STRIDE: EVALUATING SPATIOTEMPORAL REASONING IN DRIVING EDGE CASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating vision-language models (VLMs) for autonomous driving requires assessing their ability to connect spatial observations of past events with future consequences. Existing benchmarks either assess driving edge cases through basic perceptual questions or evaluate temporal understanding under relatively simple scenarios, leaving the spatiotemporal reasoning required in complex driving scenarios insufficiently tested. To address this gap, we introduce STRIDE (**S**patio**T**emporal **R**easoning **I**n **D**riving **E**dge Cases), an object-centric benchmark for evaluating spatiotemporal reasoning in long-tail and safety-critical scenarios. STRIDE defines six task families: spatial perception, spatial understanding, temporal memory, temporal extrapolation, trajectory prediction, and scene-context awareness. It pairs multiple-choice questions covering a broad range of spatiotemporal judgments with related free-form visual question answering (VQA) that examines the same capabilities through detailed scene explanations and justifications of driving decisions. We curate STRIDE by prioritizing scenes with unusual object behaviors and complex traffic interactions, constructing reference answers from structured dataset annotations through automated procedures, and applying limited language model assistance alongside human verification to support annotation quality. Evaluation of 12 proprietary, open-weight, and driving-specialized models reveals substantial limitations: even GPT-6-Astra achieves the highest average multiple-choice accuracy of 43.4%, compared with a random-choice baseline of approximately 20%. Our analysis further reveals that strong scene-description scores can mask weaknesses in safety-critical judgments, and models struggle more with reconstructing object motion than recalling past positions. These insights highlight STRIDE's diagnostic value for uncovering spatiotemporal driving limitations that aggregate VQA performance can conceal.

1 INTRODUCTION

Despite their emerging reasoning capabilities (Hwang et al., 2025; Chi et al., 2026), vision-language models (VLMs) still face challenges in perception, behavior prediction, and safe decision-making in autonomous driving (Schwarting et al., 2018; Li et al., 2024; Chen et al., 2025; Xie et al., 2025; Zhao et al., 2026). These limitations are particularly consequential in rare and safety-critical scenarios, which are underrepresented in driving data dominated by routine situations (Xu et al., 2026). Navigating such scenarios requires integrating past observations with current spatial relationships to anticipate how evolving traffic interactions may create hazards. This motivates a systematic evaluation of VLMs' spatiotemporal reasoning under these challenging scenarios.

However, existing benchmarks do not adequately assess whether VLMs can integrate spatial observations with past events to anticipate future consequences in challenging driving scenarios. Edge-case benchmarks often emphasize recognition questions that require limited reasoning about motion history or future interactions (Chen et al., 2025; Chi et al., 2026), while temporal reasoning benchmarks typically focus on comparatively simple scenarios, such as stationary vehicles and constant-speed straight driving (Cannons et al., 2025; Fruhwirth-Reisinger et al., 2026). As a result, VLMs' ability to reason jointly about spatial and temporal relationships in scenarios involving unusual behaviors and complex interactions remains insufficiently tested. These gaps motivate a unified benchmark that evaluates spatiotemporal reasoning in such scenarios.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

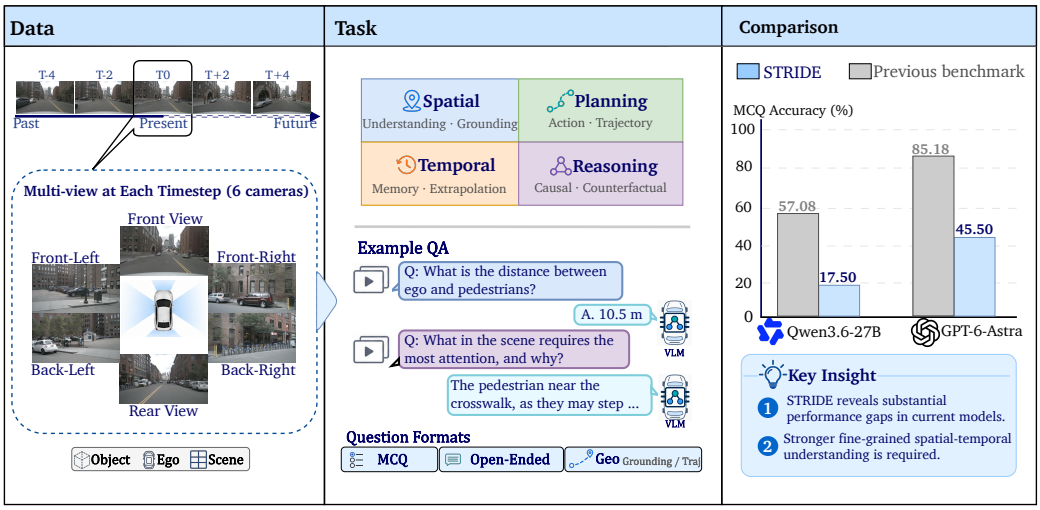


Figure 1: Overview of STRIDE. Built from multi-view temporal driving data, STRIDE evaluates spatiotemporal capabilities at the object and scene levels through complementary tasks. Qwen3.6-27B and GPT-6-Astra achieve substantially lower multiple-choice questions (MCQ) accuracy on STRIDE than on the previous benchmark (Fruhwirth-Reisinger et al., 2026).

To address this gap, we introduce STRIDE (Figure 1), an object-centric benchmark for evaluating spatiotemporal reasoning in long-tail and safety-critical driving scenarios. We prioritize scenes with unusual object behaviors and complex traffic interactions, requiring models to locate nearby objects, reconstruct their past motion, and anticipate how their future behavior may affect the ego vehicle. To support fine-grained diagnosis of spatiotemporal reasoning, the benchmark comprises 39 manually designed question templates organized into six complementary task families: spatial perception, spatial understanding, temporal memory, temporal extrapolation, trajectory prediction, and scene-context awareness. Each evaluation combines object-centric temporal context with multiple-choice questions (MCQs) that test intermediate judgments and open-ended questions (OEQs) that assess their integration into driving explanations and decisions.

Our evaluation of 12 proprietary, open-weight, and driving-specialized models highlights the difficulty of STRIDE. Qwen3.6-27B (Qwen Team, 2026) and GPT-6-Astra both achieve substantially lower MCQ accuracy than on the previous temporal reasoning benchmark (Fruhwirth-Reisinger et al., 2026) (Figure 1). Fine-grained analysis reveals difficulty connecting observations across frames into a coherent account of object dynamics: GPT-5.5 achieves 72% accuracy for previous relative position but only 18% for recent motion patterns. Our evaluation also reveals gaps that aggregate scores can conceal. High scenario explanation VQA scores coexist with weak judgments about drivable space and specific risks, suggesting that broad scene caption tasks may partly rely on generic driving knowledge. This concern echoes findings that models retain substantial performance even when visual inputs are replaced with black images (Xie et al., 2025). These findings highlight STRIDE's diagnostic value and motivate evaluating driving judgments, temporal integration, and geometric prediction alongside verbal explanations.

2 Related Work

Driving VQA Benchmarks Driving vision-language benchmarks evaluate scene understanding through questions about road entities, vehicle behavior, and driving context (Caesar et al., 2020; Marcu et al., 2024). Beyond individual questions, subsequent efforts connect perception with prediction and planning through graph-structured QA, annotated reasoning chains, and counterfactual supervision (Sima et al., 2024; Nie et al., 2024; Wang et al., 2025). Spatially grounded evaluation has also expanded toward multi-view understanding and explicit 3D localization (Wang et al., 2025; Zhao et al., 2026). However, reliability studies show that favorable QA or planning scores can reflect textual shortcuts or reliance on ego-state information, motivating closer examination of the capabilities underlying model performance (Li et al., 2024; Xie et al., 2025).

Table 1: Comparison of representative autonomous-driving vision-language benchmarks. ✅: explicitly evaluated; ❌: no dedicated evaluation reported. "Mixed" indicates dataset-dependent camera configurations. ☰: multiple-choice or fixed-vocabulary short-answer QA; ✎: open-ended QA. A capability is marked only when explicitly evaluated: Spatial concerns where objects are and how they relate, Temporal concerns reasoning across frames or about what happens next, Edge Cases concerns unusual or safety-critical situations, Grounding requires explicit localization, and Ego Traj. requires future ego-trajectory coordinates.

Benchmark	Input		QA Format	Capabilities				
	Format	View		Spatial	Temporal	Edge Cases	Grounding	Ego Traj.
NuScenes-QA (Qian et al., 2024)	Image / Video	Surround	☰	✅	❌	❌	❌	❌
DriveLM-nuScenes (Sima et al., 2024)	Image	Surround	☰ ✎	✅	✅	❌	✅	✅
Reason2Drive (Nie et al., 2024)	Video	Mixed	✎	✅	✅	❌	✅	✅
LingoQA (Marcu et al., 2024)	Video	Front	✎	✅	✅	❌	❌	❌
OmniDrive (Wang et al., 2025)	Image	Surround	✎	✅	✅	✅	❌	✅
DriveBench (Xie et al., 2025)	Image	Surround	☰ ✎	❌	✅	✅	✅	❌
NuPlanQA-Eval (Park et al., 2025)	Image	Surround	☰	✅	❌	❌	❌	❌
STSBench (Fruhwirth-Reisinger et al., 2026)	Video	Surround	☰	✅	✅	❌	❌	❌
STRIDE	**Video**	**Mixed**	☰ ✎	✅	✅	✅	✅	✅

Spatiotemporal and Long-Tail Evaluation Recent benchmarks increasingly assess temporal understanding through multi-frame traffic interactions, action recognition, event ordering, and temporal localization (Cannons et al., 2025; Fruhwirth-Reisinger et al., 2026). In parallel, corner-case and long-tail datasets broaden evaluation beyond routine driving, covering unusual objects, unstructured roads, and challenging agent interactions through perception, driving suggestions, and trajectory tasks (Chen et al., 2025; Chi et al., 2026; Xu et al., 2026). These efforts establish complementary foundations for evaluating temporal capabilities and scenario difficulty.

