., 2024; Cheng et al., 2024; Chen et al., 2024; Yang et al., 2025; Lin et al., 2025). Recent driving benchmarks like WaymoQA and DriveSpatial further target multi-view and spatiotemporal reasoning, often leveraging structured scene representations (Yu et al., 2025; Vo et al., 2026; Ishihara et al., 2025; Tian et al., 2020). LTDriveQA instead emphasizes event-centered question construction and evidence auditing, combining question generation with explicit view requirements and restricted-input probes for diagnosing shortcuts.

Automated Benchmark Construction and Evidence Auditing. Automated benchmark construction commonly relies on scene-graph templates (Grunde-McLaughlin et al., 2021; Wang et al., 2023a), human annotation (Xiao et al., 2021; Yu et al., 2025), and model-assisted generation or refinement (Chen et al., 2025; Zhang et al., 2024; Luo et al., 2025). Complementary studies probe shortcut reliance using temporal contrast, dispersed evidence, non-visual controls, and robustness tests (Liu et al., 2024b; Ben-Ami et al., 2026; Xie et al., 2025; Kong et al., 2023a). Our pipeline integrates question construction and evidence auditing in driving scenes, using generation feedback on relevance, answerability, and evidence sufficiency. Restricted-input probes then test whether finalized questions depend on the intended event evidence.

3 LTDriveQA Benchmark

Benchmark Overview. LTDriveQA is an event-centric benchmark designed to evaluate spatiotemporal understanding in long-tail, multi-view driving videos rather than generic video segments, with questions spanning Perception, Prediction, Reasoning, and Planning. LTDriveQA grounds each

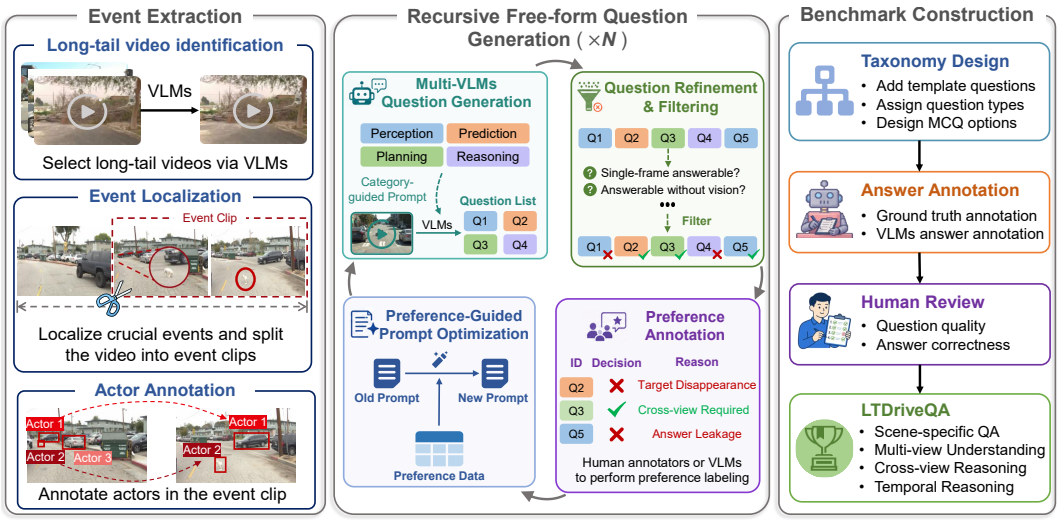


Figure 2: **Construction pipeline of LTDriveQA.** Left: localize events in driving segments and identify their actors. Middle: generate competency-specific questions, check their evidence requirements, and update generation guidance from keep/discard feedback over N rounds. Right: add eligible templates, assign competencies and answer formats, annotate answers, and review uncertain cases. The framework supports multiple question authors; the archived rounds use one author model and different models for judging and annotation.

question in a temporally localized driving event and its relevant actors, while recording its intended view requirement; restricted-input probes then assess temporal evidence use on applicable questions. This design enables systematic evaluation of whether a model can reason over event evolution, cross-view observations, and future outcomes, while reducing reliance on language, single-frame, or single-view shortcuts. As illustrated in Fig. 2, we construct LTDriveQA through three stages: long-tail event extraction, recursive scene-specific question generation, and question finalization with taxonomy design and quality assurance, as detailed in the following subsections.

3.1 Long-Tail Event Extraction

To ground questions in specific driving interactions, we use VLMs to screen Waymo E2E sequences for candidate long-tail scenarios and localize them into event-centered clips. Each clip preserves the event's temporal evolution and surrounding context, with actor records identifying the subjects of subsequent questions. For cross-view events, we verify actor continuity across cameras and derive spatial attributes from calibrated camera orientations. Accepted trajectories must follow chronological order and a geometry-consistent path around the calibrated camera ring; otherwise, no transit-derived spatial label is assigned. Ego-state records provide additional grounding for applicable motion labels. Here, long-tail denotes rare, safety-relevant events selected from a source corpus already curated for such behavior, not an estimate of their natural frequency. Appendix A.2 provides the extraction details, including the observation boundaries used for Prediction.

3.2 Recursive Scene-Specific Question Generation

Question Generation. Given an event clip, a VLM generates questions using a prompt defining the target competency, allowed video frames, question format, and generation guidance. For Prediction, the model sees only frames preceding the future outcome. Each question asks about one property of an identifiable subject, without answers, options, or answer-revealing hints. Previously generated questions from the same clip are included to avoid repetition. The model may skip generation when the evidence is insufficient for the target competency.

Question Filtering. We check whether each question is relevant to driving and answerable from the video, and flag duplicates and possible text-only or single-frame shortcuts. Prediction requires cues in the observed frames and a future outcome verifiable in withheld frames; Planning requires a scene-dependent action, and Reasoning requires a visually supported explanation rather than speculation. The checks record issues and reasons to keep or discard questions before answer annotation.

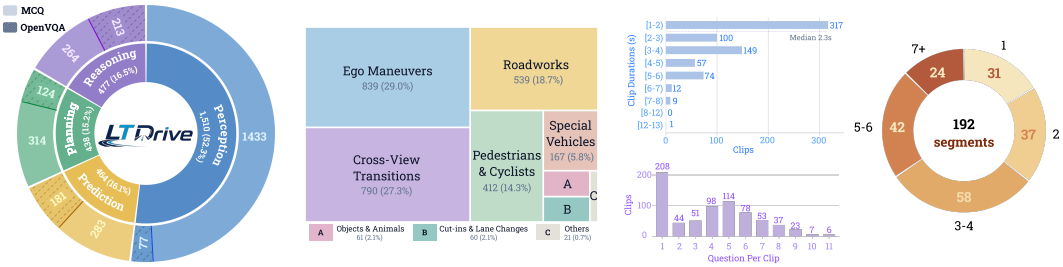


Figure 3: **Composition and statistics of LTDriveQA.** (a) Questions by competency and answer format, with competency totals and proportions shown in the inner ring. (b) Questions by event type. (c) Clip-duration distribution. (d) Questions per clip. (e) Source segments grouped by the number of clips contributed.

Recursive Guidance Refinement. After each round, we revise the generation guidance using this feedback. Let $g_c^{(r)}$ denote the guidance for competency c at round r, and $D_c^{(r)}$ the round's candidates with their keep/discard decisions and reasons. We update:

$$g_c^{(r+1)} = \mathcal{S}\left(g_c^{(r)}, D_c^{(r)}\right), \tag{1}$$

where $\mathcal{S}$ is an LLM that reads the pass rate and the retained and rejected examples of each competency, and rewrites $g_c^{(r)}$ to keep successful patterns and forbid recurring errors. Competency definitions, allowed frames, and question format stay fixed, and an invalid update keeps $g_c^{(r)}$. The process repeats within a fixed budget, retaining each round's guidance and decisions for later analysis (Appendix A.3).

3.3 Question Finalization and Quality Assurance

Coverage and Semantic Typing. Grounded templates complement free-form generation with controlled coverage of ego behavior, actor motion, and cross-view transitions when the required event records are available. After deduplication, each question is independently assigned a semantic competency, view requirement, and answer format. Perception queries an observed state or change; Prediction queries a future outcome; Planning requests a scene-dependent ego action or risk priority; and Reasoning requests a visible explanation or counterfactual consequence. Questions use MCQ when a unique set of plausible alternatives can be constructed and OpenVQA otherwise.

Grounded Answer Annotation. Reference answers are derived from visual evidence or, when applicable, ego-state records and calibrated cross-view trajectories. For MCQs, distractors query the same subject and property as the correct answer and are checked for plausibility, uniqueness, and wording symmetry. Items without a reliable single-answer option set are revised, converted to OpenVQA, or routed for review.

Evidence Auditing and Human Review. Finalized MCQs are re-answered under text-only, single-timestamp, and front-only inputs to identify possible language, temporal, or view shortcuts. These probes serve as diagnostic flags rather than automatic rejection criteria because their outcomes also depend on probe-model ability. Flagged or unresolved items enter a bounded rewriting and revalidation loop while retaining their construction lineage. Finally, human reviewers inspect every released question for clarity, answer correctness, and evidence sufficiency, revising or rejecting invalid items. Details are provided in Appendices A.5 to A.7.

3.4 Benchmark Composition and Coverage

Benchmark Composition. LTDriveQA contains 2,889 reviewed question–answer pairs derived from 719 event clips across 192 long-tail source segments, identified by screening all 1,505 test segments of the Waymo End-to-End dataset (Xu et al., 2025). As summarized in Fig. 3, the benchmark covers four competencies: Perception, Prediction, Planning, and Reasoning, and includes 2,294 MCQs and 595 OpenVQA questions across 15 annotated event types. Clip durations range from 1.0 to 12.0 seconds with a median of 2.3 seconds, and question density varies across clips and source segments. Additional statistics are provided in Appendix A.9.

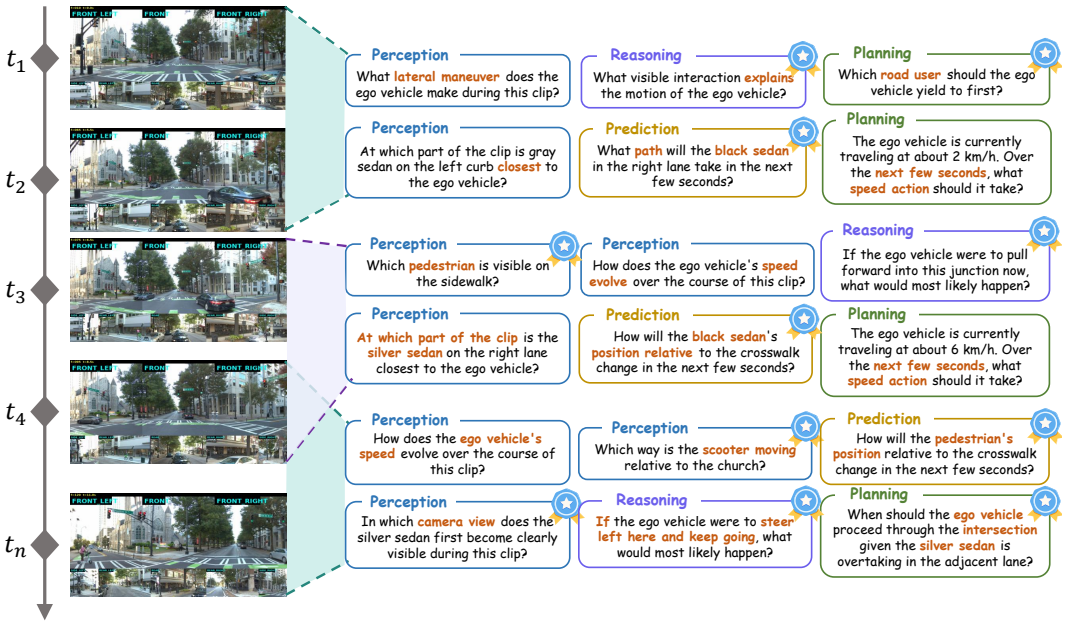


Figure 4: **Representative questions in LTDriveQA.** LTDriveQA features diverse, event-specific questions grounded in the actors, interactions, and temporal evolution of each scene. Medal-marked examples indicate VLM-generated free-form questions, highlighting scene-adaptive question design beyond generic templates.

Human Validation and Agreement. All 2,889 scored question–answer pairs undergo human review by five reviewers, who verify question clarity, answer correctness, and the sufficiency of visual and temporal evidence, revising ambiguous or incorrect instances when necessary. This verification and refinement process requires approximately 55 person-hours. We additionally conduct a human agreement study to validate the OpenVQA judging protocol, requiring approximately 12 additional person-hours. On 150 sampled model responses, the automated judge agrees with blinded human annotators on 88.7% of binary correctness decisions, with Cohen's $\kappa = 0.77$ (95% CI: $[0.66, 0.86]$). Detailed review procedures and agreement analyses are provided in Appendices A.7 and B.2.

4 Benchmark Evaluation

4.1 Experimental Setup

Construction Implementation. Qwen2.5-VL-32B is used for long-tail video identification, Qwen3-VL-30B for event localization and actor annotation, and Qwen3-VL-30B/235B for question generation, with Claude Fable 5 as the guidance updater $\mathcal{S}$. InternVL3.5-30B and Qwen3-VL-32B perform refinement and filtering, while Gemma-3-27B and Claude Opus 4.8 support taxonomy design and answer annotation (details in Appendix A.3).

Evaluated Models. We comprehensively evaluate 25 VLMs in three groups: The proprietary group comprises Claude Opus 4.8 (Anthropic, 2026), GPT-5.6-Sol (OpenAI, 2026), and Gemini 3.5 Flash (Google, 2026). General-purpose open-source models comprise Qwen3-VL (2B–32B) (Bai et al., 2025), InternVL3.5 (2B–38B) (Wang et al., 2025b), Gemma 3 (4B–27B) (Team et al., 2025), GLM-4.6V-Flash (GLM-V Team et al., 2025), and LLaVA-1.6 (Liu et al., 2024a). The driving and embodied group comprises Cosmos-Reason2 (2B–32B) (Agarwal et al., 2026), ReCogDrive (2B/8B) (Li et al., 2026), MiMo-Embodied-7B (Hao et al., 2025), and Qwen-Drive-1.0 (Zhou et al., 2026). We use public checkpoints without adaptation to LTDriveQA. Appendix B.1 records endpoints, serving settings, and model-specific constraints. Full results for all 25 models appear in Tab. 9.

Evaluation Protocol. Following Xu et al. (2025), we combine the eight labeled camera views into a fixed 1536×820 resolution and sample the resulting multi-view images at 5 Hz over the observation window to construct the input sequence. All models are evaluated zero-shot with greedy decoding, while Claude Opus 4.8 and GPT-5.6-Sol use their default API settings. For MCQ, we report accuracy

Table 2: **Evaluation results on LTDriveQA.** Each competency reports MCQ, OpenVQA, and their mean (Avg.); Average spans all eight cells, or the four MCQ cells for Human. **Green** marks the best per group.

Model	Size	Perception			Prediction			Planning			Reasoning			Average
		MCQ	Open	Avg.	MCQ	Open	Avg.	MCQ	Open	Avg.	MCQ	Open	Avg.	
Human	–	88.0	–	–	85.3	–	–	93.3	–	–	97.3	–	–	91.0
Proprietary VLMs														
Gemini 3.5 Flash	–	44.0	48.1	46.0	47.3	47.5	47.4	39.8	52.4	46.1	68.9	49.8	59.4	49.7
Claude Opus 4.8	–	61.0	48.1	54.5	46.3	63.0	54.6	43.3	**65.3**	**54.3**	67.8	**64.8**	**66.3**	57.4
GPT-5.6-Sol	–	**64.2**	**58.4**	**61.3**	**56.2**	**67.4**	**61.8**	**47.1**	58.9	53.0	**71.2**	56.8	64.0	**60.0**
Open-Source VLMs														
LLaVA-1.6	7B	27.6	26.0	26.8	36.4	14.9	25.6	36.9	25.0	30.9	41.7	20.7	31.2	28.6
GLM-4.6V-Flash	9B	50.5	41.6	46.0	44.2	24.9	34.5	43.9	42.7	43.3	56.4	34.7	45.5	42.4
Gemma 3	4B	22.4	18.2	20.3	30.4	10.5	20.4	35.4	16.9	26.1	41.3	12.7	27.0	23.5
Gemma 3	12B	33.6	24.7	29.1	32.5	12.2	22.4	40.8	32.3	36.5	47.7	19.2	33.5	30.4
Gemma 3	27B	38.9	23.4	31.1	33.6	21.0	27.3	34.7	29.0	31.9	51.9	16.0	34.0	31.1
InternVL3.5	2B	36.8	24.7	30.8	30.0	17.7	23.9	**45.2**	33.1	39.2	43.6	18.3	31.0	31.2
InternVL3.5	4B	44.3	33.8	39.0	31.4	15.5	23.4	35.7	33.9	34.8	50.0	26.8	38.4	33.9
InternVL3.5	8B	52.1	32.5	42.3	33.2	27.6	30.4	36.6	41.1	38.9	53.4	28.6	41.0	38.2
InternVL3.5	38B	51.1	24.7	37.9	37.8	26.0	31.9	38.5	**46.0**	42.2	62.1	33.3	47.7	39.9
Qwen3-VL	2B	40.5	18.2	29.4	31.1	16.6	23.9	38.9	28.2	33.5	48.5	21.6	35.0	30.4
Qwen3-VL	4B	49.9	28.6	39.2	36.0	23.8	29.9	40.4	37.1	38.8	56.1	35.7	45.9	38.4
Qwen3-VL	8B	50.5	37.7	44.1	35.7	26.0	30.9	42.7	37.1	39.9	58.0	36.6	47.3	40.5
Qwen3-VL	32B	**55.1**	**48.1**	**51.6**	**50.9**	**42.0**	**46.5**	43.6	44.4	**44.0**	**67.4**	**44.1**	**55.8**	**49.4**
Open-Source Driving and Embodied Models														
Qwen-Drive-1.0	4B	51.4	**40.3**	**45.8**	39.9	**40.9**	**40.4**	37.3	**50.0**	43.6	59.8	**45.5**	**52.6**	**45.6**
MiMo-Embodied	7B	53.5	36.4	45.0	38.2	28.2	33.2	39.2	46.0	42.6	58.0	26.3	42.1	40.7
ReCogDrive	8B	49.0	20.8	34.9	38.9	28.7	33.8	41.1	43.5	42.3	49.6	29.6	39.6	37.6
Cosmos-Reason2	8B	**58.3**	22.1	40.2	39.6	23.8	31.7	42.7	32.3	37.5	61.0	26.3	43.6	38.2
Cosmos-Reason2	32B	54.4	31.2	42.8	**45.2**	24.9	35.0	**43.3**	46.0	**44.6**	**65.9**	35.7	50.8	43.3

based on the predicted option letter, with answer positions balanced across question groups. For OpenVQA, following prior open-end evaluation (Yu et al., 2023; Nie et al., 2024; Marcu et al., 2024; Xie et al., 2025), we use gpt-5.4-mini as an LLM judge to assess whether each response correctly captures the key claim of the reviewed reference, achieving substantial agreement with human evaluation (Cohen's kappa $\kappa = 0.77$). OpenVQA accuracy is defined as the resulting acceptance rate. Further details are provided in Appendix B.

4.2 Overall Benchmark Results

Tab. 2 shows that LTDriveQA remains challenging for current VLMs. GPT-5.6-Sol achieves the highest overall score of 60.0, followed by Claude Opus 4.8 at 57.4, while the strongest open-source model, Qwen3-VL-32B, reaches 49.4. In contrast, three human experts with driving experience achieve 91.0% accuracy on a balanced subset of 100 MCQs, where the best model reaches only 65.0%. This gap holds in every competency and is largest on Planning, where humans reach 93.3% and the best model 68.0%, indicating substantial room for improving scene-dependent understanding and decision making on the long-tail driving scenarios targeted by LTDriveQA.

The results also reveal complementary effects of model scale, domain specialization, and answer format. Scaling Qwen3-VL from 2B to 32B improves the overall score from 30.4 to 49.4, although gains are not uniform across competencies. Domain-specialized models show no consistent advantage, with Qwen-Drive-1.0 reaching 45.6 overall but remaining below Qwen3-VL-32B. Performance differences also vary across answer formats: the gap between GPT-5.6-Sol and Qwen3-VL-32B is 5.4 points on MCQ but 15.7 on OpenVQA, while GPT-5.6-Sol and Claude Opus 4.8 exchange rankings on Reasoning across the two formats. These results motivate evaluating both answer selection and free-form generation across distinct driving competencies.

4.3 Capability-Level Evaluation

Competency Differences. Model performance varies unevenly across driving competencies. Among general-purpose open-source models, Perception MCQ accuracy ranges from 22.4% to 55.1%, spanning 32.7 percentage points, whereas Planning scores occupy a narrower range of 34.7%–45.2%. Substantial differences in scene perception therefore coexist with more similar performance on scene-dependent action selection. Moreover, InternVL3.5-2B achieves the highest Planning MCQ accuracy



Figure 5: **Three representative failure modes of strong models.** The examples reveal persistent weaknesses in temporal grounding and motion reasoning: all eight strongest models fail in (a), seven of eight fail in (b), and none of the 25 evaluated models answers (c) correctly.

in this group at 45.2%, exceeding Qwen3-VL-32B at 43.6%, despite their markedly different overall scores of 31.2 and 49.4. These results highlight that *competency-level evaluation reveals model strengths that aggregate rankings can obscure*, allowing LTDriveQA to characterize the distribution of capabilities across driving tasks.

From Perception to Decision Making. The relationship between competencies further highlights the value of evaluating Planning separately. Across the 25 evaluated models, Perception MCQ accuracy has Spearman correlations of $\rho = 0.80$ with Reasoning and 0.67 with Prediction, compared with 0.51 with Planning. This weaker association is reflected in individual model comparisons: Cosmos-Reason2-32B substantially outperforms InternVL3.5-2B on Perception (54.4% vs. 36.8%), while InternVL3.5-2B achieves higher Planning accuracy (45.2% vs. 43.3%). These distinct rankings indicate that *Planning provides complementary information about model capability, supporting language-level assessment of driving decisions alongside perception*.

4.4 Qualitative Failure Analysis

Even the strongest models exhibit a shared failure in temporal localization. Across the eight top-performing proprietary and open-weight models, "near the beginning of the clip" is selected in 41.1% of predictions despite comprising only 12.7% of the ground truth. Controlled ablations on Qwen3-VL-8B further show that multiple observations matter far more than their order: on 133 questions asking where an actor *first* becomes visible, accuracy drops from 61.7% to 29.3% with only the final frame, but only to 57.1% when the same frames are randomly reordered. Predictions are thus often more consistent with where an actor becomes most prominent than where it first appears. Fig. 5(a) illustrates this behavior: an articulated bus appears first in `FRONT`, yet all eight models select the later, more prominent `FRONT_RIGHT` view. A related temporal failure appears in ego-speed estimation: in Fig. 5(b), the logged ego speed decreases monotonically from 11.5 to 3.3 m/s, yet seven of the eight models report a roughly constant speed.

Multi-view errors are similarly structured. Among incorrect camera predictions, 84.8% of proprietary-model errors and 84.7% of open-weight-model errors fall on an adjacent camera, while opposite-view confusions account for under 1% and left–right substitutions for only 3–4%. Models therefore generally recover the correct side but struggle to distinguish neighboring views, particularly near overlapping fields of view where the same actor may be visible in multiple cameras. This is not because surround-view input is ignored: on questions requiring multiple views, all eight cameras improve accuracy over `FRONT` alone by 8.2, 10.0, and 6.7 points for Qwen3-VL-8B, Qwen3-VL-32B, and InternVL3.5-8B, respectively. A related motion error appears in Fig. 5(c): a sedan remains stationary at a junction in the `SIDE_LEFT` view, yet every model describes it as approaching, receding, or turning—a pattern consistent with incomplete ego-motion compensation. Together, these results show that strong models benefit from additional temporal and multi-view evidence, yet use the relationships among those observations much less reliably.

Table 3: **Free-form question yield and quality across recursive generation rounds.** All rounds use the same 553 event clips, author model, and intent contract; round 0 receives no learned guidance. *Pass rate* denotes the fraction of new candidates passing all applicable content checks, while the last three columns report independent diagnostics (Appendix A.4). Retained counts reflect deduplication and later annotation decisions; historical candidates and template additions are excluded.

Round	New questions	Pass rate (%)	Retained questions	Evidence dependence (%)	Determinacy (%)	Grounding (%)
0	2,115	43.78	902	67.7	66.1	76.6
1	1,965	50.08	775	70.3	65.3	79.4
2	2,132	53.66	437	71.4	66.4	81.3

Table 4: **Restricted-input controls on MCQs, averaged over the 22 open-source models.** Question only includes the stem and options without images. Final frame only uses the last eight-camera composite of the observation window. Chance is the mean of $1/K$ over questions with K options.

Input	Perception (%)	Prediction (%)	Planning (%)	Reasoning (%)	Overall
Video (5 Hz)	45.6	36.7	39.3	53.4	44.5
Question only	28.1	31.0	34.0	37.2	30.3
Final frame only	41.0	38.8	38.1	51.8	41.6
Chance	27.0	27.4	21.0	23.8	25.9

5 Benchmark Analysis

Recursive Generation Effectiveness. As shown in Tab. 3, recursive feedback raises the content-check pass rate from 43.78% in round 0 to 53.66% in round 2. Because these checks also produce the feedback, we verify the gain with diagnostics that play no role in generation: evidence dependence rises from 67.7% to 71.4% and grounding from 76.6% to 81.3%, both significant under clip-level resampling, while determinacy stays at about 66%. Later rounds therefore ask more evidence-dependent and better-grounded questions without becoming more ambiguous. Because all rounds draw on the same 553 clips and competency contract, later rounds must also avoid a growing pool of earlier questions, so the fall in retained questions reflects deduplication rather than lower quality. These results characterize the recursive procedure as a whole rather than the guidance update alone. Its well-posed rate rises from 46.0% to 48.7%, but the clip-clustered interval spans zero (Appendix A.4).

Input Evidence Ablations. Prior work shows that VLMs can answer driving questions well even without valid visual input (Xie et al., 2025). We therefore test whether answers on LTDriveQA depend on visual evidence. As shown in Tab. 4, removing all images lowers the average MCQ accuracy of open-source models from 44.5% to 30.3%, with drops in every competency. Free-form answers show the same dependence: removing images lowers OpenVQA acceptance by 12.9 and 16.0 points for Qwen3-VL-8B and InternVL3.5-8B. Keeping only the final frame lowers accuracy to 41.6%, most on Perception and on questions missed by single frame probes (Tab. 16). On Prediction, however, the final frame alone outperforms the full video, suggesting that current models still exploit video input poorly: earlier frames can distract them from the latest state that determines the future outcome. Both effects hold under segment-level resampling. Residual priors also remain, especially for Planning, where accuracy without images stays close to the video result. Per-model results are in Appendix C.2.

6 Conclusion

We introduce LTDriveQA, an event-centered driving VQA benchmark for evaluating spatiotemporal intelligence in long-tail, multi-view scenarios. LTDriveQA covers Perception, Prediction, Planning, and Reasoning with both MCQ and OpenVQA, supported by explicit temporal and cross-view evidence labels. A scalable recursive pipeline generates and refines scene-specific questions through diagnostic feedback and human review, with modest gains in evidence dependence and grounding and stable determinacy. Evaluation across 25 VLMs reveals substantial headroom, while input ablations show broad reliance on visual input, stronger temporal gains in Perception, and consistent surround-view gains on multi-view questions across the tested models. Future work can use this scalable construction pipeline to extend LTDriveQA to broader video collections, more diverse driving events, and additional scene-specific questions.

Reproducibility Statement

Section 3 describes the benchmark formulation, event extraction, question generation, answer annotation, and human quality assurance, while Section 4 specifies the evaluated models, multi-view input protocol, decoding settings, and scoring procedures. The appendix further provides detailed construction and review procedures, benchmark statistics, model serving settings and input constraints, the MCQ prompt and parsing rules, the OpenVQA judging rubric and its human and cross-judge validation, and full per-model results and controlled input analyses. The source driving data and open-source model checkpoints are available under their respective access terms. Upon publication, we will release the benchmark annotations, which include questions, options, reference answers, source segment identifiers, and timestamps, together with the evaluation code.

AI Use Statement

We used LLMs to improve the clarity and readability of the manuscript and to assist in identifying relevant literature. All AI-assisted content and retrieved references were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Ethics Statement

LTDriveQA is built from the Waymo Open Dataset End-to-End test split and follows the Waymo Dataset License Agreement for Non-Commercial Use. We release only text annotations, namely questions, options, reference answers, and the segment identifiers and timestamps they refer to; no images or sensor data are redistributed, and users obtain the frames directly from Waymo after registering and accepting its license. The annotations are released under the same non-commercial terms with the attribution the license requires. Waymo removes or blurs personally identifiable information such as faces and license plates in the source data, and our annotations describe traffic behavior without identifying individuals. Finally, LTDriveQA evaluates language-level scene understanding and decision reasoning; its Planning answers are not validated driving actions and should not be used in safety-critical systems without action-level testing.

References

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Anthropic. Introducing claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, May 2026. Accessed: 2026-09-17.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. HERBench: A benchmark for multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ami_HERBench_A_Benchmark_for_Multi-Evidence_Integration_in_Video_Question_Answering_CVPR_2026_paper.html.

Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21819–21830, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7817–7826. IEEE, 2025.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2088–2098, 2019.

Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2024.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Enhancing mllms with multi-scale high-resolution details for autonomous driving. *International Journal of Computer Vision*, pp. 1–17, 2025.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, 2024.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Google. What's new in gemini 3.5 flash. `https://ai.google.dev/gemini-api/docs/whats-new-gemini-3.5`, 2026. Accessed: 2026-09-17.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. In *NeurIPS*, 2025.

Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, et al. MiMo-Embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*, 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *TMLR*, 2025.

Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. *arXiv preprint arXiv:2312.06352*, 2023.

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, and Yu Yamaguchi. Stride-qa: Visual question answering dataset for spatiotemporal reasoning in urban driving scenes. *arXiv preprint arXiv:2508.10427*, 2025.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–578, 2018.

Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 772–782. IEEE, 2025.

Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pp. 19994–20006, 2023a.

Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36:21298–21342, 2023b.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *ICLR*, 2026.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14864–14873, 2024.

Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

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

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

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. VideoAutoArena: An automated arena for evaluating large multimodal models in video analysis through user simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8461–8474, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Luo_VideoAutoArena_An_Automated_Arena_for_Evaluating_Large_Multimodal_Models_in_CVPR_2025_paper.html.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pp. 403–420. Springer, 2024.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pp. 252–269. Springer, 2024.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026. Accessed: 2026-09-17.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pp. 256–274, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. In *8th Annual Conference on Robot Learning*, 2024.

Yafu Tian, Alexander Carballo, Ruifeng Li, and Kazuya Takeda. Road scene graph: A semantic graph-based scene representation dataset for intelligent vehicles. 2020.

Hao Vo, Khoa Vo, Phu Loc Nguyen, Sieu Tran, Duc Minh Nguyen, Ngo Xuan Cuong, Gladys Gawugah, Sreevenkata Anjani Tishita Godavarthi, Chase Rainwater, Nghi DQ Bui, et al. Drivespatial: A benchmark for spatiotemporal intelligence in vlms for autonomous driving. *arXiv preprint arXiv:2605.23176*, 2026.

Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, and Mohammad Abdullah Al Faruque. Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. *arXiv preprint arXiv:2304.08600*, 2023a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22442–22452, 2025a.

Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5599–5606, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023b.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *International Conference on Learning Representations*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *ICCV*. IEEE, 2025.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Seungjun Yu, Seonho Lee, Namho Kim, Jaeyo Shin, Junsung Park, Wonjeong Ryu, Raehyuk Jung, and Hyunjung Shim. Waymoqa: A multi-view visual question answering dataset for safety-critical reasoning in autonomous driving. *arXiv preprint arXiv:2511.20022*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5972–5983, 2025.

Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15459–15469, 2024.

Xin Zhou, Zongchuang Zhao, Zhibo Yang, Mingsheng Li, Humen Zhong, Shuai Bai, Du Chu, Ruizhe Chen, Zhaohai Li, Jun Tang, Qiuyue Wang, Mingkun Yang, Jiazhao Zhang, Dayiheng Liu, Dingkang Liang, and Xiang Bai. Qwen-Drive-1.0: An initial step towards a vision-language foundation model for autonomous driving. *arXiv preprint arXiv:2609.00111*, 2026.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025a.

Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025b.



Figure 6: **Multi-view driving scenes in LTDriveQA.** Eight-camera observations capture surrounding actors and road conditions in varied daytime and nighttime settings. The benchmark asks event-centered questions about Perception, Prediction, Planning, and Reasoning.

A Additional Details of LTDriveQA

A.1 Task Representation and Evidence Contracts

Task Formulation and Benchmark Axes. Each example specifies the question, its permitted observations, and the answer to be evaluated. Let $V = \{I_t^v\}$ denote synchronized images indexed by timestamp t and camera v. An example contains an observation interval $[t_s, t_o]$, question q, competency c, view requirement ν, answer format f, and reference a. MCQs also contain options $O = (o_1, \ldots, o_K)$ and a correct option index; OpenVQA uses a textual reference. Prediction examples record an annotation endpoint $t_g > t_o$, with the target outcome established from $(t_o, t_g]$ and excluded from model input. Competency identifies the requested answer, while the view label records whether front-camera evidence is intended to suffice. Temporal-probe results are stored separately. These fields describe the intended task; evidence dependence must still be assessed empirically.

Benchmark Design Principles. Construction follows four criteria: questions must concern visible driving events, be relevant to driving or situational awareness, admit sufficiently determinate answers, and avoid revealing the requested evidence in their wording. Actor descriptions use only the attributes needed to identify the subject; a motion question, for example, should not identify an actor by the motion it asks the model to infer. Prediction requires a verifiable future target. Reasoning requires visible support for an explanation or counterfactual consequence, without speculation about private intentions. Automated checks operationalize these criteria, but model agreement or probe failure alone cannot guarantee that they are satisfied.

A.2 Details of Long-tail Event Extraction and Visual Inputs

Scenario Screening. A VLM inspects temporally ordered front-camera observations for unusual interactions and driving-relevant hazards, including cut-ins, vulnerable road users, obstructed lanes, and construction zones. Each proposal records scenario tags, a salient timestamp, and a rationale tied to visible evidence. Here, "long-tail" denotes targeted scenario selection within a source corpus already curated for rare behavior; the tags do not estimate natural event frequencies. Appendix A.9 quantifies the source curation and subsequent screening.