3 STRIDE

The STRIDE benchmark is designed to evaluate the spatiotemporal reasoning of autonomous-driving agents in long-tail scenarios with reliable ground truth and interpretable assessment. As summarized in Table 1, STRIDE combines spatial and temporal reasoning, edge-case evaluation, explicit grounding, and ego-trajectory prediction with both multiple-choice and open-ended QA.

3.1 Task Definition

Autonomous driving is an anticipatory decision-making problem: an agent must understand the previous and current scene, predict how it may evolve, and take preventive action before developing risks lead to dangerous interactions. Such anticipation of recognizing emerging conflicts and determining when and how the ego vehicle should respond depends jointly on spatial and temporal reasoning. Object positions, distances, and available driving space establish the current conditions for action, while changes across consecutive frames reveal motion trends and support predictions of future interactions. Motivated by these requirements, we organize STRIDE around the complementary dimensions of *space* and *time*.

Space This dimension is divided into two hierarchical levels of increasing difficulty, from individual object location to multi-object relation. **Spatial perception** evaluates basic object-level properties that are directly observable from the scene, including relative position, distance, direction of motion, and lane occupancy. **Spatial understanding** builds on these observations and evaluates whether models can interpret the driving significance of these spatial relationships. Tasks include determining whether an object restricts the drivable region, affects lane-change feasibility, constrains the ego vehicle's motion, or poses a potential risk.

Time This dimension extends this representation of the current scene backward and forward in time. Taking the current spatial configuration as the reference state, **temporal memory** evaluates whether a model can reconstruct how that state developed from previous observations. This includes retrieving previously visible or disappeared objects, earlier object locations, occlusion and

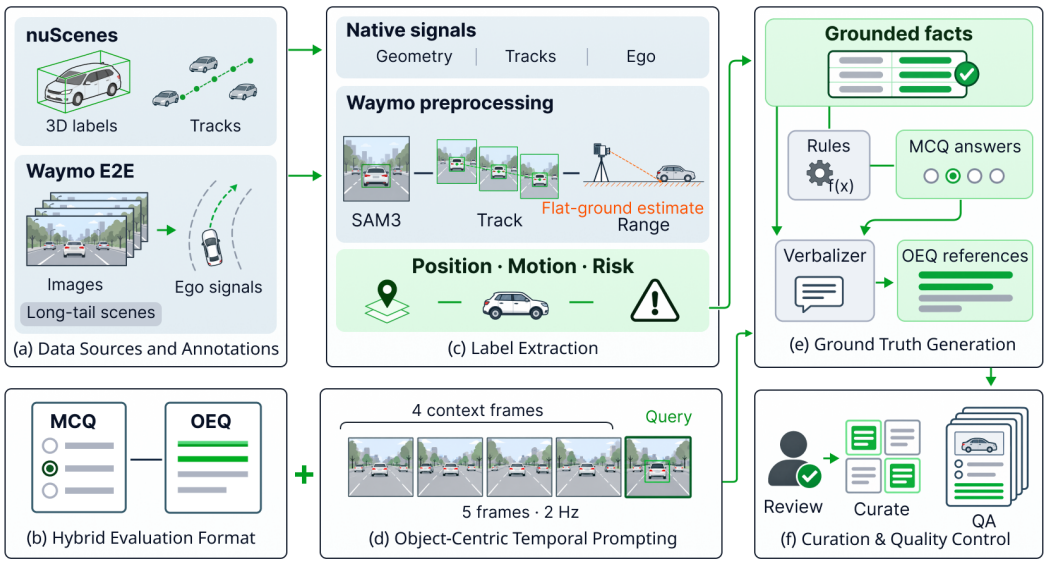


Figure 2: Data generation pipeline of STRIDE. From left to right, dataset annotations and extracted signals are converted into structured labels. Manually designed hybrid format pair these labels with five-frame context to generate linked MCQ answers and verbalized OEQ references, followed by human review and curation.

reappearance, and recent motion trends. In the opposite direction, **temporal extrapolation** evaluates whether the model can anticipate how the current scene may evolve, including future object motion, lane occupancy, collision timing, and likely subsequent events. Together, these two task families test whether a model can connect the present scene to its recent history and possible future.

Driving Context Above tasks primarily assess capabilities through questions grounded in a target object. To account for the ego vehicle's own motion and the broader environment in which object interactions occur, we therefore include the evaluation on **trajectory prediction** and **scene-context awareness**. Trajectory tasks assess the ego vehicle's past and future motion. Scene-context tasks assess environmental conditions, road structure, and surrounding activity that provide context for interpreting individual objects and their interactions.

3.2 Data Construction

Data Source and Annotation We construct STRIDE from two complementary sources: **nuScenes** (Caesar et al., 2020) and **WOD-E2E** (Xu et al., 2026). As seen in Figure 2(a), each of them provides different annotations. nuScenes is a multimodal autonomous-driving dataset comprising 1,000 driving scenes of 20 seconds each, with six-camera surround-view coverage and 3D object annotations. While providing broad coverage of everyday urban driving and rich annotations, such naturally collected datasets are dominated by nominal driving situations (Zhai et al., 2023; Li et al., 2024; Dauner et al., 2024; Tian et al., 2025). We complement nuScenes with WOD-E2E, a more recent dataset curated to emphasize challenging long-tail driving scenarios, including unusual agent interactions, cut-ins, construction zones, encounters with vulnerable road users, unexpected obstacles, and other safety-critical situations. For WOD-E2E, we supplement the released annotations with object-centric signals derived through detection, tracking, and geometric distance estimation, as detailed in Appendix A.1.

Data Sampling Strategy To keep evaluation computationally manageable while preserving scenario diversity, we construct a compact evaluation set through template-based curation. For each question template, we sample up to 50 valid instances from scenes in which the required annotations and visual evidence are available. Since the two source datasets provide different annotation types



Figure 3: **Coverage of STRIDE. (a)** Share of tagged clips containing each challenging situation in nuScenes and Waymo E2E. **(b)** Share of clips within each task family carrying the same scenario tag (SC: scene context, SP: spatial perception, SU: spatial understanding, TM: temporal memory, TE: temporal extrapolation, TRJ: trajectory). Tags are multi-label, so columns need not sum to 100%. **(c)** Future ego-maneuver composition by speed change over unique ground-truth motion patterns; bars are normalized within each maneuver.

and scenario characteristics, not every template is applicable to both datasets due to missing annotations required. To address this issue and ensure that our ground truth is trustworthy, each template is instantiated only on the dataset or datasets that provide sufficient information to support reliable ground-truth generation. Within the eligible candidates, we prioritize informative and challenging examples using a combination of complementary signals. These include scene-intrinsic factors such as traffic-participant density and interactions, weather and visibility conditions, hazardous elements, unusual agent behavior, and overall scene complexity such as involvement of vulnerable road users.

Hybrid Evaluation STRIDE combines multiple-choice and free-form VQA to assess shared spatial and temporal capabilities through complementary response formats (Figure 2(b)). Within each task family, questions are aligned around shared scene attributes, with the two formats evaluated independently. For example, the **spatial understanding** family pairs multiple-choice questions on physical constraints, available drivable space, lane-change feasibility, and risk with the free-form question: *Describe the overall spatial situation of the ego vehicle in the scene, including existing physical constraints, available drivable space, risk level, and feasible maneuvers.* The multiple-choice questions serve as intermediate judgments of individual attributes, which are then integrated in the linked open-ended question for a broader description or justification of the driving situation. This alignment supports complementary analysis of judgment accuracy and explanatory quality.

Quality Control After construction, we remove question types lacking sufficient source annotations for reliable ground truth. OEQ references are further refined using human scores and VLM assisted error analysis to correct unsupported or incomplete answers. After human review of the selected subset, we replace weak references with stronger model-assisted answers based on human scoring of the reference answer and the corresponding model response generated by GPT-5.5. More details can be found in Appendix A.6.

3.3 Dataset Statistics

Our benchmark comprises 2,350 questions spanning six task families and 1,452 distinct clips. Of these, 1,150 questions are derived from nuScenes (Caesar et al., 2020) using 23 templates, while 1,200 come from Waymo E2E (Xu et al., 2026) using 24 templates; 1,700 are multiple-choice and 650 are open-ended.

Figure 3(a) highlights the complementary corner-case coverage: nuScenes contains a larger proportion of pedestrian, intersection, special-vehicle, and construction scenarios, whereas Waymo E2E contributes relatively more foreign-object and cyclist cases. Figure 3(b) shows that these tags are not confined to particular question types. Pedestrians, intersections, and special vehicles recur across all six families, so nearly every capability is evaluated under challenging conditions. Behaviorally, the benchmark includes various future-motion patterns (Figure 3(c)). Pairing every maneuver with multiple speed profiles prevents models from inferring speed from motion direction alone.

Table 2: Results on the nuScenes and Waymo E2E splits of STRIDE. MC accuracy is reported in percentage. OEQ performance is reported with scale 0-10. Best and second-best model results are shown in **bold** and underline, respectively.

	nuScenes				Waymo E2E			
Model	**MCQ ↑**	**OEQ ↑**	**ADE ↓**	**FDE ↓**	**MCQ ↑**	**OEQ ↑**	**ADE ↓**	**FDE ↓**
RANDOM CHANCE	19.8	–	–	–	20.3	–	–	–
HUMAN PERFORMANCE	79	–	–	–	69	–	–	–
DRIVING EXPERT MODELS								
Dolphins (Ma et al., 2024)	18.5	3.24	8.16	3.81	22.2	1.86	18.43	8.60
UniAD 2.0 (Hu et al., 2023)	–	–	**0.56**	**0.79**	–	–	–	–
OPEN-WEIGHT MODELS								
Qwen3.6-27B	17.5	5.99	5.81	4.06	28.4	5.87	20.37	19.25
Qwen3-VL-8B-Thinking	23.4	5.50	8.00	3.80	32.0	4.72	22.36	19.37
Qwen3-VL-30B-A3B-Thinking	26.8	5.45	7.12	3.42	31.8	4.98	**11.83**	**6.18**
Qwen3.6-35B-A3B	31.6	5.32	7.42	5.45	34.3	5.83	18.55	8.86
PROPRIETARY MODELS								
Claude-Sonnet-5	29.4	6.42	0.87	1.30	39.2	**7.01**	13.16	10.27
GPT-5.6-Sol	38.5	5.65	0.69	0.99	37.7	6.48	13.81	12.39
Gemini-3.1-Pro	37.8	4.53	1.42	1.84	39.7	4.03	14.22	14.21
Gemini-3.6-Flash	38.9	6.50	0.69	1.01	38.8	6.20	13.53	11.37
GPT-5.5	39.3	**6.68**	0.65	0.97	38.7	7.00	13.84	12.57
GPT-6-Astra	**45.5**	6.41	0.60	0.98	**41.2**	6.77	13.64	12.03

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate STRIDE with a diverse set of VLMs spanning proprietary general-purpose models including GPT-6-Astra, Gemini-3.1-Pro, and Claude-Sonnet-5, open-weight models of different scales including Qwen3-VL-8B-Thinking, Qwen3.6-27B, and Qwen3.6-35B-A3B, and driving-specialized models including Dolphins and UniAD. This model suite allows us to compare both broad multimodal capabilities and domain-specific expertise under a unified evaluation protocol.

Implementation Details All models are evaluated using the same benchmark inputs and prompting protocol whenever their interfaces permit. Open-weight models are downloaded from `HuggingFace` and evaluated using `vLLM` (Kwon et al., 2023) on NVIDIA A6000 GPUs, while proprietary models are accessed through their official APIs. Unless otherwise stated, we use deterministic or near-deterministic decoding settings to reduce evaluation variance. More details about experimental settings can be found in Appendix C.

Evaluation Metrics. For multiple-choice tasks, we report accuracy. For open-ended questions, we report the mean score assigned by a vision-language judge on a 0–10 scale. For trajectory prediction, each predicted waypoint is compared with the ground-truth waypoint at the same timestep by Euclidean (ℓ_2) distance. We report two summaries of that distance: Average Displacement Error (ADE), the mean ℓ_2 over timesteps, and Final Displacement Error (FDE), the ℓ_2 at the final timestep. More details about scoring can be found in Appendix C.4

4.2 MAIN RESULTS

Table 2 reveals substantial limitations even in the strongest evaluated models. GPT-6-Astra achieves the highest MCQ accuracy on both nuScenes and Waymo E2E, yet reaches only 45.5% and 41.2%, respectively. Although above the random-choice baselines of 19.8% and 20.3%, these results leave more than half of the questions incorrectly answered, highlighting persistent difficulties with the benchmark's spatial and temporal judgments. For OEQ evaluation, performance also varies across response formats and tasks. GPT-5.5 leads on nuScenes with a score of 6.68, while Claude-Sonnet-5 achieves the highest score on Waymo E2E at 7.01, closely followed by GPT-5.5 at 7.00. Meanwhile, the driving expert UniAD 2.0 excels at trajectory prediction on nuScenes, achieving an ADE/FDE

Table 3: Representative per-task results used in the analysis. GPT-5.6: GPT-5.6-Sol; G-Fl.: Gemini-3.6-Flash; G-Pro: Gemini-3.1-Pro; Claude: Claude-Sonnet-5; Q35: Qwen3.6-35B-A3B; Q30-T/Q8-T: Qwen3-VL-30B-A3B-Thinking/Qwen3-VL-8B-Thinking.

(a) nuScenes	**Model**	**Representative tasks**			
Capability		**Task 1**	**Score**	**Task 2**	**Score**
Spatial perception	G-Pro	SP-3: Distance Estimate	64%	SP-7: Spatial desc.	2.10
Spatial understanding	GPT-5.5	SC-6: Scene desc.	9.22	SU-2: Drivable region	22%
Temporal memory	GPT-5.5	TM-2: Prev. position	72%	TM-3: Recent motion	18%
Temporal extrapolation	Best	TE-1: Future motion	40%	TE-6: Scene evolution	6.98

(b) WOD-E2E	**Model**	**Representative tasks**			
Capability		**Task 1**	**Score**	**Task 2**	**Score**
Scene/risk reasoning	GPT-5.5	SC-4: Scene/risk/action	8.08	SC-2: Risk source	28%
Spatial perception	G-Fl. / Claude	SP-3: Distance Estimate	62%	SP-3a: Metric distance	6.28
Temporal memory	Q30-T	TM-2: Prev. position	10%	TM-3: Recent motion	32%
Temporal extrapolation	Best	TE-1: Future motion	36%	TE-5: Next event	34%

(c) Trajectory	**Split / Task**	**Parsing Rate**	**ADE valid**	**ADE all**	**FDE all**
Q35	nuScenes / TRJ-6	16%	0.32	8.19	4.61
Dolphins	WOD-E2E / TRJ-5,6	0%	–	18.43	8.60
GPT-5.5	WOD-E2E / TRJ-5	100%	23.44	23.44	20.10

of 0.56/0.79, compared with 0.60/0.98 for GPT-6-Astra and 0.65/0.97 for GPT-5.5. In summary, these results expose a key limitation: strong aggregate VQA performance does not establish reliable spatiotemporal judgment or superior trajectory prediction. Improving the consistency of these capabilities remains an important direction for future work.

5 Analysis

We examine the models in Table 2 across four capability categories using the selected per-task results in Table 3. Complete results can be found in Appendix Tables 9 and 10. Scene and risk interpretation are discussed under Section 5.2, while future-trajectory tasks are included under Section 5.4.

5.1 Spatial Perception

Performance varies across spatial attributes and response formats. On nuScenes, the best relative-position and distance accuracies are 62% and 64%, compared with 38% for trajectory direction and 44% for occlusion state. Integrated spatial descriptions remain difficult: GPT-5.5 leads with 3.66/10, while Gemini-3.1-Pro scores 2.10/10 despite leading distance MCQs. On Waymo E2E, distance MCQ and OEQ leaders also differ: Gemini-3.6-Flash achieves 62% accuracy, whereas Claude-Sonnet-5 scores 6.28/10. Success on individual spatial attributes therefore does not consistently extend across response formats or to descriptions integrating multiple relations.

Output-format compliance further affects measured performance. For Qwen3.6-27B on nuScenes SP-3, only 34% of responses are of the valid format; accuracy is 22% overall and 64.7% among the 17 parsed responses. This suggests insufficient instruction-following, although performance on this selected subset cannot establish accuracy on the full task.

5.2 Spatial Understanding

Our benchmark further reveals that strong scene descriptions do not ensure reliable driving judgments. On nuScenes, GPT-5.5 scores 9.22/10 for scene description but only 22% for both drivable-region identification and lane-change clearance. Qwen3-VL-8B-Thinking also exceeds all proprietary models on physical constraints, reaching 42%.

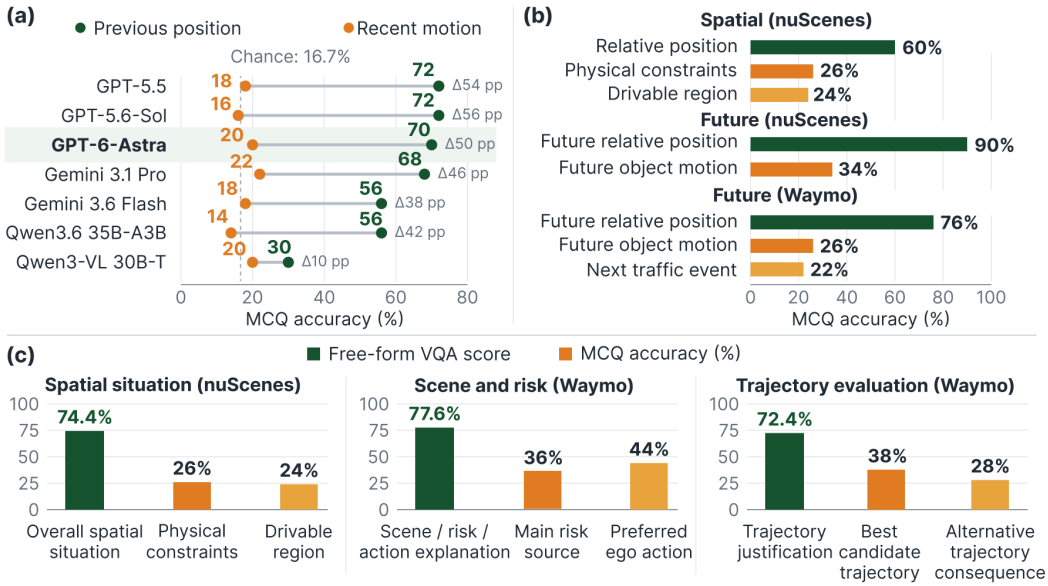


Figure 4: Analysis of GPT-6-Astra's driving capabilities. (a) Accuracy gaps between recalling previous object positions and reconstructing recent motion; Δ denotes the difference in percentage points. (b) Stronger positional judgments along with weaker spatial-constraint and future-event reasoning. (c) Favorable explanation scores with low accuracy on related driving judgments. Judge scores are multiplied by 10 for visualization on a shared 0–100 scale.