Temporal Localization and Actor Records. Event boundaries are recorded as frame indices in the 10 Hz source sequences. A coarse pass proposes an interval and its principal actor; denser

visual inspection then refines the boundaries or rejects unsupported proposals. Overlapping proposals concerning the same event and actor are merged. The standard clip-preparation configuration targets 1–5 s windows, while finalized clips span 1–12 s. Ego speed and longitudinal acceleration reconstructed from supplied ego-state records corroborate longitudinal motion labels, with visual observations determining event boundaries. Ordinary clips may include up to three identifiable secondary actors, with notes distinguishing their motion from apparent motion induced by the ego vehicle.

Cross-View Verification. Proposed sightings of the same actor are checked using temporal strips from individual cameras and calibrated camera geometry. The eight camera orientations define an azimuthal ring: accepted sightings must follow chronological order and a consistent direction around the ring, with at most two ring steps between consecutive sightings. Verified paths determine entry, exit, and passing-side attributes. The records retain camera identities and supporting timestamps as visibility annotations. Clips without a verified transit receive no transit-derived geometric labels. Construction stages use a shared renderer with labeled camera tiles.

Prediction Observation Boundary. For Prediction, the splitter first seeks a future interval after the event clip, targeting the larger of 1.5 s and 35% of the clip duration. If insufficient footage remains, an eligible clip is divided into a 65% observation prefix and a 35% annotation suffix. Items without a valid future holdout are excluded from Prediction generation. Question authoring and preference assessment use only the observation window; future-outcome verification and reference annotation may inspect the holdout. Evaluation inputs stop at the recorded observation endpoint.

A.3 Additional Implementation Details of the Construction Pipeline

Long-Tail Video Identification. Qwen2.5-VL-32B identifies long-tail videos among all 1,505 source segments using ten uniformly sampled front-camera frames plus four observations from the final four seconds, at temperature 0.2. The resulting long-tail videos provide the input to event localization in Fig. 2.

Event Localization and Actor Annotation. Qwen3-VL-30B-A3B-Instruct-FP8 localizes temporal events and annotates their actors. Front-camera event localization uses 12 coarse observations and ten observations for boundary refinement. Actor annotation identifies the principal actor and eligible secondary actors; cross-view actor continuity is checked using visual evidence and calibrated camera geometry, as detailed in Appendix A.2.

Multi-VLM Question Generation. The initial three recursive rounds use Qwen3-VL-30B-A3B-Instruct-FP8, with six temporal observations and temperature 0.5. For each eligible clip and competency, the model proposes two alternatives and retains at most one after structural checks. In these rounds, each candidate queries one property of an identifiable subject without an answer, options, or wording that reveals the answer; the model may abstain when evidence is insufficient. Prediction generation uses only the observation prefix. Later Prediction and Reasoning additions use Qwen3-VL-235B-A22B-Instruct-FP8 with eight temporal observations. Thus, multiple VLMs contribute to question generation across construction batches, while the initial recursive rounds use one author. Prior questions from the same clip serve as exclusions; token-set Jaccard matching at 0.85 and semantic subject–property matching remove duplicates. Candidate records retain their generation round, prompt version, and rejection history.

Question Refinement and Filtering. InternVL3.5-30B-A3B checks driving relevance, visual answerability, answer leakage, and single-frame or text-only shortcuts, using four temporal observations and temperature 0.1. Prediction checks require an informative prefix and a verifiable future outcome; Planning requires a scene-dependent action, and Reasoning requires visible explanatory evidence without hidden intentions. Filter labels and keep/discard rationales provide feedback for refinement. Qwen3-VL-30B-A3B also rewrites questions flagged during subsequent answer-quality checks. The later additions use Qwen3-VL-32B-Instruct for visual verification and text-only probes; Prediction additionally uses InternVL3.5-8B for text-only probing. These automated judgments remain subject to human review.

Tab. 5 provides the operational checklist for the initial recursive rounds, separating shared checks from the evidence required by each competency. Perception uses the general relevance and visual-

Table 5: **Question-filtering checklist by competency.** The shared checks apply across competencies, with the stated Prediction exception. Each competency row specifies the evidence required for a favorable content judgment before answer annotation.

Scope	Checklist
Shared	The answer matters for ego driving or safety awareness. The wording and generic driving knowledge alone do not determine the answer. Check for token-level duplicates and semantic overlap with questions about the same subject and property. Check whether a single middle front-camera frame suffices, subject to the Prediction exception described above.
Perception	The requested fact can be determined from the visible clip. Reject questions about incidental scenery or properties that are not visible or cannot be established from the observations. This category uses the general content checker rather than a separate competency-specific checker.
Prediction	The prefix supplies meaningful temporal evidence rather than only a generic prior or an arbitrary guess. The held-out future clearly establishes one objective answer. The requested future outcome is not already revealed in the prefix.
Planning	A specific visible actor, conflict, or traffic condition grounds the decision. The answer goes beyond universal advice such as "be cautious," "slow down," or "maintain safety." The evidence supports a defensible action or risk judgment.
Reasoning	Both the effect to be explained and its candidate cause are visually supported. One explanation is substantially better supported than other plausible explanations. Answering does not require guessing a driver's or pedestrian's private intention, motivation, or another non-visual state.

answerability checker; the other three competencies have dedicated checks. Prediction verification inspects both the observation prefix and the held-out future, although generation sees only the prefix. A positive single-frame flag contributes to filtering for Perception, Planning, and Reasoning, but is not itself a rejection reason for Prediction. The resulting model verdicts remain available for preference adjudication; they are distinct from the diagnostic probes applied after answer annotation.

Preference Annotation. Candidates are grouped by competency with their filter labels and keep/discard rationales. In the archived recursive rounds, preference annotations combine filter decisions with recorded adjudications. Although the framework supports a VLM preference panel, its outputs are empty in these runs; the annotations therefore do not represent a multi-model vote. Rejected candidates remain available as negative feedback for the next round.

Preference-Guided Prompt Optimization. Preference summaries contain per-competency statistics and retained/rejected examples, which guide revisions to the next round's generation prompt. Updates retain useful prior guidance and preserve the competency definitions, evidence boundaries, and output constraints, without updating model weights. Guidance contains four competency-specific sections within 700 words; an invalid update leaves the previous guidance unchanged. Claude Fable 5 performs these updates: it receives the current guidance and the preference summaries and returns the revised guidance.

Tab. 6 illustrates how feedback changes the guidance across three rounds. For example, Prediction feedback first encourages more varied target properties, then restricts speed-change questions after they become overused on distant or irrelevant vehicles. The table summarizes archived instructions and feedback; it does not imply that the identified failure modes are eliminated.

Taxonomy Design. Grounded templates add questions about eligible ego behavior, actor motion, and cross-view transitions. Gemma-3-27B-it assigns semantic competencies and answer formats and constructs initial MCQ options. For the later Reasoning OpenVQA-to-MCQ conversion, Qwen3-VL-30B-A3B and InternVL3.5-8B construct and check options using model-proposed distractors and template banks. Converted questions undergo text-only probing, visual option checks, and three answer-agreement queries using the conversion model. The Prediction additions instead use fixed outcome banks with alternatives describing the same subject and property.

Answer Annotation. Gemma-3-27B-it annotates initial MCQ answers using three samples and a strict-majority rule; available ego-state and calibrated camera-path labels take precedence. Claude

Table 6: **Evolution of question-generation guidance.** Rows alternate between the guidance used in each round and feedback informing the next update. Entries summarize the archived records; competency contracts remain fixed.

Stage	Perception	Prediction	Planning	Reasoning
Round 0: initial guidance	Prefer temporal changes, interaction ordering, relative motion, and cross-camera continuity. Avoid static trivia, counting, and answer-revealing premises.	Predict one verifiable property of a visible moving agent, supported by motion history and a clear held-out outcome. Avoid generic continuation and ambiguous targets.	Ask for scene-specific ego actions, risks, or priorities. Prefer plausible alternatives distinguishable through temporal evidence; avoid universal caution and descriptions of what ego actually did.	Explain an observed effect through visible, temporally linked evidence. Avoid hidden intentions, equally plausible causes, and questions that reveal their own explanation.
Feedback after round 0	Static adjacency and timestamp-specific lookups remained common. Some valid temporal questions involving small or dim subjects were incorrectly rejected.	Questions targeted stationary objects or ambiguously identified agents. Relative-position templates dominated, with additional camera references and repeated questions.	Generic "safest action" questions admitted universal answers such as slowing down or yielding. Some premises were not visually supported.	Questions explained motion through the same motion, relied on static infrastructure, confused parked and stopped vehicles, or assumed events that never occurred.
Round 1: updated guidance	Exclude static adjacency, timestamps, and ego telemetry. Retain small or dim subjects when their motion spans frames; emphasize temporal transitions and cross-camera ordering.	Require a moving, uniquely identifiable subject. Diversify properties across speed, turning, stopping, and crossing; exclude static objects, ego prediction, camera references, and overlapping-clip repeats.	Require a concrete visible conflict and distinguishable action alternatives. Favor decision timing and yielding priority; skip questions fully answered by generic caution.	Require a visible, temporally linked trigger distinct from the effect being explained. Exclude static causes and verify that the claimed effect actually occurs.
Feedback after round 1	First/last/after-pass variants repeatedly queried the same transit event, escaping lexical deduplication.	Diversification overcorrected toward speed-change templates for distant or irrelevant vehicles. Some targets left view before the observation prefix ended.	Priority questions sometimes had only one candidate. Vague scene references, parked candidates, and unsupported ego intentions persisted.	Relative-motion tautologies persisted, alongside camera-visibility explanations, static causes, hidden motivations, and unsupported event premises.
Round 2: updated guidance	Allow only one question per transit event; prohibit first/last/after-pass paraphrases. Continue excluding single-frame identification and poorly visible signal states.	Select the property whose uncertainty is supported by the prefix. Restrict speed prediction to nearby, relevant agents interacting with ego's path; require visibility at the prefix endpoint.	Require at least two moving candidates for yielding priority. Name specific actors and states; exclude invented ego intentions and questions with a single obvious answer.	Focus on actual behavior changes explained by a different visible dynamic trigger. Prohibit position/visibility tautologies, static causes, private motivations, and camera-based wording.

Opus 4.8 annotates OpenVQA references through Azure AI Foundry using six observations, default API sampling, and a 2,048-token output budget. Later additions derive answers from structured scene records or held-out outcomes using Qwen3-VL-235B, with ego-state records supplying applicable Prediction labels. Visual future-answer annotation uses six held-out observations and three option-permuted confirmation queries to Qwen3-VL-32B. Held-out observations are excluded from question generation and benchmark evaluation.

Human Review. Human reviewers inspect every released question, including low-confidence items flagged by the automated stages, and revise or reject questions, options, and references based on clarity, answer correctness, and evidence sufficiency. Appendix A.7 describes this final review. All settings above concern construction; benchmark evaluation follows the separate protocol in Appendix B.

A.4 Gold-Free Quality Probe across Generation Rounds

The Rounds Are Already Matched. The three archived rounds form a controlled comparison without any re-generation. All three ran over the *same* 553 event clips with one author model and one fixed intent contract, producing 2,115, 1,965 and 2,132 new free-form candidates; the only thing that differs between them is the guidance learned from the previous round's keep and discard decisions. Round 0 uses initial seed guidance without updates from prior-round feedback and is therefore the no-recursion arm of the comparison. What the round-over-round pass rates in Tab. 3 lack is not a matched design but an independent criterion, since those checks also supply the feedback. The probe below supplies one.

Why the Metrics Are Gold-Free. A round's archived candidates carry question text and filter labels only: by design the author returns a question without an answer, options, or hints, and answers are annotated later, for retained items alone. Scoring a round by accuracy is therefore impossible without re-annotating it, and would be uninformative even then. Accuracy moves in the same direction for a question that is harder and for a question that is ambiguous, so a falling accuracy would not distinguish a better benchmark from a noisier one. We instead measure two properties that separate those cases, using no gold answer and no human rating.

Probe and Judge. A probe model, Qwen3-VL-8B-Instruct, answers each candidate question twice: once from the question alone, and once with the 5 Hz eight-camera clip, both at temperature 0 with a 256-token budget. The clip window is exactly the evidence the question's author saw, for Prediction intents as well, so the probe is never shown frames the author did not have. A separate local text model, Qwen3-8B, then decides whether two answers assert the same central claim, applying the key-claim criterion of the OpenVQA rubric in Appendix B.2: wording, length and extra detail are ignored, while a reversed direction, side, motion state or subject actor counts as a different claim. Both models are served locally; the probe never sees a reference answer, and the judge never sees the video and never decides correctness.

The Two Measured Properties. *Evidence dependence* is the fraction of candidates whose blind and video answers differ: a question that survives this test cannot be answered from a generic driving prior. *Determinacy* is the mean pairwise agreement among four video answers sampled at temperature 1, which asks whether the question has a settled answer once the evidence is visible. The two together separate the cases that accuracy conflates: a well-posed question is one the probe cannot guess blind but does converge on when it can see, whereas an ambiguous question fails the second test as well as the first. We report their conjunction as the well-posed rate. Determinacy needs four video answers per question, four times the inference of every other column, so it is measured on a per-round stratified sample of about 600 candidates drawn with a fixed seed and balanced across generation intents; the other columns use every candidate the round produced. Two further columns need no model at all: grounding is the share of questions naming the clip record's actor or event, and the near-duplicate rate is the share sharing at least 80% of their content words with an earlier candidate in the same round, measured before deduplication.

Results and Residuals. The probe supports a modest, genuine effect and rules out a large one. Evidence dependence is 67.7% in round 0, 70.3% in round 1 and 71.4% in round 2; against round 0 those are gains of 2.60 and 3.69 points, with clip-clustered 95% intervals of $[0.20, 5.01]$ and $[1.23, 6.15]$

and bootstrap p values of 0.03 and 0.002. Grounding moves the same way, from 76.6% to 79.4% and 81.3% ($[1.27, 4.25]$ and $[3.10, 6.25]$, both $p < 0.001$). Determinacy does not move at all: 66.1%, 65.3% and 66.4%, so the later rounds do not buy their harder questions by making them vaguer. The well-posed rate, the conjunction of the two, rises from 46.0% to 48.9% and 48.7%, but it is measured on about 600 candidates per round and its intervals span zero ($[-2.61, 8.30]$ and $[-2.69, 8.05]$), so we report it without claiming a difference. Intervals are clip-clustered because a clip contributes several questions to every round; the rounds ask different questions, so these are unpaired comparisons over a shared clip pool rather than paired tests.

Two residuals are worth stating plainly. First, the gain accumulates unevenly: evidence dependence and grounding rise over both recursive rounds, but most of the improvement arrives in the first recursive step ($+2.60$ and $+2.8$ points), with a smaller further gain in round 2 ($+1.09$ and $+1.9$ points). Second, the independent gain is far smaller than the internal one. The content-check pass rate rises 9.88 points across the same rounds, while evidence dependence rises 3.69; a little over a third of the apparent improvement survives a criterion that did not also generate the feedback. This is the quantity the pass-rate column cannot establish on its own, and it is the reason we report both. Near-duplicate rates do not improve either, at 21.2%, 28.6% and 20.7% before deduplication, with round 1 the worst of the three; the falling retained counts in Tab. 3 reflect that redundancy against a growing exclusion pool rather than falling candidate quality. These retained counts are construction-stage counts; the released free-form set (908 questions; Fig. 8) additionally reflects final human review (Appendix A.7) and the later targeted additions described in Appendix A.3.

Scope. This probe measures evidence dependence and answer determinacy, which is what the benchmark's design principles ask of a question; it is not a measure of difficulty, of factual correctness, or of whether the eventual reference answer is right. One probe model and one agreement judge are used throughout, so the absolute rates would shift with other models even though the round ordering is the quantity of interest. Rounds 1 and 2 also receive earlier questions as exclusions, so their candidates are not independent draws from the same distribution as round 0's; the comparison estimates the effect of the recursive procedure as run, not of the guidance text in isolation.

A.5 Taxonomy, Answer Formats, and Options

Semantic Competency Assignment. Competency is determined by the requested answer. Perception concerns states and changes within the observation interval, including cross-camera tracking. Prediction concerns an outcome after the observation endpoint. Planning requests an ego action or risk priority; an action request remains Planning when it also asks for a rationale. Reasoning requests a visible explanation or a scene-grounded counterfactual consequence. Competency, view requirement, and answer format are assigned separately. For free-form questions, a semantic classifier applies these definitions with wording checks to avoid unsupported Prediction, Planning, or Reasoning labels.

Template Eligibility and Answer Formats. Templates require the records needed to answer them: ego-state records for ego behavior, verified camera-path attributes for transit, and an identified primary or secondary actor for actor questions. Each template specifies its competency, view requirement, and format. Format assignment selects MCQ when a reliable option set with one supported answer can be constructed and OpenVQA otherwise. The selected free-form configuration enables these two formats, although the implementation also supports legacy binary questions. Captioning remains auxiliary and is excluded from the reported benchmark.

Answer Sources and Option Verification. Applicable ego-state and calibrated-path labels take precedence over VLM selections. Structured properties, such as speed trend, turn direction, and camera entry or exit, use canonical option sets. Other options are proposed from the annotation window and checked visually for multiple valid answers, no valid answer, or insufficient evidence. VLM-based MCQ annotation uses three samples and requires a strict majority; otherwise, the label remains unresolved. OpenVQA uses a separate visual reference annotator, configured as Claude Opus 4.8 in the archive, with one reference sample. Prediction reference annotation may inspect the held-out future, as specified in Appendix A.2. All resulting labels still require quality checks.

Option Symmetry and Controlled Alternatives. Option checks target wording cues and ambiguous alternatives. Deterministic checks reject identical options, near-duplicates with token overlap of at

least 0.8, and inconsistent naming of the subject. After annotation, they flag correct answers that differ from the median distractor length by more than six words or substantially repeat the question. When an open reference exists, they also flag a distractor with at least as much reference overlap as the designated answer. These lexical heuristics complement visual assessment. Prediction options describe alternative outcomes of one property, such as longitudinal action or crossing completion, and must be mutually exclusive over the specified horizon: "slows but keeps moving" and "comes to a stop" cannot both be correct under the question's interpretation.

Answer-Position Control. MCQs are grouped by template, option count, and competency, then processed in a deterministic order to limit template-specific answer-position cues. The correct answer is assigned to the least occupied feasible position, and distractors fill the remaining positions. The implementation remaps the correct letter and checks that answer content is unchanged; summary alternatives that depend on preceding options remain last. Grouping by option count prevents assignments to nonexistent positions. This procedure controls position frequencies but does not remove semantic answer priors.

A.6 Evidence Probes and Diagnostic Rewriting

Question-Level and Answer-Level Checks. Before annotation, judges assess whether a question can be answered from text alone or a middle front-camera frame, without a reference answer. After annotation, the standard MCQ path repeats text-only answering four times and flags an item with at least three options if every response matches the reference. Single-timestamp probes inspect the beginning, middle, and end of the visible interval, preserving the clip's camera layout and recording answer matches. Front-only video probes test recovery without other cameras. OpenVQA receives question-level checks and visual reference annotation, but not this answer-aware MCQ audit.

Probe results are diagnostic flags interpreted against the question's stated evidence requirements, not universal rejection rules. A reference match may flag a shortcut, whereas a mismatch does not establish evidence dependence because it also depends on probe-model ability. These probes do not constitute a factorial temporal/view experiment.

The archived front-only construction probe uses the full clip for clip-split Prediction items and may therefore include held-out future frames. Its inputs must be corrected to the observation endpoint before its flags can support a view-dependence claim. Temporal Prediction probes already use that endpoint. This limitation concerns construction-time flags and is separate from the permitted-input protocol for benchmark evaluation.

Rewriting and Unresolved Records. Diagnostic rewriting revises individual items after answer-quality checks. This loop is separate from recursive free-form generation: it supplies failure reasons to a rewriter, revalidates the revision, and repeats applicable checks within a bounded retry budget. Records retain their lineage and unresolved status. Missing answers are not replaced by guessed labels, and invalid MCQ option sets may be routed to OpenVQA annotation. Human review procedures and outcomes are described in Appendix A.7.

A.7 Human Review Protocol and Outcomes

Protocol. Every released question was reviewed by at least two reviewers. Reviewers used a browser interface shown in Fig. 7, which shows the eight-camera clip at 5 Hz, the question, its options, and the draft reference. They could *confirm* the item, *modify* its question, options, or reference, or *reject* it, and could also add hand-written questions. Reviewers worked blind to each other's decisions. A fixed reviewer precedence, set before review, was used to resolve keep/discard conflicts.

Coverage. Five reviewers inspected all 3,556 candidate items from the main construction run over 192 segments, most often revising options or reference answers.

Outcome. The final benchmark contains 2,889 scored questions after human review and refinement. Reviewers edited the question, options, or reference of about 44% of them, and each item keeps its edit record. This construction review is distinct from the judge–human agreement study in Appendix B.2, which evaluates the reliability of the OpenVQA scorer.

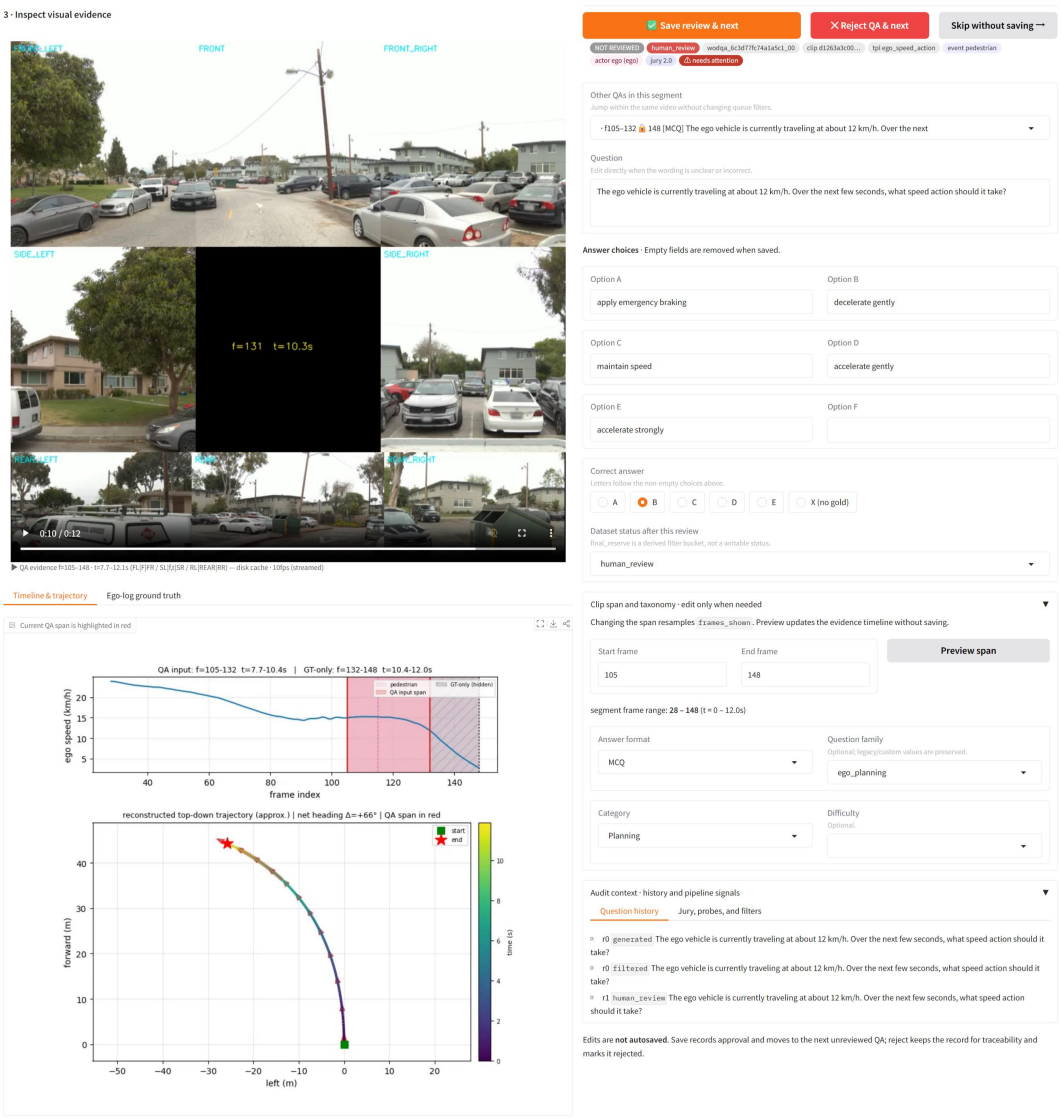


Figure 7: **Screenshot of LTDriveQA review interface.**

A.8 Criteria for the Benchmark Comparison

Tab. 1 marks a property only when the original paper documents it for the evaluation set. Long-tail requires source data selected for rare or safety-critical events, such as crashes, near-misses, or curated corner cases, rather than routine driving logs. Video requires multiple time-ordered frames per question, and Multi-view requires synchronized cameras facing different directions. Temporal and Cross-view require a dedicated group of questions whose answers need several timestamps or several camera views. Scene-specific requires questions written for the events and actors of each scene rather than instantiated from a shared template bank, and Free-form requires open question wording. Audit requires construction-time checks of each question's evidence dependence, such as text-only or single-frame probes that decide whether it is kept or revised; answer verification and evaluation-time ablations do not count. Closed requires answers from a fixed option set scored by exact match, and Open requires free-text answers scored by text metrics, judges, or humans.

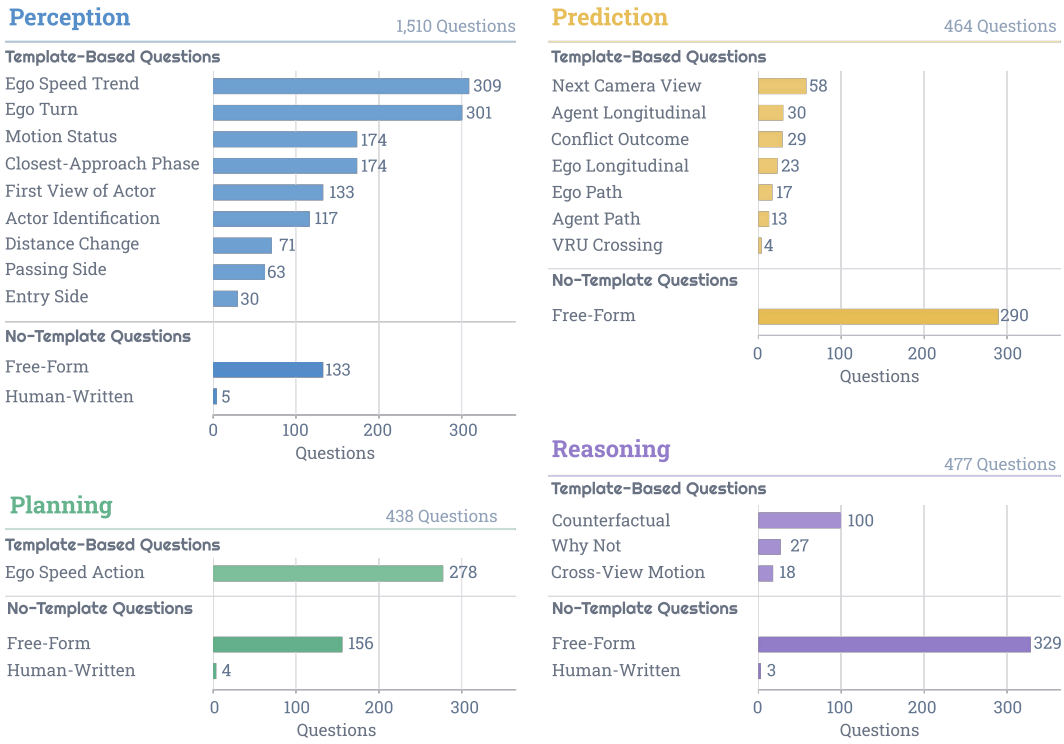


Figure 8: **Question counts by subtype within each competency.** Named template subtypes specify the queried property; Free-form denotes questions from recursive generation, and Human-written denotes separately authored additions. Free-form questions form the largest subtype in Prediction and Reasoning, whereas ego-speed actions dominate Planning. Counts combine MCQ and OpenVQA.

A.9 Dataset Statistics

What "Long-Tail" Means Here. Our use of "long-tail" is a statement about nested selection, not an estimate of event frequency in natural driving. The source corpus already carries the first selection: Waymo E2E is not a uniform sample of driving, but a set of segments its authors curated for rare and challenging behavior (Xu et al., 2025), mined from a far larger fleet corpus. Any segment we could have drawn is therefore already a tail sample with respect to ordinary driving, and we take that property from the source rather than re-deriving it. LTDriveQA applies a second selection on top of the first. We screened all 1,505 test segments, 5.7 hours of front-camera video, and retained 203 as candidate long-tail segments; 192 survive localization and review. The retained event windows are narrower still: taking the union of overlapping windows, the 719 clips cover 1,122 seconds, 5.5% of the screened video. Within a selected segment the retained windows cover 44.1% of its duration, so the reduction comes from discarding whole segments and then trimming each to its event, not from sampling short windows at random. The surviving categories are correspondingly skewed: 111 of the 203 segments are primarily construction zones, 41 pedestrians, and 18 complex intersections (Fig. 11). What this does and does not establish should be read precisely. It establishes that LTDriveQA sits two selection stages below ordinary driving, and it quantifies the second stage: 13.5% of already-curated segments and 5.5% of their video duration. It does not estimate how often these events occur per kilometer driven, does not fit a frequency distribution, and does not claim that every retained clip is individually rare. Readers who need an absolute rarity figure should take it from the source corpus's own curation, which we do not re-measure.

Question Subtype Distribution. The four competencies also exhibit distinct subtype distributions (Fig. 8). Ego-motion templates contribute substantially to Perception and Planning, whereas free-form questions constitute the largest subtype in Prediction and Reasoning. These differences reflect the distinct question structures and generation strategies underlying each competency.

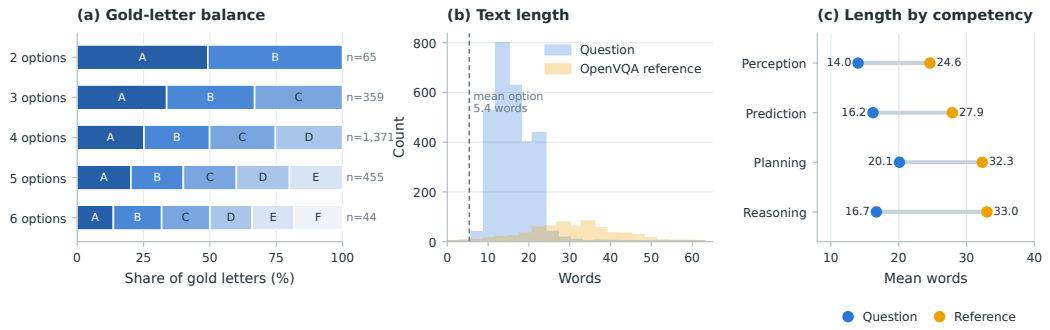


Figure 9: **Answer positions and text lengths.** (a) Correct-option frequencies within each option count, using MCQs only. (b) Word-count distributions for questions and OpenVQA references, with mean option length indicated. (c) Mean question and OpenVQA reference lengths by competency. Captioning items are excluded.



Figure 10: **Content-word frequencies across questions, options, and reference answers after removing stop words and common framing terms.** Font size is proportional to the square root of frequency.

Answer Positions and Text Lengths. Fig. 9 summarizes answer-position balance and text length. Correct letters are distributed approximately uniformly within each option count (panel a), limiting the advantage of a constant-letter strategy. OpenVQA references average about twice the length of questions (panel b). Planning questions are longest on average and often include current ego speed; Planning and Reasoning references tend to be longer because they describe an action or explanation with supporting scene details (panel c).

Vocabulary. The benchmark's frequent content words describe spatial relations, motion, and actors (Fig. 10). We count words in questions, options, and references after removing stop words and shared framing terms such as ego, vehicle, clip, and seconds. Frequent terms include left, right, lane, speed, slows, stationary, truck, sedan, and pedestrian. This vocabulary is consistent with the benchmark's emphasis on actor motion and position, although word frequencies alone do not measure evidence dependence.

From Segments to Events. Segment screening and clip localization describe different levels of the data (Fig. 11). Each of the 203 selected segments receives a primary screening category, most commonly construction or pedestrians (panel a). Localization produces the 719 retained clip windows, whose event labels include cross-view transits and ego braking as the largest groups (panel b). A construction segment can therefore contribute clips about braking, stopping, or pedestrians. The distributions of clips and distinct event types per represented segment (panel c) show why segment categories should not be treated as clip-event counts.

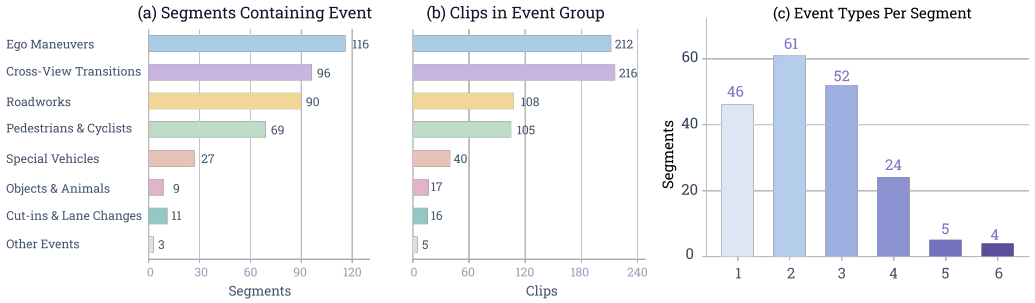


Figure 11: **From selected segments to retained event clips.** (a) Primary categories of the 203 screened segments. (b) Event labels of the 719 retained clip windows. (c) Distributions of clip count and distinct event-type count across the 192 represented segments. The panels use different counting units, as indicated.