On Waymo E2E, GPT-5.5 scores 8.08/10 for scene/risk/action explanations, alongside accuracies of 44%, 28%, and 48% for scene classification, risk identification, and preferred action. Fine-grained prompts and task-specific scoring criteria in STRIDE make these comparisons more informative by separating broad explanatory quality from precise driving judgments. As a result, differences between OEQ and MCQ performance should not be interpreted as contradictory answers to equivalent questions, but rather as complementary evidence about distinct aspects of model capability.

5.3 Temporal Memory

Recalling earlier positions is easier than reconstructing motion on nuScenes. GPT-5.5 and GPT-5.6-Sol both achieve 72% for previous object position but only 18% and 16%, respectively, for recent motion, near its 16.7% chance baseline. Memory of currently occluded objects is also limited: Qwen3.6-35B-A3B leads at 36%, against a 25% baseline. These results suggest difficulty integrating observations into a consistent motion history.

Waymo E2E shows a less uniform pattern: Qwen3-VL-30B-A3B-Thinking achieves 32% for recent motion but only 10% for previous position. Past ego-maneuver classification is comparatively stronger, reaching 74% for Claude-Sonnet-5. Strong ego-motion summaries therefore do not necessarily translate into reliable object-level memory, particularly when models must maintain object identities and reconstruct motion histories across multiple frames.

5.4 Temporal Extrapolation

Anticipating behavior and its driving consequences remains difficult. On nuScenes, the best future-object-motion accuracy is 40%, and the best scene-evolution OEQ score is 6.98/10. On Waymo E2E, the best accuracies for future object motion, ego response, and the next traffic event are 36%, 34%, and 34%. Trajectory evaluation also yields different rankings: Qwen3-VL-30B-A3B-Thinking achieves the lowest Waymo ADE/FDE of 11.83/6.18 despite weaker aggregate VQA performance.

Parsing success is essential to interpreting displacement errors. On nuScenes TRJ-6, Qwen3.6-35B-A3B parses successfully on only 16% of examples; ADE is 0.32 on valid outputs but 8.19 overall. Dolphins parses 18% of outputs on each nuScenes waypoint task and none on Waymo. Its

finite Waymo FDE of 8.60 reflects the stationary-at-origin fallback for invalid outputs, rather than successful waypoint generation. Conversely, GPT-5.5 achieves 100% parsing on Waymo TRJ-5 but still incurs an ADE of 23.44. Format and trajectory accuracy therefore require joint assessment.

5.5 A Look into Failure Cases of GPT-6-Astra

Although GPT-6-Astra achieves the highest aggregate MCQ accuracy on both nuScenes and Waymo E2E split of STRIDE, its fine-grained results reveal several remaining limitations, most of which apply to all the VLMs we have tested.

Weak spatial reasoning and temporal tracking across frames GPT-6-Astra performs substantially better at identifying an object's state at a particular moment than at reconstructing how that state changes over time. As shown in Figure 4(a), it achieves 70% accuracy for previous object position (TM-2) on nuScenes, but only 20% for recent motion pattern (TM-3), a gap of 50 percentage points. This pattern extends across all seven models shown, with gaps ranging from 10 to 56 percentage points, suggesting a shared limitation in integrating observations across frames. Despite differences in positional recall, recent-motion accuracy remains consistently low at 14–22%, close to the 16.7% chance baseline. Future research could investigate memory mechanisms that explicitly maintain object identities, update their states, and retrieve relevant motion history, enabling driving agents to use past observations more reliably as a step toward practical deployment.

Figure 4(b) contrasts GPT-6-Astra's positional judgments with its performance on spatial constraints and future developments. On nuScenes, relative-position accuracy reaches 60%, compared with 26% for physical constraints and 24% for drivable regions. Models perform well on simple recognition questions but struggle with safety-critical, action-relevant judgments that require a deeper understanding of how the scene evolves over time. A similar gap appears in future reasoning: future relative-position accuracy reaches 90% on nuScenes and 76% on Waymo E2E, whereas future object-motion accuracy is only 34% and 26%, respectively. On Waymo E2E, next-traffic-event accuracy further falls to 22%.

Strong scene-description scores coexist with weak safety-critical judgments GPT-6-Astra receives favorable scores for integrated scene explanations while remaining unreliable on specific safety-critical questions, as illustrated in Figure 4(c). On nuScenes, its overall spatial-situation explanations receive an average judge score of 7.44/10, alongside the low physical-constraint and drivable-region accuracies noted above. On Waymo E2E, scene/risk/action explanations receive 7.76/10, yet MCQ accuracy is only 36% for identifying the main risk source and 44% for selecting the preferred ego action. This pattern also extends to trajectory evaluation: trajectory justifications receive 7.24/10, while best-candidate trajectory selection and alternative-trajectory consequence judgments achieve only 38% and 28% accuracy, respectively. These aggregate results suggest that evaluations centered on broad scene descriptions may present an overly optimistic picture of driving competence, obscuring weaknesses in the detailed capabilities required for safety-critical judgments.

6 Conclusion

We present STRIDE, a benchmark designed to evaluate whether vision-language models can connect past observations, current spatial relationships, and future consequences in challenging driving scenarios. By combining object-centric temporal prompting, six complementary task families, and linked MCQ–OEQ evaluation grounded in structured driving annotations, STRIDE enables fine-grained diagnosis beyond broad scene-description quality. Across 12 proprietary, open-weight, and driving-specialized models, our results show that current systems remain far from reliable: even the strongest model answers fewer than half of the multiple-choice questions correctly, while detailed analysis reveals pronounced gaps between explanation quality and safety-critical judgments, and between recalling past object positions and reconstructing motion history. Together, these findings highlight the need to evaluate driving VLMs through precise spatial and temporal judgments in addition to fluent scene-level explanations. We hope STRIDE will serve as a shared evaluation resource for the research community, supporting systematic comparison of models and diagnosis of their spatiotemporal limitations in challenging driving scenarios.

AI USE STATEMENT

In this work, we used generative AI tools during dataset construction solely to verbalize computed labels as natural-language answers. The resulting dataset was subsequently verified by human reviewers. We also used generative AI tools to create small icons for scientific figures; the overall figures were designed and edited by the authors in SVG format. Additionally, we used generative AI tools to polish the writing, and the manuscript underwent multiple rounds of proofreading and revision by the authors. Beyond these uses, we did not use generative AI for any other tasks requiring disclosure. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates vision-language models using existing driving datasets, with reference answers grounded in dataset annotations and verified by human reviewers. Despite these measures, annotation errors and biases inherited from the source datasets may remain. The benchmark is intended for research evaluation, and its results should be used responsibly, with careful consideration of these limitations and the safety requirements of real-world autonomous driving.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the data sources, preprocessing procedures, and annotation-grounded question and answer construction in Section 3. The evaluated models, inference settings, and evaluation metrics are documented in Section 4 and Appendix C. Question templates are provided in Appendix E.

REFERENCES

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Kevin Cannons, Saeed Ranjbar Alvar, Mohammad Asiful Hossain, Ahmad Rezaei, Mohsen Gholami, Alireza Heidarikhazaei, Zhou Weimin, Yong Zhang, and Mohammad Akbari. From segments to scenes: Temporal understanding in autonomous driving via vision-language model. *arXiv e-prints*, pp. arXiv–2512, 2025.

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.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *Advances in Neural Information Processing Systems*, 38, 2026.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Christian Fruhwirth-Reisinger, Dušan Malić, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. *Advances in Neural Information Processing Systems*, 38, 2026.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-end multimodal model for autonomous driving. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=kH3t5lmOU8.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14864–14873. IEEE, 2024.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Sung-Yeon Park, Can Cui, Yunsheng Ma, Ahmadreza Moradipari, Rohit Gupta, Kyungtae Han, and Ziran Wang. Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8066–8076. IEEE, 2025.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4542–4550, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):187–210, 2018.

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. Springer, 2024.

Thomas Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3656–3673. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/tian25b.html.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452. IEEE, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6285–6297. IEEE, 2025.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3709–3718, 2026.

Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023.

Zongchuang Zhao, Haoyu Fu, Dingkang Liang, Xin Zhou, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Extending large vision-language model for diverse interactive tasks in autonomous driving. *IEEE Transactions on Image Processing*, 2026.

APPENDIX

A Additional Benchmark Design **13**

A.1 WOD-E2E Preprocessing and Label Extraction 13

A.2 Question Template Design 14

A.3 Object-centric Temporal Prompting 14

A.4 Trustworthy Ground Truth Generation 15

A.5 Evaluation with LVLM Judges 15

A.6 More on Quality Control 15

A.6.1 Multiple-choice questions 15

A.6.2 Open-ended questions 16

B More Dataset Statistics **16**

C Additional Experimental Details **17**

C.1 Evaluated Models 17

C.2 Shared Evaluation Protocol 17

C.3 Model-Specific Settings 18

C.4 Scoring Notes 18

D Additional Per-task Performance Tables **19**

E Question Templates **21**

A ADDITIONAL BENCHMARK DESIGN

A.1 WOD-E2E PREPROCESSING AND LABEL EXTRACTION

For WOD-E2E, several object-centric signals required by our question templates are not directly provided by the released annotations. We therefore construct them through an additional preprocessing pipeline. Specifically, we apply SAM3 with ten road-user prompts to the front-camera images, filter detections by confidence, mask quality, and minimum area, and retain up to ten high-quality objects per frame according to confidence, proximity, and driving relevance. The resulting boxes are deduplicated and associated across consecutive frames using heuristic tracking, from which we derive relative motion, range rate, time-to-collision, relative-position sectors, lane-position bins, and occlusion proxies.