B Additional Evaluation Details

B.1 Model Notes and Serving Settings

Serving and Decoding. Open-source models are served with vLLM 0.11.0 on H100 GPUs, using one GPU per replica or tensor parallelism of two for checkpoints with at least 30B parameters. Decoding uses temperature 0 and output budgets of 128 tokens for MCQ and 256 for OpenVQA. We disable thinking where an explicit switch is available, including Qwen3-VL, GLM-4.6V-Flash, and Cosmos-Reason2. Claude Opus 4.8 and GPT-5.6-Sol are accessed through Azure AI Foundry, with extended thinking disabled and reasoning effort set to none, respectively. Gemini 3.5 Flash uses its OpenAI-compatible endpoint with reasoning disabled. Apart from these switches, all other API parameters are left at their provider defaults. These are configured inference settings, not a verification of providers' internal computation.

Model-Specific Input Limits. Context and endpoint limits require documented exceptions to the common visual protocol. Qwen-Drive-1.0 is evaluated using the Qwen3.5-4B VLM checkpoint inside its planning wrapper, in VQA mode with a 32k context and a 22-frame cap. InternVL3.5 and ReCogDrive use at most four dynamic tiles per image and frame caps of 30 and 24, respectively, for their 40k and 32k contexts. LLaVA-1.6 receives at most eight uniformly spaced frames. The InternVL cap affects only three clips longer than six seconds. For GPT-5.6-Sol, the endpoint's 50-image limit requires thinning one 51-frame input to 50 frames. Consequently, comparisons use a shared rendering and task protocol but do not equalize visual-token budgets.

Proprietary Evaluation Cost. At published API prices, a full inference pass over LTDriveQA costs approximately $313 for Claude Opus 4.8 (caching disabled), $122 for GPT-5.6-Sol (with cache hits), and $43 for Gemini 3.5 Flash. These usage-based estimates cover all MCQ and OpenVQA questions, excluding judging and auxiliary experiments.

MCQ Prompt. Every MCQ is presented with a fixed zero-shot template and no in-context examples. The prompt states that the images are ordered chronologically and that each image is a labeled eight-camera composite captured at one timestamp, then gives the question, the lettered options in their released order, and the single instruction "Answer with only the option letter." No chain-of-thought is requested, and the same template is used for every model and every restricted-input control; the control conditions change only the sentence describing the images. Because models differ in how closely they follow the output instruction, the scored answer is not read from the reply verbatim but recovered by the parser described next.

Answer Parsing. A reply is scored only when it contains an explicit and unambiguous answer marker. The parser first strips packaging that carries no answer content: residual `<think>` blocks, answer-box markup such as `\boxed{C}` or a model's own box tokens, code fences, and a leading "Answer:". It then applies five ordered acceptance rules. A reply that is a bare letter, optionally wrapped in brackets, asterisks, or a trailing period, is read directly. A reply that is exactly one option's text is mapped to that option. A reply of the form "B. *option text*" is accepted only when the letter and the text name the same option; when they name different options, the reply is recorded as a conflict

Table 7: **Agreement between gpt-5.4-mini and human annotators on 150 OpenVQA responses from Qwen3-VL-8B ($n = 74$) and Claude Opus 4.8 ($n = 76$).** Both assess the question, reference, and response without video; the annotators are blinded to the judge's output. Agreement is binary-verdict agreement, and κ is Cohen's kappa.

Competency	n	Agreement (%)	κ
Perception	19	78.9	0.55
Prediction	45	91.1	0.82
Planning	31	90.3	0.79
Reasoning	55	89.1	0.77
Overall	150	88.7	0.77

and left unparsed rather than resolved in favor of either. A letter introduced inside a sentence by an explicit marker, such as "the answer is B" or "Option D", is accepted only when every such marker in the reply agrees. Finally, a reply that explains itself and then places the letter alone on its last line is read from that line, which is the reply's stated conclusion. Matching is case-insensitive and ignores surrounding punctuation and markdown emphasis, and a letter outside the question's option range is rejected.

Parsing Outcomes. Across the 25 evaluated models and all 2,294 released MCQs (57,350 replies), 80.6% are bare letters, 13.3% a letter followed by option text, 4.7% a letter introduced by an explicit marker, and 0.6% a letter on the reply's final line, leaving 0.96% unparsed. A single reply in the whole set is a letter–text conflict. The unparsed replies are concentrated in two models, Gemini 3.5 Flash and Qwen-Drive-1.0, which account for 548 of the 552; for both, the dominant cause is truncation at the 128-token MCQ budget rather than a refusal to answer, and the affected runs are flagged accordingly. All other models produce a parsed letter on more than 99% of MCQs.

B.2 OpenVQA Judge

Scoring Rubric. The judge assesses whether a response commits to the reference's central claim. It receives the competency, question, reviewed reference, and model answer, without video. Key-claim correctness is scored from 0 to 3: 3 denotes the central fact stated in equivalent wording; 2 denotes heavy qualification; 1 denotes the right subject without commitment; and 0 denotes contradiction or omission. Reversed directions, sides, or motion states count as contradictions. Supporting-detail accuracy receives 0–2 points, but the binary verdict is positive only when the key-claim score is 3. Length, fluency, and style are excluded. We report the binary acceptance rate and retain the 0–5 total for diagnostics. The judge is gpt-5.4-mini with a 512-token budget. In a small pilot of six answers repeated three times, verdicts were unchanged while total scores varied by up to one point.

Human Agreement Protocol. Two annotators independently scored 150 sampled OpenVQA responses to assess agreement with the automated rubric. The sample contains 55 Reasoning, 45 Prediction, 31 Planning, and 19 Perception responses, with 74 from Qwen3-VL-8B and 76 from Claude Opus 4.8. The annotators received the same question, reference, and response as the judge, without video or access to the judge's output, and assigned both a binary verdict and a 0–5 rubric total. This protocol tests reference-based scoring agreement; the reference itself is verified against the scene in the frame-grounded audit below.

Human Agreement Results. The judge and the human annotators agree on 133 of 150 binary verdicts (88.7%), with Cohen's $\kappa = 0.768$ and a bootstrap 95% confidence interval of $[0.660, 0.864]$ (Tab. 7). Of the 17 disagreements, six are judge acceptances rejected by the annotators and eleven are judge rejections accepted by the annotators. An exact McNemar test does not detect a directional difference in this sample ($p = 0.33$). Rubric totals have Spearman correlation $\rho = 0.78$ and a mean judge-minus-human difference of $+0.05$ points. These results support the rubric's agreement with human annotation, while the small competency subsets limit broader conclusions.

Frame-Grounded Audit Protocol. Agreement against the reference cannot detect an error in the reference itself: a response that describes the scene correctly but contradicts a faulty reference is rejected by judge and reference-blind annotators alike. To bound this failure mode we audited a

Table 8: **Frame-grounded audit of the OpenVQA judge.** A 40-item subset of the agreement sample was re-assessed by the same annotators with the eight-camera clip visible, so that the response could be checked against the scene rather than against the reference alone. Top: the 30 audited items that also carry a text-only verdict, scored under both conditions. Bottom: the reference audit over all 40 items. Intervals are Wilson 95% intervals.

Paired items, text-only vs. frame-grounded assessment ($n = 30$)		
	Text only	With video
Judge accepts, annotators accept	13	12
Judge accepts, annotators reject	0	1
Judge rejects, annotators accept	5	6
Judge rejects, annotators reject	12	11
Agreement (%)	83.3	76.7
κ	0.68	0.55
Annotator acceptances	18	18
Reference audit ($n = 40$)		
Reference incomplete or wrong		2 (5.0%)
of which the verdict was affected		1 (2.5%)
Judge–annotator disagreements with video		8 (20.0%)
attributable to the reference		1 (12.5%)

40-item subset of the agreement sample, drawn at random from the same stratum before annotation began and balanced across the two answering models. The annotators re-assessed each item with the eight-camera clip at 5 Hz visible, applying the same rubric but judging the response against the scene rather than against the reference, and additionally flagged whether the reference was incomplete or wrong. This audit measures end-to-end validity—whether an accepted response is in fact correct—which is a different quantity from the rubric fidelity measured in Tab. 7.

Frame-Grounded Audit Results. The annotators flagged the reference as incomplete or wrong on 2 of 40 items (5.0%, 95% CI $[1.4, 16.5]$), and on 1 of 40 (2.5%, 95% CI $[0.4, 12.9]$) the deficient reference changed the outcome (Tab. 8). In that case the reference states that a pedestrian crosses left-to-right while the clip shows the opposite direction; the judge rejected a response that had described the crossing correctly, citing the reversed direction its rubric treats as a contradiction. The second flagged item, a Prediction question about a parked truck's changing position relative to a road-work sign, has an underspecified reference, but judge and annotators both reject the response, so the verdict is unaffected. Of the 8 judge–annotator disagreements under the frame-grounded condition, therefore, 1 is attributable to the reference and the remainder to rubric application. On the 30 items carrying both verdicts, 4 (13.3%) changed when the video was shown—two acceptances withdrawn and two rejections reversed—leaving the annotators' acceptance count unchanged at 18 of 30. Agreement on these 30 items is 83.3% without video and 76.7% with it; the two values answer different questions and the difference is well within sampling error at this size.

Read together with Tab. 7, these results indicate that reference errors are a minor contributor to OpenVQA scoring error relative to rubric application, and that the reported acceptance rates are not materially inflated or deflated by the reference set.

Repeated and Cross-Judge Scoring. We also re-score all 595 Qwen3-VL-8B OpenVQA answers with a second gpt-5.4-mini pass and with Claude Opus 4.8 under the same rubric. The two gpt-5.4-mini passes agree on 95.5% of verdicts ($\kappa = 0.90$), yielding accuracies of 33.6% and 34.8%. Claude Opus 4.8 agrees with the first pass on 88.8% ($\kappa = 0.76$) and accepts 37.2% of responses. Disagreements often involve a correct actor but a qualified or noncommittal statement about the queried property. Thus, absolute scores vary with the judge even under a fixed rubric. Because this comparison covers one answering model, it does not establish ranking stability across all 25 models. Stored responses and verdicts permit alternative scoring without repeating model inference.

Table 9: **Accuracy (%) of 25 VLMs on LTDriveQA: 2,294 MCQs and 595 OpenVQA questions.** MCQ uses strict option-letter scoring; Open uses reference-based judge acceptance. Each competency reports MCQ, OpenVQA, and their mean (Avg.); the final Average is the unweighted mean of the eight competency–format cells, not accuracy over all questions. Bold and green mark the best performance in each column within each group. †: frame caps of 30 for InternVL3.5, 24 for ReCogDrive, and 22 for Qwen-Drive-1.0; the first two also use at most four dynamic tiles per image.

Model	Size	Perception			Prediction			Planning			Reasoning			Average
		MCQ	Open	Avg.	MCQ	Open	Avg.	MCQ	Open	Avg.	MCQ	Open	Avg.	
Proprietary VLMs														
Gemini 3.5 Flash	–	44.0	48.1	46.0	47.3	47.5	47.4	39.8	52.4	46.1	68.9	49.8	59.4	49.7
Claude Opus 4.8	–	61.0	48.1	54.5	46.3	63.0	54.6	43.3	**65.3**	**54.3**	67.8	**64.8**	**66.3**	57.4
GPT-5.6-Sol	–	**64.2**	**58.4**	**61.3**	**56.2**	**67.4**	**61.8**	**47.1**	58.9	53.0	**71.2**	56.8	64.0	**60.0**
Open-Source VLMs														
LLaVA-1.6	7B	27.6	26.0	26.8	36.4	14.9	25.6	36.9	25.0	30.9	41.7	20.7	31.2	28.6
GLM-4.6V-Flash	9B	50.5	41.6	46.0	44.2	24.9	34.5	43.9	42.7	43.3	56.4	34.7	45.5	42.4
Gemma 3	4B	22.4	18.2	20.3	30.4	10.5	20.4	35.4	16.9	26.1	41.3	12.7	27.0	23.5
Gemma 3	12B	33.6	24.7	29.1	32.5	12.2	22.4	40.8	32.3	36.5	47.7	19.2	33.5	30.4
Gemma 3	27B	38.9	23.4	31.1	33.6	21.0	27.3	34.7	29.0	31.9	51.9	16.0	34.0	31.1
InternVL3.5†	2B	36.8	24.7	30.8	30.0	17.7	23.9	**45.2**	33.1	39.2	43.6	18.3	31.0	31.2
InternVL3.5†	4B	44.3	33.8	39.0	31.4	15.5	23.4	35.7	33.9	34.8	50.0	26.8	38.4	33.9
InternVL3.5†	8B	52.1	32.5	42.3	33.2	27.6	30.4	36.6	41.1	38.9	53.4	28.6	41.0	38.2
InternVL3.5†	14B	45.1	35.1	40.1	42.0	26.0	34.0	36.3	36.3	36.3	53.8	24.9	39.3	37.4
InternVL3.5†	30B-A3B	46.1	33.8	40.0	37.8	21.0	29.4	37.6	41.9	39.8	55.7	32.4	44.0	38.3
InternVL3.5†	38B	51.1	24.7	37.9	37.8	26.0	31.9	38.5	**46.0**	42.2	62.1	33.3	47.7	39.9
Qwen3-VL	2B	40.5	18.2	29.4	31.1	16.6	23.9	38.9	28.2	33.5	48.5	21.6	35.0	30.4
Qwen3-VL	4B	49.9	28.6	39.2	36.0	23.8	29.9	40.4	37.1	38.8	56.1	35.7	45.9	38.4
Qwen3-VL	8B	50.5	37.7	44.1	35.7	26.0	30.9	42.7	37.1	39.9	58.0	36.6	47.3	40.5
Qwen3-VL	32B	**55.1**	**48.1**	**51.6**	**50.9**	**42.0**	**46.5**	43.6	44.4	**44.0**	**67.4**	**44.1**	**55.8**	**49.4**
Open-Source Driving and Embodied Models														
Cosmos-Reason2	2B	46.3	20.8	33.5	32.9	18.2	25.5	39.5	27.4	33.5	51.1	23.5	37.3	32.5
Cosmos-Reason2	8B	**58.3**	22.1	40.2	39.6	23.8	31.7	42.7	32.3	37.5	61.0	26.3	43.6	38.2
Cosmos-Reason2	32B	54.4	31.2	42.8	**45.2**	24.9	35.0	**43.3**	46.0	**44.6**	**65.9**	35.7	50.8	43.3
ReCogDrive†	2B	45.1	26.0	35.5	28.6	30.4	29.5	35.0	38.7	36.9	41.7	26.3	34.0	34.0
ReCogDrive†	8B	49.0	20.8	34.9	38.9	28.7	33.8	41.1	43.5	42.3	49.6	29.6	39.6	37.6
MiMo-Embodied	7B	53.5	36.4	45.0	38.2	28.2	33.2	39.2	46.0	42.6	58.0	26.3	42.1	40.7
Qwen-Drive-1.0†	4B	51.4	**40.3**	**45.8**	39.9	**40.9**	**40.4**	37.3	**50.0**	43.6	59.8	**45.5**	**52.6**	**45.6**

C Additional Experimental Results

C.1 Comprehensive Results of Evaluated Models

Tab. 9 shows the comprehensive results of 25 evaluated VLMs.

Overall Model Comparison. The strongest models still make substantial errors on LTDriveQA (Tab. 2). GPT-5.6-Sol leads the eight-cell macro-average at 60.0%, followed by Claude Opus 4.8 at 57.4%. The Average weights the eight competency and format cells equally. Question-level accuracy over all 2,889 items would instead give Perception MCQs, half of the benchmark, half of the weight; it ranks models similarly (Spearman $\rho = 0.95$) and keeps GPT-5.6-Sol and Claude Opus 4.8 on top, but lifts models that are strong mainly on Perception MCQs, such as Cosmos-Reason2-8B from 13th to 5th. Qwen3-VL-32B is the strongest open-source model at 49.4%, a gap of 10.6 points from GPT-5.6-Sol. Overall MCQ accuracy over all 2,294 questions reaches 61.7% for GPT-5.6-Sol and 54.4% for Qwen3-VL-32B, compared with the option-count chance rate of 25.9%. Gemini 3.5 Flash scores 46.7% on MCQ, with 15.6% invalid responses under the fixed output budget; Qwen-Drive-1.0 also has a substantial invalid-response rate of 8.3%. Their scores therefore reflect instruction following and answer completion as well as scene understanding. All other models produce valid letters on more than 99% of MCQs.