Object distance is not obtained from Waymo 3D labels. Instead, for a detected box (x_1, y_1, x_2, y_2), we take its bottom-center pixel

$$u = \frac{x_1 + x_2}{2}, \qquad v = y_2,$$

and back-project it through the calibrated front camera. The corresponding camera-frame ray is

$$\mathbf{r}_c = \begin{bmatrix} 1 \\ -\frac{u - c_x}{f_x} \\ -\frac{v - c_y}{f_y} \end{bmatrix},$$

where (f_x, f_y, c_x, c_y) are the camera intrinsic parameters. Using the camera-to-vehicle rotation $\mathbf{R}_{vc}$, we obtain $\mathbf{r} * v = \mathbf{R} * vc\mathbf{r}_c$ and intersect this ray with the assumed ground plane $z = 0$. If $\mathbf{o}_v$ denotes the camera origin in vehicle coordinates, the intersection scale and point are

$$\lambda = -\frac{o_z}{r_{v,z}}, \qquad \mathbf{p} = \mathbf{o}_v + \lambda \mathbf{r}_v,$$

from which the planar object distance is estimated as

$$d = \sqrt{p_x^2 + p_y^2}.$$

This is therefore a flat-ground ray-intersection estimate, assuming that the bottom center of the detected box corresponds approximately to the object's contact point with the road surface.

We additionally extract camera calibration and timestamps, ego past position, velocity, and acceleration, ego future trajectory and driving intent, preference trajectories, and released long-tail scenario-cluster labels directly from WOD-E2E (Figure 2(c)). These signals are combined with the derived object-level annotations to support rule-based answer construction for the Waymo split of STRIDE.

A.2 Question Template Design

Each instance contains five consecutive 360° views formed by stitching synchronized camera images and sampled at 2Hz, with four context frames followed by a query frame in which a bounding box identifies the target for object-centric questions. To support capability-level analysis, we manually design question templates aligned with the task families introduced in Section 3.1, each specifying the question wording and expected answer structure. For example, spatial-perception templates cover trajectory direction, relative position to the ego vehicle, distance, lane occupancy, occlusion state, and ground contact. Representative multiple-choice templates include *What is the spatial relationship between* `<obj>` *and the ego vehicle?* and *What is the approximate distance between* `<obj>` *and the ego vehicle?*, with discretized answer choices derived from the underlying annotations. For each five-frame temporal unit, we select a valid target object and instantiate the corresponding template with the highlighted traffic participant in the query frame. A template is applied only when the required visual evidence and structured annotations are available. The same object may support multiple complementary questions, allowing us to probe different aspects of its spatial and temporal state while maintaining a consistent visual context. Free-response templates combine related capabilities into a more comprehensive query, such as describing the target object's relative position, direction of motion, and approximate distance to the ego vehicle in a single response.

A.3 Object-centric Temporal Prompting

As illustrated in Figure 2(d), our prompting scheme is designed to reflect the inherently dynamic nature of autonomous driving. We first construct a unified 360° representation of each driving observation by stitching the synchronized multi-camera views provided by the source datasets. This gives the model a more complete view of the surrounding environment rather than restricting evaluation to a single front-facing camera. While some existing VQA benchmarks evaluate models using a single static image, such a setting is insufficient for many driving-related questions. Unlike some other embodied agent tasks, autonomous driving is fundamentally temporal: properties such as whether a nearby vehicle is moving, accelerating, approaching the ego vehicle, or changing lanes often cannot be determined reliably from an isolated frame. We therefore exploit the short driving clips provided by both source datasets and standardize their sampling rate to 2Hz. Each benchmark instance contains five consecutive observations separated by 0.5s. The first four frames provide short-term temporal context, while the fifth serves as the query frame. For object-centric questions, we visually highlight the target traffic participant with a bounding box in the query frame and ask the model to reason about that object using both its current spatial configuration and the preceding temporal context. Other questions may instead target broader scene-level properties. This formulation constrains the model's attention to a clearly defined subject while preserving the temporal evidence needed

for reasoning about motion and interaction. Compared with broad scene-level prompting, it also reduces ambiguity in both the question and its expected answer, making model behavior easier to interpret and evaluate consistently.

A.4 Trustworthy Ground Truth Generation

To provide reliable supervision, STRIDE grounds its question templates in original dataset annotations, restricting their scope to questions supported by observable or annotated (Figure 2(e)). Ground truth is derived hierarchically from geometric signals to driving semantics: annotated longitudinal and lateral offsets between the target object and ego vehicle are converted into distance and bearing, then mapped to labels such as *front*, *front-left*, and *rear-right*. Higher-level labels, such as risk, combine relative position, distance, motion direction, speed, and interaction with the ego trajectory. For example, an object directly ahead, approaching at high relative speed within a collision-relevant region, receives a higher risk label than one moving away. For free-form questions, a language model acts solely as a *verbalizer*, composing the source annotations, derived labels, and corresponding multiple-choice answers into natural-language references. This pipeline supports scalable automated construction while minimizing unsupported or hallucinated supervision and preserving traceability from each answer to the underlying driving data.

Table 4: Agreement of different automatic evaluation methods with human judgments.

Method	**Pearson**	**Spearman**	**Mean $\lvert\Delta\rvert$**
BLEU	-0.290	-0.120	7.573
METEOR	0.098	0.356	5.580
BLEURT	0.193	0.195	—
Qwen3-VL-32B (text-only)	0.780	0.804	1.320
GPT-5.5 (vision)	0.870	0.781	0.880
Qwen3-VL-32B (vision)	0.788	0.845	0.900

A.5 Evaluation with LVLM Judges

Traditional machine translation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee & Lavie, 2005) primarily measure lexical or surface-level similarity and are therefore poorly suited for evaluating open-ended driving reasoning, where semantically equivalent answers may differ substantially in wording (Nie et al., 2024). We instead employ a scalable model-based judge for free-response evaluation, providing the judge with the visual observations, question, reference answer, and model response so that correctness can be assessed in the context of the underlying driving scene.

We note that the choice of judge itself requires validation. For example, Chen et al. (2025) report that a text-only LLM judge achieved stronger human alignment than the vision-language judge evaluated in CODA-LM. Rather than assuming that multimodal judging is inherently superior, we compare conventional NLP metrics, text-only LLM judges, and LVLM judges against human annotations on a held-out subset of STRIDE. As shown in Table 4, contemporary LVLM judges achieve strong agreement with human evaluation while retaining direct access to the visual evidence required by our spatial and temporal questions. We evaluate both proprietary and open-weight judges and adopt `Qwen3-VL-32B` as our default judge due to its competitive human alignment and greater reproducibility and accessibility.

A.6 More on Quality Control

A.6.1 Multiple-choice questions

Each multiple-choice template in the pre-review set of STRIDEwas represented by a sample of 50 items. Across all 31 MCQ templates, this review set contained 1,550 questions. A template was retained only when the sample had a well-posed question, a usable option set, and ground truth that could be supported from the frames. Sixteen templates passed and fifteen failed, a pass rate of $16/31 = 51.6\%$. Because every template contributes the same number of items, the same rate

applies to the questions themselves: $800/1{,}550$ (51.6%) were kept and $750/1{,}550$ (48.4%) were removed. A failed template was removed in full, rather than edited item by item. Table 5 groups the removed templates by the reason recorded in the review.

Table 5: Multiple-choice templates removed after manual review of the 50-item sample. Each template contains 50 questions.

Template	Question	Reason for removal
SC-1	Weather and environmental condition	No suitable choice for dark scenes.
SC-2	Time of day	Samples are mostly daytime, so the label is not reliable.
SC-3	Road type and driving environment	Parked cars do not imply a parking lot, and the sample contains no highway.
TE-3	Future lane of the target object	The frames do not provide enough lane information.
TE-4	Ego response to the target object	Ground truth is incorrect.
TE-5	Next traffic event	Ground truth is incorrect and the question requires lane identification.
SC-4	Traffic density	
SC-5	Baseline scene risk	
SP-4	Lane of the target object	
SP-6	Whether the target is on the drivable plane	
SU-3	Risk posed by the target object	Rejected in manual review for insufficient/unsatisfactory evidence.
SU-5	Geometrically feasible maneuver	
SU-6	Spatial density around the ego vehicle	
TE-2	Time to collision under constant motion	
TM-1	Objects that are no longer visible	

A.6.2 Open-ended questions

The same template review removed one open-ended template, TM-6 (50 questions), because its descriptions of recent motion were inaccurate. Five templates remained: SC-6, SP-7, SU-7, TE-6, and TRJ-7, for 250 questions in total. Their original labels had been derived from rules or from multiple-choice annotations. nuScenes does not provide weather or lane labels, and comparison with the frames showed that these labels were often wrong.

We then rewrote the labels in two stages. First, we drew a stratified sample of 50 questions, 10 from each template (20% of the open-ended set), and scored each GPT-5.5 answer by hand on a 0–10 scale against the images (mean 7.74). The score determined the rewrite. On $19/50$ items (38%) the human score was 10, and the model answer replaced the ground truth unchanged. On $17/50$ (34%) the score was 7–9: a vision-language model (Qwen3-VL-32B-Instruct) corrected the model answer using the human note and the frames, and that correction became the new ground truth. On the remaining $14/50$ (28%) the score was below 7, so the same model revised the original ground truth from the note and the frames, without starting from the model answer.

Second, the other 200 questions were audited by the same vision-language model. It marked $189/200$ (94.5%) inconsistent with the images and proposed a replacement, which we accepted, and marked $11/200$ (5.5%) acceptable. Altogether, $239/250$ open-ended labels (95.6%) were rewritten. The most frequent audit errors were wrong object identity, unsafe driving claims, incorrect distance, incorrect risk, and wrong weather or lighting.

B More Dataset Statistics

Here we provide a table showing unique scene counts of STRIDE.

Table 6: nuScenes question and scene counts by task family. Scene counts use `scene_id`. Scenes are reused across question IDs, so family scene counts do not sum to the split total (310 scenes appear in more than one family).