Model Scale and Specialization. Larger models generally perform better within a family, but the gains are uneven. Qwen3-VL's macro-average rises from 30.4% at 2B to 49.4% at 32B, while InternVL3.5 rises from 31.2% at 2B to 39.9% at 38B, with little separation among its 8B–30B variants. Gemma 3 improves from 23.5% to 31.1% across 4B–27B. Driving or embodied models do not consistently exceed general-purpose models: Cosmos-Reason2 rises from 32.5% at 2B to 43.3% at 32B, above Qwen3-VL at 2B but below it at 8B and 32B, while Qwen-Drive-1.0 reaches 45.6%, above Qwen3-VL-4B's 38.4%. Because these comparisons also differ in backbone and training data,

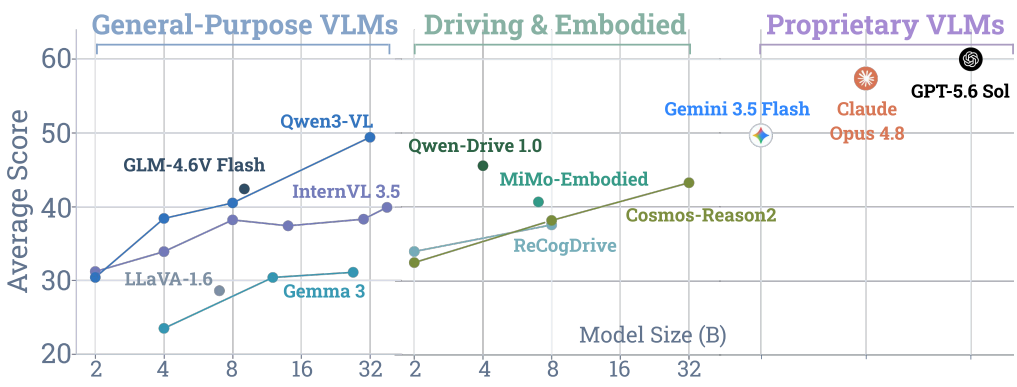


Figure 12: **Model performance versus model size.** The vertical axis reports the eight-cell macro-average score in Tab. 2, and the horizontal axis denotes model size (B) for models with disclosed parameter counts. Models are grouped into general-purpose VLMs, driving and embodied models, and proprietary VLMs, with lines connecting different scales within the same family. Proprietary models with undisclosed parameter counts are shown separately.

they do not isolate the effect of domain post-training. Fig. 12 places the results in the context of model size.

Performance across Answer Formats. The proprietary–open-source gap is larger on OpenVQA than on MCQ. Unlike the main-text comparison, which averages each format over the four competencies (a GPT-5.6-Sol–Qwen3-VL-32B gap of 5.4 points on MCQ and 15.7 on OpenVQA; Tab. 2), the accuracies here are computed over all questions of each format. The best MCQ results differ by 7.3 points (61.7% for GPT-5.6-Sol versus 54.4% for Qwen3-VL-32B). On OpenVQA, Claude Opus 4.8 reaches 62.2% and GPT-5.6-Sol 60.7%, while the best open-source results are 44.4% for Qwen-Drive-1.0 and 44.0% for Qwen3-VL-32B. Because Claude Opus 4.8 also annotated the OpenVQA reference answers (Appendix A.3), its OpenVQA scores may partly benefit from proximity to the references' wording; the judge scores only the key claim and excludes length and style (Appendix B.2), which limits but does not rule out this effect. OpenVQA requires generating the queried fact or decision without a supplied option set, but the two formats also contain different questions and use different scoring rules. Their gap therefore cannot be attributed to response format alone. Competency rankings can also reverse: Claude Opus 4.8 scores lowest on Planning MCQs (43.3%) but highest on Planning OpenVQA (65.3%). Although overall model rankings are strongly correlated across formats (Spearman $\rho = 0.86$), separate reporting preserves these differences.

Human Baseline on a Matched Subset. The Human row of Tab. 2 is measured on 100 MCQs, 25 per competency, drawn from the released set with a fixed seed and shown with the same 5 Hz eight-camera input as the models. Tab. 10 scores all 25 models on the same questions. Humans reach 91.0%, whereas the best model, Qwen3-VL-32B, reaches 65.0% and GPT-5.6-Sol 59.0%. The gap holds in every competency, from 17.3 points on Reasoning to 25.3 points on Planning. Model rankings on this subset agree with those on all MCQs (Spearman $\rho = 0.81$), although 25 questions per competency make individual cells noisy.

C.2 Restricted-Input Controls across Open-Source Models

Question-Only Controls. Full-video input improves MCQ accuracy over question-only input for every evaluated open-source model (Tab. 11). Question-only scores range from 25.1% to 36.3%, compared with the full-set chance rate of 25.9%. Video gains range from 1.7 points for LLaVA-1.6 to 23.3 points for Cosmos-Reason2-8B. Cosmos-Reason2-32B, Qwen3-VL-32B, GLM-4.6V-Flash, and InternVL3.5-38B also gain 18.4–20.6 points. These comparisons use the full 2,294-question set with the main-run prompts and decoding budgets, and provide the per-model results behind the averages in Tab. 4.

Current-Frame Controls. Temporal context helps many models, but the gain over a single observation is not universal. Video exceeds the final eight-camera composite by 9.9 points for Cosmos-

Table 10: **MCQ accuracy (%) of humans and all 25 models on the human subset.** Humans and models receive the same questions, options, and 5 Hz eight-camera input. Bold marks the best model per column; Chance is the mean of $1/K$.

Model	Size	Perception	Prediction	Planning	Reasoning	Overall
Human	–	88.0	85.3	93.3	97.3	91.0
Gemini 3.5 Flash	–	20.0	56.0	40.0	60.0	44.0
Claude Opus 4.8	–	60.0	52.0	32.0	72.0	54.0
GPT-5.6-Sol	–	48.0	56.0	60.0	72.0	59.0
LLaVA-1.6	7B	24.0	52.0	48.0	48.0	43.0
GLM-4.6V-Flash	9B	48.0	60.0	48.0	48.0	51.0
Gemma 3	4B	16.0	44.0	44.0	52.0	39.0
Gemma 3	12B	16.0	48.0	48.0	60.0	43.0
Gemma 3	27B	40.0	48.0	32.0	68.0	47.0
InternVL3.5	2B	44.0	36.0	**68.0**	56.0	51.0
InternVL3.5	4B	40.0	28.0	44.0	44.0	39.0
InternVL3.5	8B	60.0	44.0	24.0	68.0	49.0
InternVL3.5	14B	56.0	40.0	52.0	60.0	52.0
InternVL3.5	30B-A3B	52.0	48.0	36.0	56.0	48.0
InternVL3.5	38B	60.0	40.0	56.0	76.0	58.0
Qwen3-VL	2B	40.0	28.0	48.0	52.0	42.0
Qwen3-VL	4B	56.0	40.0	40.0	72.0	52.0
Qwen3-VL	8B	48.0	48.0	56.0	64.0	54.0
Qwen3-VL	32B	**64.0**	**64.0**	56.0	76.0	**65.0**
Cosmos-Reason2	2B	40.0	40.0	48.0	60.0	47.0
Cosmos-Reason2	8B	52.0	24.0	48.0	**80.0**	51.0
Cosmos-Reason2	32B	60.0	56.0	52.0	76.0	61.0
ReCogDrive	2B	48.0	16.0	40.0	44.0	37.0
ReCogDrive	8B	56.0	32.0	44.0	44.0	44.0
MiMo-Embodied	7B	48.0	56.0	52.0	68.0	56.0
Qwen-Drive-1.0	4B	40.0	56.0	44.0	68.0	52.0
Chance	–	26.5	27.0	20.7	23.0	24.3

Reason2-32B, 8.1 for Cosmos-Reason2-8B, 7.5 for InternVL3.5-14B, and 7.2 for Qwen3-VL-32B. Several models gain little or score slightly lower with video, including LLaVA-1.6, InternVL3.5-2B, Gemma 3 27B, Cosmos-Reason2-2B, and ReCogDrive-2B. A single timestamp therefore accounts for a substantial share of performance, and the benefit of additional frames depends on the model. These descriptive differences alone do not establish significance for each model or necessity for each question.

Controls by Competency. Tab. 12 splits both gains by competency. Removing images lowers mean accuracy in every competency, and lowers it for all 22 models on Perception and Reasoning. The gain over the final frame is concentrated in Perception, where 19 of the 22 models improve, and it grows with scale within the Qwen3-VL and Cosmos-Reason2 families. Planning and Reasoning show small mean gains whose intervals include zero. Prediction reverses the pattern: 17 of the 22 models score higher with the final frame alone, and the mean gain is negative with an interval that excludes zero. The reversal comes mainly from questions about the next camera view or the future path of an actor or the ego vehicle, whose answers depend on the most recent position; free-form Prediction questions show almost no difference. Current models thus make limited use of video history, and earlier frames can pull their predictions away from the latest observed state.

C.3 Restricted-Input Controls for OpenVQA

The controls in Tab. 11 cover the multiple-choice half of LTDriveQA only. Because MCQ accuracy can be raised by option-set priors that have no counterpart in free-form answering, the evidence-dependence of the OpenVQA half does not follow from them. We therefore repeat the two restricted-input conditions on all 595 OpenVQA questions for Qwen3-VL-8B and InternVL3.5-8B, the two models whose full-protocol OpenVQA rows in Tab. 2 we reuse as the video baseline.

Table 11: **MCQ accuracy (%) on 2,294 questions under full-video, question-only, and current-frame inputs.** Video uses the 5 Hz protocol with the model-specific caps in Tab. 9; Current frame uses the final permitted eight-camera composite. The last column reports the video-minus-question-only difference in percentage points. Proprietary video scores are included for reference; dashes indicate unrun controls.

Model	Size	Video	Question only	Current frame	Video − question only
Claude Opus 4.8	–	57.5	–	–	–
GPT-5.6-Sol	–	61.7	–	–	–
Gemini 3.5 Flash	–	46.7	–	–	–
LLaVA-1.6	7B	31.6	29.9	32.4	+1.7
GLM-4.6V-Flash	9B	49.5	29.8	46.8	+19.7
Qwen3-VL	2B	40.1	25.5	38.0	+14.6
Qwen3-VL	4B	47.6	33.2	43.7	+14.4
Qwen3-VL	8B	48.4	31.4	43.3	+17.0
Qwen3-VL	32B	54.4	34.6	47.2	+19.8
InternVL3.5	2B	37.9	30.8	39.3	+7.1
InternVL3.5	4B	42.2	28.9	41.8	+13.3
InternVL3.5	8B	47.8	30.5	44.4	+17.3
InternVL3.5	14B	44.5	27.1	37.0	+17.4
InternVL3.5	30B-A3B	45.0	30.8	41.5	+14.2
InternVL3.5	38B	49.0	30.6	42.9	+18.4
Gemma 3	4B	27.3	25.1	27.2	+2.3
Gemma 3	12B	36.1	28.9	34.4	+7.1
Gemma 3	27B	39.2	29.3	40.5	+9.9
Cosmos-Reason2	2B	44.2	29.9	44.6	+14.3
Cosmos-Reason2	8B	54.1	30.9	46.0	+23.3
Cosmos-Reason2	32B	53.1	32.5	43.2	+20.6
ReCogDrive	2B	41.3	26.3	41.5	+15.0
ReCogDrive	8B	46.7	31.8	43.5	+14.9
MiMo-Embodied	7B	50.1	33.2	47.3	+16.9
Qwen-Drive-1.0	4B	49.0	36.3	47.7	+12.7

Protocol. Only the visual input changes. The *video* condition is the 5 Hz eight-camera protocol of the main run; *current frame* sends the final permitted eight-camera composite alone; *question only* sends no image and drops the video preamble from the prompt. The question text, the 256-token output budget, the temperature, the reviewed reference answers and the judge rubric are identical across conditions, and prediction inputs still stop at the recorded observation endpoint. Responses are generated once and stored, so every condition is scored by the same judge pass described in Appendix B.2. Unlike MCQ, OpenVQA has no option set and therefore no chance rate: a question-only score above zero measures how often a driving prior alone reproduces the reference claim.

Visual and Temporal Evidence in Free-Form Answers. Removing the video lowers OpenVQA acceptance by 12.94 points for Qwen3-VL-8B (33.61% to 20.67%) and by 15.97 points for InternVL3.5-8B (31.43% to 15.46%); exact paired McNemar tests on the binary verdicts give $p = 1.2 \times 10^{-8}$ and $p = 8.7 \times 10^{-15}$. These gaps are not larger than the MCQ gaps of $+17.00$ and $+17.30$ points that the same two models show in Tab. 11, so the free-form format does not by itself widen the measured dependence on visual evidence. The two figures are not on a common scale: MCQ question-only accuracy is bounded below by the 25.86% option-count chance rate, a free-form answer has no such floor, and the video baselines differ (48.43% and 47.80% for MCQ against 33.61% and 31.43% for OpenVQA). Measured against each model's own video baseline the free-form losses are the larger of the two, 38.5% and 50.8% of the baseline against 35.1% and 36.2% for MCQ. We therefore read these controls as showing that the MCQ gaps do not overstate evidence dependence, rather than as establishing that they understate it.

Evidence from a Single Observation. Replacing the clip with its final composite costs Qwen3-VL-8B 5.88 points (33.61% to 27.73%, $p = 0.0058$), a loss concentrated in Reasoning, which falls 11.27 points (36.62% to 25.35%, $p = 9.4 \times 10^{-4}$); Perception falls 6.49 points but is inconclusive on its 77 items ($p = 0.36$). InternVL3.5-8B shows no detectable overall loss (31.43% to 32.77%, a -1.34-point change with $p = 0.54$): its Perception and Planning acceptance is identical under both inputs, and its Prediction acceptance is 7.73 points higher from the single frame ($p = 0.059$). A

Table 12: **Per-competency MCQ gains (percentage points) of full video over question-only and final-frame inputs for the 22 open-source models.** Bold marks $p < 0.05$ under an exact paired McNemar test. Mean averages the per-question gains over models; its 95% interval resamples source segments. The last row counts models with a positive gain.

Model	Size	Gain over question only				Gain over final frame			
		Perc.	Pred.	Plan.	Reas.	Perc.	Pred.	Plan.	Reas.
LLaVA-1.6	7B	+0.4	+4.6	−2.2	**+10.6**	−1.7	−1.1	+0.3	+3.0
GLM-4.6V-Flash	9B	**+23.7**	**+12.0**	**+8.9**	**+18.9**	**+4.9**	**−7.1**	−2.5	**+7.6**
Qwen3-VL	2B	**+19.3**	+4.2	+1.6	**+15.2**	**+4.2**	−3.9	+0.0	−0.8
Qwen3-VL	4B	**+20.0**	+1.1	−7.3	**+23.9**	**+6.4**	−2.8	+0.6	+1.5
Qwen3-VL	8B	**+20.4**	+3.9	+6.7	**+25.0**	**+7.5**	−2.1	+2.2	+3.4
Qwen3-VL	32B	**+22.9**	**+12.0**	**+8.6**	**+25.0**	**+11.4**	−3.5	+4.1	−0.4
InternVL3.5	2B	**+8.0**	+1.8	**+5.1**	**+10.2**	−1.4	−4.2	+0.0	−0.4
InternVL3.5	4B	**+17.5**	+3.2	+3.5	**+12.9**	+1.4	+0.0	−2.2	−1.1
InternVL3.5	8B	**+23.8**	+3.9	+1.3	**+15.9**	**+5.7**	−1.8	−2.9	+4.5
InternVL3.5	14B	**+20.4**	**+10.6**	**+8.6**	**+18.9**	**+12.1**	−0.7	−1.6	+1.9
InternVL3.5	30B-A3B	**+16.7**	**+11.3**	+2.9	**+17.4**	**+4.1**	+0.4	**+4.8**	+1.9
InternVL3.5	38B	**+22.6**	+4.9	**+8.0**	**+22.3**	**+9.8**	−2.8	+2.2	−0.4
Gemma 3	4B	+0.6	−1.8	**+10.8**	+5.3	+0.3	−4.9	−0.6	**+6.1**
Gemma 3	12B	**+6.1**	+2.1	**+9.9**	**+14.8**	+2.0	−1.8	+2.2	+2.7
Gemma 3	27B	**+11.9**	+1.4	+6.4	**+12.1**	−2.8	−1.8	+1.6	+3.4
Cosmos-Reason2	2B	**+17.3**	**+6.4**	+5.4	**+17.0**	+0.3	−3.9	+0.0	−0.4
Cosmos-Reason2	8B	**+30.9**	**+8.5**	**+7.0**	**+17.0**	**+12.6**	−1.4	+1.0	+2.3
Cosmos-Reason2	32B	**+26.7**	**+7.1**	**+6.7**	**+18.9**	**+13.7**	+0.7	**+7.3**	+1.9
ReCogDrive	2B	**+20.0**	+3.9	**+6.7**	**+9.8**	+0.3	+0.7	+0.3	**−4.2**
ReCogDrive	8B	**+17.4**	**+10.2**	**+9.9**	**+12.1**	**+2.7**	+4.6	**+7.0**	−0.4
MiMo-Embodied	7B	**+22.1**	+6.7	+0.3	**+19.3**	**+4.2**	−2.1	+0.6	+3.0
Qwen-Drive-1.0	4B	**+14.9**	+5.7	**+7.6**	**+14.4**	**+3.6**	**−8.5**	+1.6	−0.8
Mean		+17.4	+5.6	+5.3	+16.2	+4.6	−2.2	+1.2	+1.6
95% CI		[+16.2, +18.7]	[+3.4, +7.9]	[+2.8, +7.7]	[+12.7, +20.0]	[+3.5, +5.7]	[−3.6, −0.8]	[−0.2, +2.5]	[−0.1, +3.3]
Models with a gain		22	21	20	22	19	4	14	13

single well-chosen timestamp therefore supports most reference claims that the full clip supports, echoing the MCQ result of Tab. 4, while the residual temporal requirement is model-dependent rather than a fixed property of the question set.

Competencies with a Free-Form Prior. Planning retains the highest question-only acceptance for both models, 32.26% for Qwen3-VL-8B and 33.06% for InternVL3.5-8B against video values of 37.10% and 41.13%, and neither drop is significant ($p = 0.42$ and $p = 0.099$). This mirrors the above-chance Planning prior in the MCQ averages of Tab. 4: a generic action recommendation often reproduces the reference claim without scene-specific evidence, so the absolute Planning numbers in Tab. 2 should not be read as evidence-grounded planning competence. Qwen3-VL-8B shows the same pattern on Prediction, where question-only acceptance of 27.07% is statistically indistinguishable from its 26.00% video acceptance (-1.07 points, $p = 0.89$), indicating no measurable use of the video on that subset. InternVL3.5-8B does not share it, losing 18.78 points on Prediction without video ($p = 1.2 \times 10^{-6}$), so the effect is a property of a particular model on a particular competency rather than of the Prediction questions alone.

Statistics and Scope. All p values above are exact paired McNemar tests over the 595 shared questions, computed on the judge's binary verdicts from unrounded accuracies. Because the judge scores every condition against the same reviewed reference under the same rubric, a difference between conditions reflects the model's response and not a change in the scoring criterion. Two models do not establish this pattern for the other 23 rows of Tab. 9, and the per-competency subsets are small enough (77–213 items) that inconclusive cells should not be read as evidence of no effect.

Question Composition and Difficulty. Benchmark frequency alone does not explain event-level accuracy. The rank correlation between event question count and mean MCQ accuracy is 0.43 ($p = 0.14$), providing no clear evidence of an association in these slices. Other descriptive differences are substantial: mean accuracy is 44.6% for template questions versus 52.0% for free-form questions, and 52.9% for questions labeled multi-view versus 43.5% for front-view questions. Templates often ask about fine-grained ego motion or temporal localization, whereas many multi-view questions ask about

Table 13: **OpenVQA acceptance rate (%) on 595 free-form questions under full-video, question-only, and current-frame inputs.** All conditions share the question text, the reviewed reference, the 256-token budget and the judge rubric of Appendix B.2; only the visual input changes. Video reuses the main-run rows of Tab. 2. Competency sizes are 77/181/124/213 for Perception/Prediction/Planning/Reasoning. The last row is the video-minus-question-only difference in percentage points, computed before rounding.

Model	Input	Perception	Prediction	Planning	Reasoning	Overall
Qwen3-VL-8B	Video (5 Hz)	37.7	26.0	37.1	36.6	33.6
	Question only	7.8	27.1	32.3	13.1	20.7
	Current frame	31.2	22.1	37.9	25.4	27.7
	Video − question only	+29.9	−1.1	+4.8	+23.5	+12.9
InternVL3.5-8B	Video (5 Hz)	32.5	27.6	41.1	28.6	31.4
	Question only	6.5	8.8	33.1	14.1	15.5
	Current frame	32.5	35.4	41.1	25.8	32.8
	Video − question only	+26.0	+18.8	+8.1	+14.6	+16.0

structured camera transitions. These differences show that more views or more elaborate wording do not automatically make a question harder. Appendix C.4 reports the model-level breakdowns, and the view ablation in Appendix C.6 tests camera coverage on the same questions.

C.4 Performance by Event Type, View Requirement, and Question Source

Event Types. Tab. 14 reports MCQ accuracy for the eleven largest event groups. Event types are grouped for reporting: Fig. 3(b) merges the 15 annotated event types into eight groups, whereas this table uses a finer grouping of the MCQs. Cross-view transits are relatively well answered across many models, while ego-start and special-vehicle questions often have lower scores. Exceptions matter: Claude Opus 4.8 reaches 59.2% on ego-start questions, whereas Qwen3-VL-32B reaches 38.0%. GPT-5.6-Sol's advantage over the strongest open-source result exceeds ten points on ego braking and foreign objects. Event groups differ in size and question composition, so fine-grained rankings, especially on smaller groups, should be treated as descriptive.

View Requirement. Questions labeled multi-view are not uniformly harder than front-view questions (Tab. 15). Twenty-four of the 25 models score higher on the multi-view MCQ subset; Qwen-Drive-1.0 is the exception. These subsets contain different questions, including many structured transit questions in the multi-view group, so their score difference does not measure the effect of adding cameras. OpenVQA differences vary in direction and can exceed ten points, as for Qwen3-VL-4B. Appendix C.6 instead compares camera coverage on the same questions.

Question Source. Free-form MCQs generally receive higher accuracy than template MCQs, with ReCogDrive-2B the only exception in Tab. 15. Templates concentrate on properties such as ego motion and closest-approach timing, while free-form generation produces a different question mix. The comparison therefore describes the resulting subsets rather than the quality or difficulty of a construction method in isolation. The contrast also offers a check on construction bias: free-form questions were authored by Qwen3-VL models, yet Qwen3-VL-8B and 32B gain 6.9 and 7.2 points on them over template MCQs, close to the median of 7.5 points across the other 21 models. Gemma 3 drafted the initial MCQ answers, and its models still rank near the bottom of Tab. 9. We therefore find no clear advantage for the construction model families. OpenVQA includes 594 free-form items and one separately human-written item, with no template subset to compare.

C.5 Temporal Evidence Dependency

Frames and Temporal Order. Current-frame and frame-order controls test different uses of video evidence. Replacing video with its final observation removes both earlier scene content and temporal ordering; gains in Tab. 11 can reflect either source. Shuffling instead preserves the set of frames while changing their order, providing a more direct test of sensitivity to temporal presentation.

Table 14: **MCQ accuracy (%) for the eleven largest event groups, with counts in the Questions row.** These columns cover 2,243 of 2,294 MCQs; the omitted groups are turn/lane change ($n = 32$) and other events ($n = 19$). Event labels describe the underlying clip, and each group can contain multiple competencies.

Model	Size	Cross-view transit	Construction	Ego brake	Pedestrian	Ego stop	Special vehicle	Ego accelerate	Ego start	Cyclist	Foreign object	Cut-in
questions		484	486	337	287	162	146	117	71	61	50	42
Claude Opus 4.8	–	67.8	61.9	47.2	54.4	51.9	53.4	62.4	59.2	52.5	44.0	61.9
GPT-5.6-Sol	–	68.8	64.4	64.4	56.8	63.0	52.1	50.4	52.1	49.2	62.0	69.0
Gemini 3.5 Flash	–	55.8	45.7	41.5	49.1	47.5	39.0	35.9	38.0	49.2	42.0	57.1
GLM-4.6V-Flash	9B	60.3	51.0	48.7	43.9	43.8	41.8	40.2	36.6	47.5	46.0	52.4
LLaVA-1.6	7B	43.8	26.1	34.4	30.3	30.9	24.7	20.5	14.1	29.5	30.0	31.0
Qwen3-VL	2B	51.9	35.8	45.4	35.2	35.8	36.3	29.1	28.2	42.6	42.0	38.1
Qwen3-VL	4B	60.1	46.5	41.5	49.5	48.1	41.1	38.5	45.1	44.3	28.0	38.1
Qwen3-VL	8B	58.9	48.1	47.2	49.5	51.2	34.2	41.9	40.8	39.3	30.0	47.6
Qwen3-VL	32B	64.9	52.3	52.8	56.4	54.3	41.8	46.2	38.0	59.0	46.0	54.8
InternVL3.5	2B	49.4	30.9	38.9	36.2	42.0	32.2	29.9	26.8	41.0	30.0	38.1
InternVL3.5	4B	53.7	42.6	35.9	39.0	38.3	41.8	40.2	35.2	34.4	30.0	33.3
InternVL3.5	8B	55.2	51.2	43.9	46.3	50.0	41.1	38.5	46.5	41.0	42.0	40.5
InternVL3.5	14B	55.0	47.1	38.0	41.1	40.7	40.4	35.9	36.6	41.0	44.0	40.5
InternVL3.5	30B-A3B	56.6	46.1	39.2	42.9	42.0	37.7	34.2	29.6	42.6	44.0	40.5
InternVL3.5	38B	55.4	52.7	47.8	47.7	51.2	35.6	35.0	39.4	47.5	50.0	50.0
Gemma 3	4B	37.2	21.8	34.1	27.2	22.2	21.9	19.7	16.9	31.1	20.0	16.7
Gemma 3	12B	47.9	32.1	36.2	31.0	37.0	31.5	33.3	22.5	39.3	26.0	40.5
Gemma 3	27B	53.5	37.9	40.7	32.1	35.2	26.0	33.3	28.2	36.1	36.0	31.0
Cosmos-Reason2	2B	52.1	43.8	41.2	44.3	46.3	40.4	38.5	38.0	36.1	36.0	52.4
Cosmos-Reason2	8B	58.3	60.5	51.3	51.2	54.3	45.2	42.7	50.7	47.5	48.0	66.7
Cosmos-Reason2	32B	65.3	51.9	48.4	54.4	50.6	45.9	41.9	36.6	45.9	50.0	71.4
ReCogDrive	2B	45.9	44.2	39.5	39.4	43.2	38.4	41.0	28.2	34.4	32.0	33.3
ReCogDrive	8B	52.5	50.0	41.8	43.6	46.9	43.8	47.0	36.6	41.0	32.0	52.4
MiMo-Embodied	7B	58.3	51.4	48.4	48.4	47.5	40.4	48.7	47.9	42.6	40.0	57.1
Qwen-Drive-1.0	4B	49.8	51.0	54.3	53.0	49.4	42.5	34.2	35.2	55.7	40.0	50.0

Temporal Subset from Construction Probes. Probe outcomes recorded during construction identify the questions on which video helps most. Before any evaluation run, Gemma-3-27B answered 1,078 MCQs from the first, middle, and last frame of the observation window, one frame at a time. We group these MCQs by how many probed frames reproduce the reference and evaluate each group with the 21 open-source models other than Gemma-3-27B (Tab. 16). When no probed frame reproduces the reference, video exceeds the final frame by 5.7 points and improves 19 of the 21 models. When every probed frame reproduces it, the gain is 1.9 points and 10 models improve. The difference of 3.8 points holds under segment resampling, with a 95% interval of $[1.3, 6.2]$. These questions are not simply harder: their video accuracy is close to that of the questions every frame answers, and adding images to the question helps them less. The probe outcome thus marks a subset on which models benefit from several observations.

Temporal Ordering. Shuffling frames lowers Qwen3-VL-8B accuracy from 48.4% to 44.6% on the full MCQ set (3.9 points before rounding; paired McNemar $p < 10^{-6}$). The comparison applies one random permutation with a fixed seed and keeps the frames, prompt, and option order unchanged. Perception loses 5.4 points ($p < 10^{-7}$), and questions about ego turns lose 20.9 points (55.5% to 34.6%, $p < 10^{-13}$), although the final frame alone nearly matches video on them (53.5%). Questions asking where an actor first appears show the opposite pattern: shuffling lowers accuracy from 61.7% to 57.1%, whereas the final frame alone lowers it to 29.3%. The number of observations and their order therefore affect different questions, and LTDriveQA contains questions sensitive to each.

C.6 Multi-View Evidence Dependency

View-Coverage Protocol. The view ablation compares camera coverage on the same 2,294 MCQs (Tab. 17). We render the same 5 Hz timestamps with the front camera alone or all eight cameras; Qwen3-VL-8B also receives a three-forward-camera condition. Composite width, option order, and decoding settings remain fixed. This design changes per-camera resolution as well as coverage: a front-only image spans 1,536 pixels in width, whereas the front tile in the eight-camera layout is one

Table 15: **Accuracy (%) by intended view requirement and question source.** MCQ view subsets contain 1,773 Front and 510 Multi items; source subsets contain 1,969 Template and 314 Free-form items. OpenVQA contains 493 Front, 101 Multi, and 594 Free-form items, with no Template items. Eleven MCQs and one OpenVQA item lack view labels and are human-written, so they are excluded from the displayed subsets. These are comparisons between different questions, not input ablations.

Model	Size	MCQ				OpenVQA			
		Front	Multi	Template	Free-form	Front	Multi	Template	Free-form
Claude Opus 4.8	–	55.8	64.1	56.5	65.0	61.7	64.4	–	62.1
GPT-5.6-Sol	–	61.1	63.9	60.7	68.2	60.0	63.4	–	60.6
Gemini 3.5 Flash	–	45.2	52.5	44.1	64.0	49.1	51.5	–	49.5
GLM-4.6V-Flash	9B	47.3	57.6	49.0	53.2	33.9	35.6	–	34.2
LLaVA-1.6	7B	28.0	44.1	30.5	38.5	21.7	14.9	–	20.5
Qwen3-VL	2B	37.6	48.8	38.4	50.6	21.1	20.8	–	21.0
Qwen3-VL	4B	45.1	56.7	46.4	55.7	29.6	40.6	–	31.5
Qwen3-VL	8B	46.2	56.5	47.6	54.5	32.3	40.6	–	33.7
Qwen3-VL	32B	51.7	64.7	53.6	60.8	44.6	40.6	–	43.9
InternVL3.5	2B	35.2	47.5	36.5	47.1	22.9	17.8	–	22.1
InternVL3.5	4B	39.9	50.6	41.2	49.4	24.7	29.7	–	25.6
InternVL3.5	8B	46.9	51.6	47.7	49.4	31.4	30.7	–	31.3
InternVL3.5	14B	41.7	54.5	44.0	47.8	28.8	28.7	–	28.8
InternVL3.5	30B-A3B	42.5	53.7	44.2	50.0	30.0	35.6	–	31.0
InternVL3.5	38B	47.8	53.5	48.5	53.2	32.5	32.7	–	32.5
Gemma 3	4B	24.8	36.3	25.6	38.2	13.8	12.9	–	13.6
Gemma 3	12B	33.1	46.1	34.3	46.2	20.9	17.8	–	20.4
Gemma 3	27B	35.6	51.6	38.1	45.9	21.9	17.8	–	21.2
Cosmos-Reason2	2B	43.1	48.6	43.0	52.5	24.1	13.9	–	22.4
Cosmos-Reason2	8B	53.8	56.1	53.5	59.6	26.8	23.8	–	26.3
Cosmos-Reason2	32B	50.1	63.5	52.4	58.0	34.7	30.7	–	34.0
ReCogDrive	2B	40.3	44.5	41.3	40.8	30.8	26.7	–	30.1
ReCogDrive	8B	46.0	49.2	46.5	48.4	32.0	25.7	–	31.0
MiMo-Embodied	7B	48.4	56.5	49.7	53.8	32.5	30.7	–	32.2
Qwen-Drive-1.0	4B	49.5	47.6	47.9	56.4	44.2	44.6	–	44.3

Table 16: **MCQ accuracy (%) grouped by the construction probe.** Groups count how many of the three probed single frames reproduce the reference. Scores average the 21 open-source models other than the probe model. Gain is Video minus Final frame with a 95% segment bootstrap interval, and Models counts models with a positive gain.

Frames reproducing reference	n	Video	Final frame	Question only	Gain	Models
None	494	46.2	40.5	33.3	+5.7 [3.8, 7.5]	19/21
Some	199	51.0	45.2	34.8	+5.8 [3.3, 8.7]	19/21
All	385	47.2	45.3	18.9	+1.9 [0.1, 3.7]	10/21

third as wide. The experiment therefore measures the practical coverage–resolution tradeoff of this rendering protocol.

View-Coverage Results. Surround views improve the labeled multi-view subset despite small changes in overall accuracy. Qwen3-VL-8B scores 48.4% with eight cameras, 49.4% with three, and 49.5% with the front camera alone; the eight-versus-front difference is not significant ($p = 0.33$). On 510 multi-view questions, however, eight cameras improve accuracy by 8.2 points ($p = 0.003$), including 14.6 points on 226 cross-view transit templates ($p = 0.002$). Front-only input is 3.7 points better on the 1,773 front-view questions ($p < 0.001$), consistent with the resolution tradeoff. Qwen3-VL-32B and InternVL3.5-8B also gain on multi-view questions, by 10.0 and 6.7 points, respectively, without significant changes on front-view questions or overall. These results support the labels at the subset level across the three tested models. The cross-view evidence dependence summarized in the main text therefore refers to the labeled multi-view subset rather than to overall accuracy.

Table 17: **MCQ accuracy (%) under changes in camera coverage at fixed timestamps and composite width.** Front three is evaluated only for Qwen3-VL-8B. Δ is All eight minus Front, in percentage points computed before rounding; p is the exact paired McNemar value for that comparison. Cross-view transit templates form a subset of multi-view questions. The 11 MCQs without view labels contribute only to All MCQs.

Model	Subset	n	Front	Front three	All eight	Δ	p
Qwen3-VL-8B	all MCQs	2,294	49.5	49.4	48.4	-1.0	0.33
	front-view questions	1,773	49.9	48.7	46.2	-3.7	< 0.001
	multi-view questions	510	48.2	52.4	56.5	$+8.2$	0.003
	cross-view transit templates	226	54.4	61.1	69.0	$+14.6$	0.002
Qwen3-VL-32B	all MCQs	2,294	53.1	–	54.4	$+1.3$	0.23
	front-view questions	1,773	52.8	–	51.7	-1.2	0.30
	multi-view questions	510	54.7	–	64.7	$+10.0$	< 0.001
	cross-view transit templates	226	55.3	–	70.4	$+15.0$	< 0.001
InternVL3.5-8B	all MCQs	2,294	47.3	–	47.8	$+0.5$	0.60
	front-view questions	1,773	48.1	–	46.9	-1.2	0.23
	multi-view questions	510	44.9	–	51.6	$+6.7$	0.010
	cross-view transit templates	226	47.8	–	62.4	$+14.6$	0.001

Table 18: **Sampling-rate comparison for Qwen3-VL-8B-Instruct on all 2,294 released MCQs.** Entries are accuracy (%); Overall weights questions equally. Competency counts are 1,433/283/314/264 for Perception/Prediction/Planning/Reasoning. Only sampling rate changes; observation windows, eight-camera rendering, options, and decoding settings are fixed. The 5 Hz row is the main-protocol run of Tab. 2. Bold marks column maxima.