Task family	Questions	MCQ	OEQ	Scenes	Question IDs
Scene context	50	0	50	50	SC-6
Spatial perception	250	200	50	242	SP-1, SP-2, SP-3, SP-5, SP-7
Spatial understanding	200	150	50	186	SU-1, SU-2, SU-4, SU-7
Temporal memory	200	200	0	187	TM-2, TM-3, TM-4, TM-5
Temporal extrapolation	100	50	50	100	TE-1, TE-6
Trajectory	350	200	150	316	TRJ-1–TRJ-7
Split total	1150	800	350	761	23 IDs

Table 7: Waymo question and scene counts by task family. Scene counts use `scene_id`. The split also contains 691 unique bundles (`scene_id` plus `group_id`). Family scene counts do not sum to the split total (183 scenes appear in more than one family).

Task family	Questions	MCQ	OEQ	Scenes	Question IDs
Scene context	200	150	50	117	SC-1, SC-2, SC-3, SC-4
Spatial perception	150	100	50	108	SP-2, SP-3, SP-3a
Temporal memory	150	150	0	109	TM-2, TM-3, TM-5
Temporal extrapolation	200	200	0	158	TE-1, TE-2, TE-4, TE-5
Trajectory	500	300	200	135	TRJ-1–TRJ-10
Split total	1200	900	300	335	24 IDs

C Additional Experimental Details

This appendix records the evaluated models, decoding settings, and model-specific exceptions used in the main experiments. The shared question template is described separately in Appendix E.

C.1 Evaluated Models

Table 8 lists every model reported in the main results. Proprietary models are accessed through official APIs (OpenAI Chat Completions for GPT-5.5; Vertex AI for Gemini and Claude). Open-weight models are loaded from Hugging Face and served with vLLM (Kwon et al., 2023) on NVIDIA A6000 GPUs, except Dolphins (official OpenFlamingo inference) and UniAD (end-to-end planner, not a VLM).

Unless noted, we use deterministic or near-deterministic decoding. For models with an explicit reasoning/thinking control, we use a *low* setting so that hidden chain-of-thought does not dominate latency or empty the visible-token budget, while still allowing the model to think. Claude is the exception: `effort=low` produced little or no reasoning in our setup, so we use `medium`.

C.2 Shared Evaluation Protocol

All VLMs receive the same visual inputs and the same task instruction whenever the interface allows. Non-SC-6 items use five consecutive 360° multi-camera grid frames; the fifth (query) frame highlights the target with a red box when an object is specified. SC-6 uses up to 40 chronological scene grids. Grid images are resized to at most 640×640 pixels, except SC-6 frames, which are resized to at most 360×360 to keep the long clip in context.

Table 8: Evaluated models. Reasoning knobs are API- or family-specific; "–" means none. UniAD is scored only on TRJ-5.

Model	Identifier	Reasoning / thinking
Proprietary LVLMs		
GPT-5.5	`gpt-5.5`	`reasoning_effort=low`
GPT-5.6	`gpt-5.6`	`reasoning_effort=low`
GPT-6-Astra	`gpt-6-Astra`	`reasoning_effort=low`
Gemini-3.6-Flash	`gemini-3.6-flash`	`thinking_level=low`
Gemini-3.1-Pro	`gemini-3.1-pro-preview`	`thinking_level=low`
Claude-Sonnet-5	`claude-sonnet-5`	`effort=medium`
Open-weight VLMs		
Qwen3.6-35B-A3B	`Qwen3.6-35B-A3B`	thinking off
Qwen3-VL-30B-A3B-Thinking	`Qwen3-VL-30B-A3B-Thinking`	native; after `</think>`
Qwen3-VL-8B-Thinking	`Qwen3-VL-8B-Thinking`	native; after `</think>`
Qwen3.6-27B	`Qwen3.6-27B`	thinking off
Driving expert models		
Dolphins	`gray311/Dolphins`	–
UniAD 2.0	`UniAD v2.0 (base e2e)`	n/a (TRJ-5 only)

C.3 Model-Specific Settings

Proprietary APIs GPT family is queried with `image_detail=low` and `max_completion_tokens` $= 4096$ for both MCQ and OEQ. The large completion budget is required because GPT-5.x spends hidden tokens on reasoning; a short MCQ cap often returns an empty visible answer. Gemini and Claude images are JPEG-encoded at quality 85 and resized so the long side is at most 768 pixels, matching the approximate token budget of OpenAI `image_detail=low`. Gemini uses `thinking_level=low`; Claude uses `effort=medium` via Vertex `output_config`. Proprietary SC-6 inputs use all 40 frames.

Open-weight Qwen models Qwen3.6-27B and Qwen3.6-35B-A3B are run with chat-template thinking *off* (`enable_thinking=false`) so they remain non-reasoning open-weight baselines rather than a second thinking family. For these models and Qwen2.5-VL-3B, SC-6 is evenly subsampled to 8 frames to fit the 8k context window.

Thinking variants Qwen3-VL Thinking models keep native chain-of-thought. We raise both MCQ and OEQ budgets to 4096 tokens and score only the span after the last `</think>` tag. The same strip is applied before OEQ judging and trajectory parsing.

Dolphins We follow the official inference API: frames are stacked as a video tensor and wrapped as `USER: <image> is a driving video. {instruction} GPT:<answer>`. Decoding uses beam search with 3 beams, `do_sample=false`, `no_repeat_ngram_size=3`, and at most 64 / 512 new tokens for MCQ / OEQ. SC-6 is capped at 16 frames (the official demo length).

UniAD UniAD is not a VLM and is evaluated only on TRJ-5. We take the official base end-to-end planner, convert lidar-frame waypoints to ego coordinates (flip x as in UniAD's planning metric, then apply `lidar2ego`), and keep the first 5 of the 6 predicted 0.5s steps to match the miniset ground truth.

C.4 Scoring Notes

The reported MCQ accuracy is a micro-average over items. Let $\mathcal{Q}$ be the set of multiple-choice items in a split, and let $c_i \in \{0, 1\}$ indicate whether item i is answered correctly. Then

$$\text{Acc} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} c_i = \frac{\#\{\text{correct MCQ items}\}}{\#\{\text{MCQ items}\}}. \tag{1}$$

The denominator is 800 on nuScenes v6 and 900 on Waymo v1. Free-response items are excluded, including all of nuScenes scene context (SC-6 is OEQ only).

Every MCQ question ID contains 50 items, so equation 1 equals the unweighted mean of the per-question-ID accuracies. It is not the unweighted mean of the per-family accuracies: families contain different numbers of MCQ items, and a family with more items contributes more to Acc. Family-level accuracy uses the same pooling, restricted to that family's MCQ items. For GPT-5.5, the pooled rate is 0.3925 on nuScenes (314/800) and 0.3867 on Waymo (348/900). Averaging the five family accuracies equally would give 0.375 and 0.378.

An MCQ response that cannot be parsed to an option letter is scored incorrect ($c_i = 0$) and remains in the denominator. The random-guess baseline reported alongside accuracy is the mean of $1/|\mathrm{choices}_i|$ over the same items (0.197991 on nuScenes v6 and 0.203241 on Waymo v1). It is not subtracted inside Acc.

Open-ended OEQs (SC-6, SP-7, SU-7, TE-6, TM-6, TRJ-7) are scored with a vision LLM judge (`Qwen/Qwen3-VL-32B-Instruct`, temperature 0); the judge sees the same frames as the candidate, with SC-6 subsampled to 8 frames. Waypoint OEQs (TRJ-5/TRJ-6) are parsed as JSON $[x, y]$ lists and scored with ADE-style L2 metrics. Unparseable MCQ or trajectory outputs are counted as incorrect / failed, not dropped.

D Additional Per-task Performance Tables

The results of fine-grained scores across models in each specific task are in Table 9 and Table 10.

Table 9: Fine-grained results on the nuScenes portion of STRIDE. MCQ accuracy is reported in percentage. OEQ responses are evaluated by the LVLM judge. For ADE, lower is better; for all other metrics, higher is better. Best and second-best results in each row are shown in **bold** and underline, respectively.

			Proprietary						Open-weight					Driving Expert
ID	Task	Metric	GPT-5.5	GPT-5.6	GPT-6	G-Fl.	G-Pro	Claude	Q35	Q2.5	Q30-T	Q8-T	Q27	Dolphins
Scene Context														
SC-6	Environment description	Judge ↑	**9.22**	8.94	8.94	8.50	7.22	8.94	8.84	7.16	9.14	9.10	8.94	4.58
Space Perception														
SP-1	Trajectory direction	Acc. (%) ↑	24.0	32.0	**38.0**	36.0	24.0	16.0	14.0	20.0	20.0	18.0	6.0	26.0
SP-2	Relative position	Acc. (%) ↑	56.0	42.0	60.0	**62.0**	54.0	40.0	46.0	32.0	40.0	34.0	52.0	10.0
SP-3	Object distance	Acc. (%) ↑	40.0	42.0	44.0	50.0	**64.0**	42.0	40.0	18.0	32.0	22.0	22.0	20.0
SP-5	Occlusion state	Acc. (%) ↑	36.0	42.0	**44.0**	40.0	40.0	30.0	28.0	22.0	24.0	24.0	24.0	12.0
SP-7	Spatial description	Judge ↑	**3.66**	1.96	2.94	3.22	2.10	3.42	2.74	0.78	2.16	2.28	2.86	1.12
Space Understanding														
SU-1	Physical constraints	Acc. (%) ↑	32.0	32.0	26.0	28.0	30.0	32.0	38.0	36.0	36.0	**42.0**	36.0	26.0
SU-2	Drivable region	Acc. (%) ↑	22.0	**32.0**	24.0	26.0	26.0	16.0	6.0	18.0	10.0	22.0	0.0	16.0
SU-4	Lane-change clearance	Acc. (%) ↑	22.0	30.0	**40.0**	18.0	28.0	34.0	30.0	36.0	26.0	24.0	20.0	26.0
SU-7	Overall spatial situation	Judge ↑	7.18	6.68	**7.44**	6.36	3.90	6.48	5.86	3.92	6.70	6.86	6.84	3.48
Time Extrapolation														
TE-1	Future object motion	Acc. (%) ↑	32.0	34.0	34.0	28.0	**40.0**	24.0	30.0	24.0	28.0	20.0	16.0	16.0
TE-6	Future scene evolution	Judge ↑	5.44	4.46	5.36	**6.98**	4.32	5.14	4.16	1.76	4.20	4.58	5.82	1.66
Time Memory														
TM-2	Previous object position	Acc. (%) ↑	**72.0**	**72.0**	70.0	56.0	68.0	52.0	56.0	42.0	30.0	22.0	32.0	10.0
TM-3	Recent motion pattern	Acc. (%) ↑	18.0	16.0	20.0	18.0	22.0	20.0	14.0	**30.0**	20.0	18.0	2.0	14.0
TM-4	Previously visible objects	Acc. (%) ↑	32.0	28.0	26.0	32.0	22.0	26.0	**36.0**	20.0	20.0	8.0	8.0	26.0
TM-5	Previous lane occupancy	Acc. (%) ↑	42.0	32.0	**52.0**	44.0	38.0	28.0	42.0	38.0	36.0	32.0	26.0	26.0
Trajectory Prediction														
TRJ-1	Past ego maneuver	Acc. (%) ↑	42.0	40.0	**50.0**	42.0	38.0	34.0	30.0	26.0	26.0	22.0	30.0	10.0
TRJ-2	Future ego maneuver	Acc. (%) ↑	**56.0**	44.0	54.0	38.0	32.0	28.0	32.0	22.0	28.0	26.0	2.0	18.0
TRJ-3	Future relative position	Acc. (%) ↑	52.0	46.0	**90.0**	46.0	38.0	26.0	30.0	28.0	30.0	24.0	2.0	18.0
TRJ-4	Future speed trend	Acc. (%) ↑	50.0	52.0	56.0	**58.0**	40.0	22.0	34.0	26.0	22.0	16.0	2.0	22.0
TRJ-5	Future trajectory	ADE ↓	**0.70**	0.77	0.72	0.74	1.91	0.91	6.66	–	5.51	6.08	10.13	8.59
TRJ-6	Conditioned future trajectory	ADE ↓	0.61	0.62	**0.49**	0.63	0.92	0.83	8.19	–	8.73	9.91	1.49	7.73
TRJ-7	Future path description	Judge ↑	7.90	6.20	7.36	7.42	5.10	**8.12**	5.02	4.46	5.04	4.70	5.50	5.34

GPT-5.6: GPT-5.6-Sol; GPT-6: GPT-6-Astra; G-Fl.: Gemini-3.6-Flash; G-Pro: Gemini-3.1-Pro; Q35: Qwen3.6-35B-A3B; Q2.5: Qwen2.5-VL-3B; Q30-T: Qwen3-VL-30B-A3B-Thinking; Q8-T: Qwen3-VL-8B-Thinking; Q27: Qwen3.6-27B.

Table 10: Fine-grained results on the Waymo E2E split of STRIDE. MCQ accuracy is reported in percentage. OEQ responses are evaluated by the LVLM judge. For ADE, lower is better; for all other metrics, higher is better. Best and second-best results in each row are shown in **bold** and underline, respectively.

			Proprietary						Open-weight						Driving Expert
ID	**Task**	**Metric**	**GPT-5.5**	**GPT-5.6**	**GPT-6**	**G-Fl.**	**G-Pro**	**Claude**	**Q35**	**Q2.5**	**Q30-T**	**Q8-T**	**Q8-I**	**Q27**	**Dolphins**
Scenario / Corner Case															
SC-1	Scene description	Acc. (%) ↑	44.0	46.0	**48.0**	40.0	38.0	40.0	32.0	42.0	36.0	38.0	44.0	40.0	32.0
SC-2	Main risk source	Acc. (%) ↑	28.0	24.0	**36.0**	20.0	30.0	**36.0**	22.0	24.0	26.0	34.0	26.0	30.0	20.0
SC-3	Preferred ego action	Acc. (%) ↑	**48.0**	46.0	44.0	42.0	44.0	38.0	36.0	46.0	38.0	42.0	38.0	44.0	26.0
SC-4	Scene/risk/action explanation	Judge ↑	8.08	7.80	7.76	7.46	4.22	**8.28**	7.14	4.02	6.92	6.98	7.40	7.76	2.38
Space Perception															
SP-2	Relative position	Acc. (%) ↑	32.0	36.0	34.0	40.0	34.0	32.0	**44.0**	36.0	26.0	28.0	14.0	40.0	16.0
SP-3	Object distance	Acc. (%) ↑	50.0	36.0	54.0	**62.0**	48.0	52.0	46.0	18.0	38.0	40.0	22.0	56.0	20.0
SP-3a	Distance estimation	Judge ↑	5.14	5.42	5.82	5.84	4.72	**6.28**	5.12	1.40	3.98	2.70	3.16	4.28	0.20
Time Extrapolation															
TE-1	Future object motion	Acc. (%) ↑	28.0	30.0	26.0	20.0	32.0	26.0	**36.0**	24.0	28.0	30.0	28.0	22.0	24.0
TE-2	Collision timing	Acc. (%) ↑	36.0	**42.0**	**42.0**	**42.0**	40.0	38.0	38.0	**42.0**	38.0	18.0	38.0	38.0	20.0
TE-4	Ego response	Acc. (%) ↑	28.0	**34.0**	22.0	26.0	26.0	30.0	28.0	20.0	22.0	16.0	16.0	26.0	20.0
TE-5	Next traffic event	Acc. (%) ↑	28.0	24.0	22.0	**34.0**	26.0	32.0	18.0	**34.0**	26.0	30.0	22.0	20.0	18.0
Time Memory															
TM-2	Previous object position	Acc. (%) ↑	30.0	26.0	30.0	42.0	32.0	42.0	32.0	**52.0**	10.0	8.0	10.0	34.0	20.0
TM-3	Recent motion pattern	Acc. (%) ↑	24.0	30.0	22.0	22.0	18.0	20.0	20.0	26.0	**32.0**	18.0	28.0	18.0	16.0
TM-5	Previous lane occupancy	Acc. (%) ↑	40.0	42.0	**44.0**	38.0	36.0	34.0	34.0	38.0	18.0	38.0	34.0	30.0	26.0
Trajectory Prediction															
TRJ-1	Past ego maneuver	Acc. (%) ↑	62.0	54.0	72.0	60.0	72.0	**74.0**	36.0	54.0	44.0	40.0	44.0	50.0	18.0
TRJ-2	Future ego maneuver	Acc. (%) ↑	56.0	42.0	**64.0**	54.0	60.0	54.0	46.0	52.0	54.0	52.0	54.0	24.0	18.0
TRJ-3	Future relative position	Acc. (%) ↑	62.0	60.0	**76.0**	64.0	68.0	56.0	50.0	58.0	58.0	48.0	38.0	2.0	18.0
TRJ-4	Future speed trend	Acc. (%) ↑	36.0	46.0	40.0	34.0	**48.0**	30.0	30.0	28.0	26.0	28.0	28.0	2.0	22.0
TRJ-5	Future trajectory	ADE ↓	23.44	23.26	23.04	22.59	24.01	21.85	20.00	–	**19.23**	38.69	33.58	29.47	19.83
TRJ-6	Conditioned future trajectory	ADE ↓	**4.25**	4.36	4.25	4.47	4.44	4.46	17.10	–	4.44	6.04	5.69	11.28	17.02
TRJ-7	Future path description	Judge ↑	6.22	6.22	**6.26**	5.98	4.28	5.40	5.44	3.96	5.02	4.38	4.30	6.00	3.30
TRJ-8	Best candidate trajectory	Acc. (%) ↑	34.0	30.0	38.0	26.0	34.0	**40.0**	**40.0**	38.0	30.0	38.0	28.0	6.0	**40.0**
TRJ-9	Trajectory justification	Judge ↑	**8.56**	6.46	7.24	5.52	2.88	8.08	5.60	0.90	3.98	4.80	4.98	5.42	1.54
TRJ-10	Alternative trajectory consequence	Acc. (%) ↑	30.0	30.0	28.0	32.0	28.0	32.0	30.0	**42.0**	22.0	30.0	30.0	30.0	26.0

GPT-5.6: GPT-5.6-Sol; GPT-6: GPT-6-Astra; G-Fl.: Gemini-3.6-Flash; G-Pro: Gemini-3.1-Pro; Q35: Qwen3.6-35B-A3B; Q2.5: Qwen2.5-VL-3B; Q30-T: Qwen3-VL-30B-A3B-Thinking; Q8-T: Qwen3-VL-8B-Thinking; Q8-I: Qwen3-VL-8B-Instruct; Q27: Qwen3.6-27B.

E QUESTION TEMPLATES

The benchmark taxonomy contains 39 question templates across six families. Tables 11–16 list every template. Columns **N** and **W** mark use in the reported nuScenes and Waymo evaluation splits. A blank cell means the template exists in the full catalog but is not used in that split. Waymo reuses IDs SC-1–4 for a different scenario-corner-case family and rewrites TRJ-5/6 for front-view input; those variants are recorded in Table 17.

Table 11: Scene-context templates. SC-1–5 were dropped from the reviewed nuScenes split.