Rate (Hz)	Perception	Prediction	Planning	Reasoning	Overall
1	44.73	**39.58**	42.36	54.92	44.94
2	46.96	38.52	42.68	57.58	46.56
5	50.45	35.69	42.68	**57.95**	48.43
10	**51.57**	38.16	**44.27**	57.20	**49.56**

Cross-View Question Analysis. For Qwen3-VL-8B, gains are largest on questions that explicitly identify camera transitions. First-view accuracy rises from 40.6% with front-only input to 61.7% with eight views ($n = 133$), while free-form multi-view accuracy rises from 57.6% to 69.5% ($n = 59$). Passing-side questions gain 7.9 points, whereas entry-side accuracy is unchanged. The latter may sometimes be recoverable from the front image itself. Future-outcome and counterfactual subsets show no observed benefit from additional views in this analysis; this does not identify which other capability limits them. Subset gains support measurable use of surround evidence, but do not prove that every labeled question is impossible from the front camera.

D Additional Evaluation Controls

Temporal Sampling Density. Denser sampling improves Qwen3-VL-8B-Instruct's MCQ accuracy mainly through Perception (Tab. 18). We vary sampling from 1 to 10 Hz over all 2,294 released MCQs while fixing the observation window, image resolution, option order, and decoding settings. Overall accuracy rises from 44.94% to 49.56%, and Perception, which supplies most of that movement, from 44.73% to 51.57%. The other competencies do not follow: Planning gains 1.9 points only at 10 Hz, Reasoning is flat from 2 Hz onward, and Prediction is highest at 1 Hz and lowest at 5 Hz, so additional frames do not help every competency. Under paired McNemar tests with Holm correction over the six rate comparisons, 10 Hz exceeds 1 Hz by 4.62 points ($p < 10^{-6}$) and 5 Hz by 1.13 points ($p = 0.029$). We retain 5 Hz for the main protocol and the matched input controls because it keeps most clips within the context budget of the evaluated models, with the remaining exceptions capped as documented in Appendix B.1; on this larger question set the 10 Hz margin over it is small but no longer inconclusive, which we record as a protocol choice rather than an optimum.

Option-Order Sensitivity. Two additional option permutations produce small aggregate changes for Qwen3-VL-8B. Reordering every MCQ with two random seeds yields 47.3% and 48.7% accuracy,

compared with 48.4% in the released order (paired McNemar $p = 0.14$ and $p = 0.76$). Neither comparison is significant, and the largest observed change is about 1.1 points. This supports limited aggregate sensitivity for this model and these two permutations; it does not establish invariance for individual questions or other models. Judge sensitivity is examined separately in Appendix B.2.

Scope of Statistical Comparisons. Reported paired tests compare responses to the same questions, with Holm correction applied to the six sampling-rate comparisons. Other reported McNemar values are question-level comparisons and should be read alongside effect sizes and subset counts. Several questions can share a source clip or segment, so these tests do not by themselves account for within-segment dependence; Appendix D.1 recomputes the principal comparisons with the source segment as the resampling unit and reports how much each conclusion depends on that choice. The analyses diagnose performance under the stated protocol and should not be interpreted as population estimates of driving risk.

D.1 Segment-Clustered Statistics for the Paired Controls

Why Clustering Matters. The paired tests reported elsewhere in this paper compare responses to the same question, which removes between-question variance but not between-question *dependence*. Several questions share a clip and many more share a source segment, so 2,294 questions carry less independent information than 2,294 independent trials: a segment that happens to favour one condition contributes a block of correlated wins that a question-level McNemar test counts as separate evidence. Question-level p-values are therefore anti-conservative. We recompute the paper's principal comparisons with the clustering made explicit (Tab. 19), using the 192 source segments as the resampling unit.

Two Cluster-Level Procedures. The cluster bootstrap resamples segments with replacement, not questions, and recomputes the paired difference over 5,000 draws; its interval widens when an effect is carried by a few segments rather than spread across them. The cluster permutation test reduces each segment to its own paired delta and then flips the signs of whole segments at random over 20,000 draws, which is the cluster-level analogue of the reported McNemar test: under the null that the condition does not matter, a segment's delta is equally likely to point either way. We report the add-one estimator, so its floor at this number of draws is 5×10^{-5}. Neither procedure re-runs a model; both are derived from the stored responses.

Results. Clustering widens the evidence without changing any conclusion in the paper. The three temporal comparisons remain significant at the segment level: video over question-only ($+17.00$ points, cluster 95% CI $[14.58, 19.51]$), video over the current frame ($+5.10$, $[3.09, 6.96]$), and the released frame order over a shuffled one ($+3.88$, $[2.55, 5.24]$), each with a cluster permutation p at the 5×10^{-5} floor. The multi-view subset result also survives for all three tested models, with permutation p values of 0.0098, 0.0006 and 0.0176 against question-level values of 0.0030, 7.2×10^{-5} and 0.0102; clustering costs roughly half an order of magnitude of significance but leaves every interval clear of zero. The comparisons the paper reports as inconclusive stay inconclusive: overall accuracy under changed camera coverage has cluster intervals spanning zero for all three models. Only the front-view penalty for Qwen3-VL-8B (-3.67 points) and the multi-view gains retain segment-level support, which is the pattern the paper describes. We therefore keep the question-level values in the main tables for comparability with prior benchmarks and treat the cluster-level values here as the conservative reading.

E Additional Qualitative Examples

Qualitative Cases. Fig. 13 illustrates three Perception failures using two proprietary and two open-source models. In (a), the reference identifies REAR_RIGHT as the white pickup's first clear camera appearance, but all four displayed models select FRONT_RIGHT, where it is prominent later in the clip. Their responses are consistent with confusing later salience with first appearance. In (b), ego-state records show deceleration from 13.2 to 4.9 m/s over two seconds; the proprietary models select "slows down overall," while both Qwen3-VL models select "roughly constant speed." In (c), the reference places the construction zone's closest approach at the end, but both Qwen3-

Table 19: **Segment-clustered recomputation of the paper's paired controls.** Accuracies are MCQ accuracy (%) for the two conditions named in each row, and Δ is their paired difference in percentage points. "McNemar" repeats the question-level exact value reported elsewhere; "Perm." is the segment-level sign-flip permutation value over 20,000 draws, whose add-one floor is 5×10^{-5}; the interval is a 5,000-draw cluster bootstrap over source segments. Segment counts are the number of distinct source segments contributing to each comparison.

Comparison	Subset	n	seg.	Δ	McNemar	Perm.	Cluster 95% CI
Temporal evidence, Qwen3-VL-8B							
Question only → video	all MCQs	2,294	192	+17.00	8×10^{-44}	5×10^{-5}	[14.58, 19.51]
Current frame → video	all MCQs	2,294	192	+5.10	5×10^{-7}	5×10^{-5}	[3.09, 6.96]
Shuffled → released order	all MCQs	2,294	192	+3.88	1×10^{-7}	5×10^{-5}	[2.55, 5.24]
Camera coverage, front only → all eight							
	all MCQs	2,294	192	−1.05	0.334	0.395	[−3.31, 1.27]
Qwen3-VL-8B	front-view	1,773	188	−3.67	0.0009	0.0016	[−5.84, −1.56]
	multi-view	510	132	+8.24	0.0030	0.0098	[2.19, 14.26]
	all MCQs	2,294	192	+1.26	0.228	0.292	[−0.97, 3.54]
Qwen3-VL-32B	front-view	1,773	188	−1.18	0.303	0.317	[−3.43, 1.06]
	multi-view	510	132	+10.00	7×10^{-5}	0.0006	[4.77, 15.53]
	all MCQs	2,294	192	+0.52	0.604	0.609	[−1.31, 2.34]
InternVL3.5-8B	front-view	1,773	188	−1.18	0.234	0.248	[−3.04, 0.70]
	multi-view	510	132	+6.67	0.0102	0.0176	[1.66, 11.66]

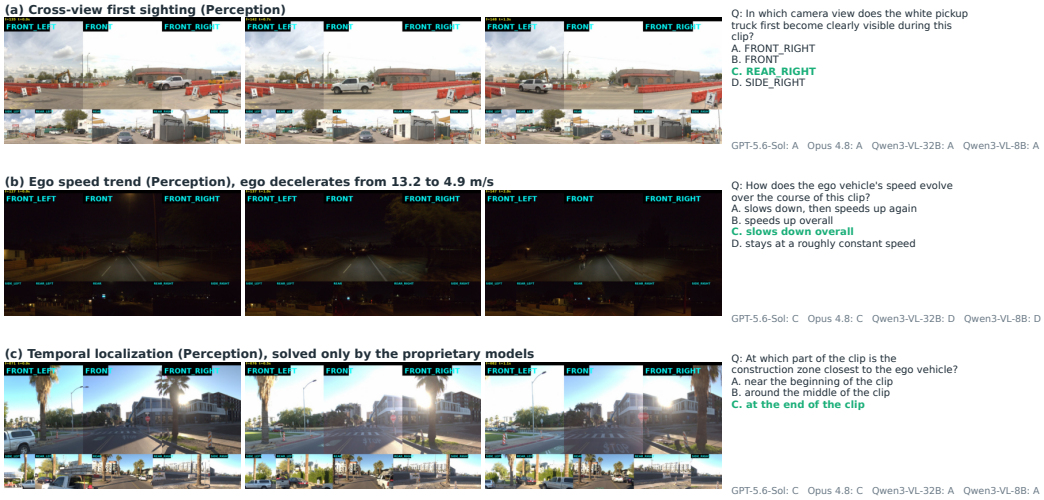


Figure 13: **Three Perception examples with the first, middle, and last sampled frames.** (a) First camera appearance of an actor. (b) Ego-speed change, with speeds from ego-state records. (c) Time of closest approach to a construction zone. Correct options are green; response lines show GPT-5.6-Sol, Claude Opus 4.8, Qwen3-VL-32B, and Qwen3-VL-8B.

VL models select the beginning. These examples illustrate camera identity, speed-change, and temporal-localization errors without claiming access to the models' internal reasoning.

F Additional Failure Cases

This appendix records how the examples in Figs. 5, 14, and 15 were chosen, and reports the full distributions behind the summary figures quoted in the main text.

Selection. Candidates are items that at most one of the eight strongest models answers correctly, which isolates errors that are shared rather than model-specific. To keep the gallery from collapsing onto a single question type, we apply a per template quota rather than taking the globally hardest items, so the released templates for camera identity, entry and passing side, temporal phase, ego speed



Figure 14: **Camera identity and approach direction: three items that none of the eight strongest models answers correctly.** Each row shows the full eight-camera composite at the first, middle and final frame of the clip. Options are coloured by role: green is the reference answer, orange the option the models selected, grey an option no model selected. In the first two rows the reference camera is `FRONT` and every model names a front-quarter camera instead; in the third the actor is ahead of the ego throughout and every model reports an approach from the left. All three are instances of the pattern quantified in Table 21: the centre of the field of view is the category that is lost, and the answers move to an immediately neighbouring view rather than to an arbitrary one. Reference answers here come from the projected actor box, so an actor near a panel boundary may be visible in two cameras while only one is named.

and heading, actor motion, range change, and open-ended Reasoning are all represented. Two further filters are practical rather than statistical. Clips whose middle frame falls below a mean luminance of $62/255$ are dropped, because a night composite prints as an unreadable block; and for ego-speed items we require the logged speed to change by at least 2.5 m/s across the clip, since a decelerating crawl makes an unconvincing illustration even when the label is correct. Within each template we prefer unanimous failures, then items whose QC record is clean, then longer clips. Every slide states the item's QC flags, so a reader can see which examples carry a caveat rather than having to trust the selection.

Camera-view errors are one panel wide. Tab. 20 decomposes every incorrect camera prediction on the 133 first-camera items by the angle between the reference camera and the chosen one. The distribution is concentrated at a single step around the rig and essentially vanishes at the opposite camera, which is the signature of a boundary effect rather than of unstructured spatial error.

Tab. 21 shows where the difficulty sits. The lateral cameras are recovered reliably; the centre camera is not, and the probability mass it loses reappears on its two neighbours. A model that confused left with right would produce a different pattern: mirror substitutions account for 3.8% (proprietary) and 3.5% (open-weight) of all camera errors, and when the actor is unambiguously to one side the side itself is recovered on 93.3% of answers for a left-hand reference and 85.9–89.7% for a right-hand one.



Figure 15: **Open-ended Reasoning: three items no model of the 25 evaluated answers correctly.** Rows show the eight-camera composite at three timestamps, the human-reviewed reference answer, and the replies of GPT-5.6-Sol and Claude Opus 4.8, truncated. In each case the reply is fluent and refers to something genuinely present in the scene – the construction barriers, the person walking ahead, the pedestrians in the crossing – but not to the interaction the reference identifies, which is a lane marking, a widening gap, and a passenger boarding respectively. The failure is therefore not a missed object but a causal attribution made to the most salient co-occurring event. Open-ended items are scored by the judge described in Sec. B.2, and the cases shown are ones where its verdict and the reference agree.

Table 20: **Share of incorrect camera predictions by angular distance from the reference camera.** Adjacent cameras are 45° apart in the eight-camera rig. Percentages are over wrong answers only.

	45°	90°	135°	180°
Proprietary (3 models, 132 wrong answers)	84.8%	9.1%	5.3%	0.8%
Open-weight (5 models, 287 wrong answers)	84.7%	10.1%	5.2%	0.0%

Frame count matters, frame order much less. Tab. 22 gives the restricted-input controls for Qwen3-VL-8B, the one model run under the complete set. On questions asking where an actor first appears, shuffling the frames costs 4.5 points, whereas reducing the clip to its final frame costs 32.3. For these questions, Qwen3-VL-8B relies mainly on seeing the actor in several observations rather than on their order.

Table 21: **Per-camera recall and selection bias on the 133 first-camera items.** The right-hand block is how often a camera is chosen minus how often it is the reference, in percentage points. FRONT is under-selected and its two neighbours absorb the difference. The side and rear rows rest on few items and are reported for completeness.

		Recall		Selected − reference (pp)	
Reference camera	Items	Prop.	Open	Prop.	Open
FRONT	48	44.3%	28.8%	−19.7	−25.5
FRONT_LEFT	29	93.8%	92.4%	+16.3	+19.9
FRONT_RIGHT	21	83.1%	77.1%	+6.2	+6.5
REAR_LEFT	11	84.4%	70.9%	+0.1	−1.8
SIDE_LEFT	8	36.4%	32.5%	−1.7	−0.8
SIDE_RIGHT	7	21.1%	54.3%	−2.6	+1.4
REAR_RIGHT	5	57.1%	52.0%	+0.3	+0.3
REAR	3	50.0%	33.3%	+1.0	0.0

Table 22: **Restricted-input controls for Qwen3-VL-8B.** Frames, their order, the surround cameras and the image itself are removed one at a time. On the first-camera subset, reordering the frames costs 4.5 points while reducing the clip to one frame costs 32.3.

Input	All MCQs (2,294)	First-camera (133)
Released order, 5 Hz	48.4%	61.7%
Shuffled frames	44.6%	57.1%
Final frame only	43.3%	29.3%
FRONT camera only	49.5%	40.6%
Question only, no image	31.4%	23.3%

Table 23: **Option distribution on the 126 temporal-localisation items**, pooled over the eight strongest models. Reference share is how often an option is the reference answer; selected share is how often a model picks it.

Option	Items	Reference share	Selected share	Difference	Recall
Near the beginning	16	12.7%	41.1%	+28.4	64.8%
Around the middle	11	8.7%	8.1%	−0.6	12.5%
At the end	90	71.4%	44.2%	−27.2	54.0%
Other phrasings	9	7.1%	0.0%	−7.1	0.0%
Unparsed reply	–	–	6.5%	–	–

A positional prior in temporal localisation. Tab. 23 gives the full option distribution for the 126 items that ask when in the clip something occurs. "Near the beginning" is selected more than three times as often as it is correct, and the middle of a clip is almost never recovered. The reference distribution is itself skewed towards the end of the clip, so a model that always answered "at the end" would score 71.4% on this slice; the over-selection of the beginning is therefore a departure from the prior in the opposite direction, not a failure to learn it.

What these examples do and do not show. Three limits apply to reading any single case. First, the reference answers do not all come from the same source: ego speed and heading are read from the vehicle's logged state and are independent of every model under test, camera identity is derived from which frustum contains the actor's box, and the remaining references are model-proposed and human-reviewed. Each slide names its source. Second, adjacent cameras overlap in field of view, so an actor near a panel boundary is genuinely visible in two of them while the reference names only one; the camera-identity cases should be read as evidence about how view identity is resolved at a boundary, not as evidence that the actor was missed. Third, a shared error is not by itself evidence about a model's internal computation, and we describe only what the answers are consistent with. The aggregate distributions above, rather than any individual case, carry the claims made in the main text.

G LIMITATIONS

LTDriveQA contains 2,889 questions over 719 clips from 192 long-tail segments of a single source dataset. Future versions can use the automatic construction pipeline to add more questions, video segments, and long-tail scenarios. VLMs also assist with event localization, question generation, and answer annotation, and their preferences may bias question style and reference wording. Human review of every released question reduces this effect but may not fully remove it.