ID	Fmt	Template	N	W
SC-1	MCQ	What is the primary weather and environmental condition in the current scene? *A* Clear/Sunny; *B* Rain/Wet; *C* Overcast; *D* Snow/Ice; *E* Fog/Haze		
SC-2	MCQ	Based on the lighting conditions and sky visibility, what is the estimated time of day? *A* Daytime; *B* Nighttime; *C* Dawn/Dusk		
SC-3	MCQ	Which of the following best describes the overall road typology and driving environment? *A* Urban intersection; *B* Straight urban/suburban road; *C* Highway; *D* Construction zone; *E* Parking lot/off-road		
SC-4	MCQ	How spatially dense or crowded is the current traffic environment considering all visible road users? *A* Sparse; *B* Moderate; *C* Dense		
SC-5	MCQ	What is the baseline safety risk level of the overall scene, independent of any single specific object? *A* Low; *B* Moderate; *C* High		
SC-6	OEQ	Describe the overall environment of the scene, including weather, lighting conditions, road type, and notable surrounding activity.	Y	

Table 12: Space-perception templates.

ID	Fmt	Template	N	W
SP-1	MCQ	What is the trajectory direction of <obj> in relation to the ego vehicle? *A* Toward ego; *B* Away from ego; *C* Left-to-right; *D* Right-to-left; *E* Static; *F* Same direction as ego	Y	
SP-2	MCQ	What is the spatial relationship between <obj> and ego vehicle? *A* In front; *B* Front-left; *C* Front-right; *D* Directly left; *E* Directly right; *F* Rear-left; *G* Rear-right; *H* Behind	Y	Y
SP-3	MCQ	What is the approximate distance between <obj> and ego vehicle? *A* 0–5m; *B* 5–15m; *C* 15–30m; *D* 30–50m; *E* >50m	Y	Y
SP-4	MCQ	Use the ego vehicle's lanes as reference, which lane is <obj> located in? *A* Same lane; *B* Left adjacent; *C* Right adjacent; *D* Two lanes left; *E* Two lanes right		
SP-5	MCQ	What is the occlusion state of <obj>? *A* Fully visible; *B* Partially occluded ($<$50%); *C* Mostly occluded ($\geq$50%); *D* Fully occluded but inferred	Y	
SP-6	MCQ	Is <obj> on the drivable ground plane? *A* On road; *B* On sidewalk; *C* Above the ground		
SP-7	OEQ	Describe the spatial relationship between the ego vehicle and <obj>, including its relative position, orientation of motion, and approximate distance.	Y	

Table 13: Space-understanding templates. This family is nuScenes-only.

ID	Fmt	Template	N	W
SU-1	MCQ	What physical constraints affect the ego vehicle given <obj>? *A* Blocking ego lane; *B* Partially obstructing; *C* No constraint; *D* Dynamic obstruction in ego lane	Y	
SU-2	MCQ	Which region around the ego vehicle is safely drivable? *A* Front only; *B* Front-left; *C* Front-right; *D* Left only; *E* Right only; *F* None; *G* All directions	Y	
SU-3	MCQ	What is the risk level posed by <obj> to the ego vehicle? *A* Low; *B* Moderate; *C* High; *D* Critical/imminent collision		
SU-4	MCQ	Does the ego vehicle have sufficient clearance to change lanes? *A* Safe left; *B* Safe right; *C* Unsafe both; *D* Safe both	Y	
SU-5	MCQ	Which maneuver is geometrically feasible at this moment? *A* Maintain lane; *B* Lane change left; *C* Lane change right; *D* Turn left; *E* Turn right; *F* Stop only		
SU-6	MCQ	How spatially dense is the environment around ego considering the number of objects and their distance to the ego? *A* Sparse; *B* Moderate; *C* Dense; *D* Highly congested		
SU-7	OEQ	Describe the overall spatial situation of the ego vehicle in the scene, including existing physical constraints, available drivable space, risk level, and feasible maneuvers.	Y	

Table 14: Time-extrapolation templates.

ID	Fmt	Template	N	W
TE-1	MCQ	What is the most likely future motion of <obj> over the next few seconds? *A* Continue straight; *B* Turn left; *C* Turn right; *D* Stop; *E* Cross ego path	Y	Y
TE-2	MCQ	If both vehicles maintain their current motion, when would a collision occur? *A* No collision; *B* >10s; *C* 5–10s; *D* 2–5s; *E* <2s		Y
TE-3	MCQ	Which lane will <obj> most likely occupy in the next few seconds? *A* Remain; *B* Left lane; *C* Right lane; *D* Enter ego lane; *E* Exit roadway		
TE-4	MCQ	What should be the ego vehicle's primary response to <obj>? *A* Maintain speed; *B* Slow down; *C* Prepare to stop; *D* Change lane; *E* Yield; *F* Emergency braking		Y
TE-5	MCQ	What traffic event is most likely to occur next? *A* Lane change; *B* Vehicle stopping; *C* Pedestrian crossing; *D* Vehicle merging; *E* No significant event		Y
TE-6	OEQ	Describe how the current scene is likely to evolve over the next few seconds and what actions the ego vehicle should take to maintain safety.	Y	

Table 15: Time-memory templates.

ID	Fmt	Template	N	W
TM-1	MCQ	Which objects previously appeared in the scene but are no longer visible in current frame? *A* None; *B* Pedestrian; *C* Vehicle; *D* Bicycle; *E* Static obstacle; *F* Multiple objects		
TM-2	MCQ	Where was `<obj>` located relative to the ego vehicle in the previous frame(s)? *A* In front; *B* Front-left; *C* Front-right; *D* Left; *E* Right; *F* Behind	Y	Y
TM-3	MCQ	What was the recent motion pattern of `<obj>`? *A* Moving straight; *B* Slowing down; *C* Accelerating; *D* Turning left; *E* Turning right; *F* Stopped previously	Y	Y
TM-4	MCQ	How many objects are currently occluded but were previously visible in the previous frame? *A* None; *B* 1; *C* 2–3; *D* More than 3	Y	
TM-5	MCQ	In which lane was `<obj>` located earlier in the sequence? *A* Same lane; *B* Left adjacent; *C* Right adjacent; *D* Outside roadway	Y	Y
TM-6	OEQ	Describe the recent history of the scene, including objects that appeared or disappeared, changes in their motion, and any interactions with the ego vehicle.		

Table 16: Trajectory-prediction templates. TRJ-5/6 on Waymo keep the ID but change the visual preamble (front-view, $\sim$2.0s); see Table 17.

ID	Fmt	Template	N	W
TRJ-1	MCQ	Based on the past 5 consecutive 360° multi-camera frames, what best describes the ego vehicle's recent maneuver over those past frames? *A* Go straight; *B* Slightly veer left; *C* Slightly veer right; *D* Turn left; *E* Turn right; *F* Slow to near stop	Y	Y
TRJ-2	MCQ	Based on the past 5 consecutive 360° multi-camera frames, what is the most likely ego-vehicle maneuver over the next 5 future frames (approximately 2.5 seconds)? *A–F* as TRJ-1	Y	Y
TRJ-3	MCQ	Based on the past 5 consecutive 360° multi-camera frames, at the 5th future frame (approximately 2.5 seconds ahead), which relative endpoint region best matches the ego vehicle trajectory in the current ego-centric coordinate system? *A* Front-left; *B* Front-center; *C* Front-right; *D* Left-side; *E* Right-side; *F* Near current position	Y	Y
TRJ-4	MCQ	Based on the past 5 consecutive 360° multi-camera frames, how does the ego vehicle speed trend evolve over the next 5 future frames (approximately 2.5 seconds)? *A* Accelerating; *B* Constant speed; *C* Decelerating; *D* Stopping/nearly stopped	Y	Y
TRJ-5	OEQ	Using past ego positions `<trajectory_history>` in the ego frame (current $=[0, 0]$), predict the next 5 future positions as a JSON array of 5 $[x, y]$ points.	Y	†
TRJ-6	OEQ	Same setup as TRJ-5, with known first future point `<first_future_position>`. Predict only the remaining 4 points.	Y	†
TRJ-7	OEQ	Based on the past 5 consecutive 360° multi-camera frames, briefly describe the ego vehicle's likely future path over the next 5 future frames (approximately 2.5 seconds). Mention the overall maneuver, endpoint region, and speed trend in 1–2 sentences.	Y	Y

Table 17: Templates that appear only on Waymo, or reuse a nuScenes ID with different wording. SC-1–4 here are *scenario-corner-case*, not scene-context. †TRJ-5/6 drop the 360° preamble and the pasted history, and ask for a front-view 2.0s horizon.

ID	Fmt	Waymo template
SC-1	MCQ	Which option best describes the current scene? *A* highway; *B* traffic intersection; *C* construction zone; *D* right turn
SC-2	MCQ	What is the main risk source in this scene? *A* Road debris; *B* Cut-in vehicle; *C* Crossing pedestrian/cyclist; *D* Unclear lane boundary; *E* No clear immediate risk
SC-3	MCQ	What is the preferred action for the ego vehicle? *A* Slow down; *B* Keep lane carefully; *C* Change lane; *D* Stop and wait; *E* Speed up
SC-4	OEQ	Briefly describe the scene, the main risk, and the preferred action.
SP-3a	OEQ	What is the approximate distance in meters between `<obj>` and the ego vehicle?
TRJ-5†	OEQ	Based on the past 5 observed front-view frames sampled over approximately 2.0 seconds, predict the next 5 future ego positions as a JSON array of 5 $[x, y]$ points. Current frame is $[0, 0]$.
TRJ-6†	OEQ	Same front-view setup, with the first future point given. Predict only the remaining 4 points.
TRJ-8	MCQ	Given the current scene and three candidate ego future trajectories, which trajectory is the best driving choice? *A–C* Candidate trajectory A/B/C
TRJ-9	OEQ	Explain why the selected trajectory is the best driving choice in this scene.
TRJ-10	MCQ	What is the most likely consequence if the ego vehicle follows the alternative trajectory to move or turn left, instead of the best trajectory? *A* Crash into another vehicle; *B* Crash into a VRU; *C* Enter opposing traffic; *D* Drift too far laterally