BENCHMARKING EVIDENCE SUFFICIENCY IN VISION–LANGUAGE MODELS FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable driving VLMs must do more than answer correctly or localize relevant scene actors: they must determine whether the available visual evidence is sufficient to support an answer. Existing evaluations largely treat grounding and answerability as separate measurements, making it difficult to distinguish well-supported predictions from those localized to relevant but insufficient evidence. We introduce **DriveAnswer**, a benchmark of 20,151 questions across seven reasoning families, that operationalizes this distinction using matched strong- and weak-evidence configurations. Each pair preserves the underlying scene fact and semantic answer while changing whether the observations support answering, making abstention the appropriate response under weak evidence. On the six-family evaluation, the strongest zero-shot baseline reaches 67.31% strong-evidence accuracy, yet only 21.55% of matched questions satisfy both answer-abstain behavior and complete grounding. With controlled adaptation, Gemma-4-12B reaches 63.29% on this joint capability, nearly tripling its zero-shot performance. Qwen3-VL-8B exhibits the same broad improvement in evidence-sensitive answering and grounding. These results show that semantic correctness, evidence sufficiency, and grounding are distinct capabilities that should be evaluated jointly.

1 INTRODUCTION

Vision–language models (VLMs) are increasingly used to interpret complex driving scenes and support perception, prediction, and planning. Hence, specialized benchmarks are increasingly necessary to systematically identify model vulnerabilities in such a critical domain. While recent benchmarks such as nuScenes-QA, DriveLM, and GraphVQA have substantially advanced the evaluation of driving VLMs, they primarily measure models' answer correctness rather than their ability to recognize and utilize required evidence when answering (Qian et al., 2024; Sima et al., 2024; Liang et al., 2021). This distinction is important because autonomous driving observations are constantly under partial observability: a road user may be occluded, outside the queried camera, or visible without temporal context. Under these circumstances, a model may still recover the correct label from priors or partial cues and receive QA credit. Robustness benchmarks thus show VLMs' undesirable responsiveness under weak inputs (Xie et al., 2025), while selective prediction encourages abstention under uncertainty (Srinivasan et al., 2024).

Motivated by this gap, we introduce DriveAnswer, a benchmark for evaluating whether VLMs know *when they have seen enough to answer*. Crucially, we do not treat visual grounding as a proxy for evidence sufficiency. A model may correctly localize the relevant road user or scene region while still lacking the temporal, cross-view, or relational evidence required to support the requested inference. DriveAnswer therefore evaluates whether models can identify relevant evidence, determine whether that evidence is sufficient, and answer or abstain accordingly.

Contributions. We make four contributions: (1) an evidence-sufficient driving evaluation protocol in which correctness depends not only on the semantic answer but also on whether the admitted observations support producing it; (2) a 20,151-question training and benchmark dataset spanning seven reasoning families, with matched strong-weak evidence interventions that preserve the underlying scene fact and question while changing answer sufficiency; (3) evaluation metrics that separately expose semantic answering error, unsupported commitment, unnecessary abstention, and failures to localize the required visual evidence; and (4) a controlled staged adaptation that separately studies

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.

answer sufficiency, grounding, and explicit matched-pair supervision. The adaptation results show substantial gains in evidence-sensitive answering, while also revealing that these capabilities respond differently to the corresponding supervision signals.

2 Related Work

Driving vision–language reasoning. Autonomous-driving VQA benchmarks increasingly evaluate capabilities beyond captioning and object recognition (Nie et al., 2024; Tian et al., 2025; Marcu et al., 2024). nuScenes-QA introduces large-scale scene-graph-derived driving QA spanning recognition, attributes, counting, and spatial relations (Qian et al., 2024), while DriveLM organizes perception, prediction, and planning through GraphVQA (Sima et al., 2024). More recent benchmarks broaden evaluation through fine-grained task taxonomies, multi-view and temporal reasoning, robustness, and cross-region generalization (Li et al., 2025; Vo et al., 2026; Wang et al., 2026; Xie et al., 2025; Wang et al., 2025a; Park et al., 2025; Fruhwirth-Reisinger et al., 2026; Cannons et al., 2025). Scene graphs have also been used to structure visual reasoning in general VQA (Hudson & Manning, 2019; Liang et al., 2021). In DriveAnswer, however, the graph is privileged construction information and is never exposed to evaluated models. DriveAnswer, therefore, is complementary to these efforts: rather than expanding the task taxonomy, it constructs matched versions of the same question with and without the visual evidence required to justify its answer.

Reliability, grounding, and selective prediction. Correct answers do not necessarily indicate reliable visual reasoning (Reich et al., 2023). Visually grounded VideoQA shows that strong QA performance can coexist with weak localization of the evidence supporting an answer (Xiao et al., 2024; Di & Xie, 2024), while DriveBench finds that driving VLMs may continue producing plausible responses from textual cues or general knowledge when visual inputs are degraded or absent (Xie et al., 2025). Reliable VQA and selective-prediction methods address this problem by allowing models to abstain when predictions are insufficiently reliable (Whitehead et al., 2022; Srinivasan et al., 2024). Related multimodal reliability benchmarks study complementary failures including unanswerable visual questions, abstention under insufficient evidence, and hallucination despite visual grounding (Guo et al., 2024; Madhusudhan et al., 2026; Zhu et al., 2026; Moratelli et al., 2026; Li et al., 2023; Guan et al., 2024; Geigle et al., 2024). Driving-safety benchmarks further evaluate robustness and safety-critical behavior under broader failure conditions (Li et al., 2025; Liao et al., 2025; Meng et al., 2025). DriveAnswer differs by defining abstention from *question-specific evidence availability*: the underlying question and scene fact are preserved while the observations needed to support the answer are intervened upon. Thus, a model may produce the correct semantic answer yet still fail if the available observations do not justify answering. Table 1 presents a comparison between DriveAnswer benchmark and the existing ones.

Table 1: Comparison with representative autonomous-driving VLM benchmarks. "✓" denotes explicit evaluation, "Limited" partial coverage, and "–" no primary evaluation.

Benchmark	QA	Multi-view / Temporal	Grounding	Decision	Matched Evidence / Sufficiency
nuScenes-QA (Qian et al., 2024)	✓	Limited	–	–	–
DriveLM (Sima et al., 2024)	✓	✓	Limited	✓	–
DriveBench (Xie et al., 2025)	✓	Limited	✓	✓	–
VLADBench (Li et al., 2025)	✓	✓	Limited	✓	–
DriveSpatial (Vo et al., 2026)	✓	✓	✓	–	–
ScenePilot-Bench (Wang et al., 2026)	✓	✓	Limited	✓	–
DriveAnswer (ours)	✓	✓	✓	✓	✓

3 Problem Formulation

Let S denote a driving scene, q a grounded question, and c an observation configuration. The benchmark constructor has access to privileged world state $W(S)$ from source annotations, calibrated geometry, persistent tracks, and semantic maps, but evaluated models do not. Their model-facing input is

$$X_{q,c} = \big(q, \mathcal{O}_c(S)\big), \tag{1}$$

where $\mathcal{O}_c(S)$ contains only the unboxed camera views and time steps and camera identities as well as temporal indices associated with each observation as detailed by configuration c.

Each question has a semantic reference answer $z_q^\star$ and a set of required evidence units $\mathcal{R}_q = \{e_1, \ldots, e_m\}$ (see Appendix A.3.2 for formal definitions). These units represent the visual support needed to justify the answer, such as grounding a queried road user, establishing a spatial or map relation, maintaining identity across time, or observing the temporal context required to infer motion. Under a particular configuration c, the evidence unit may either be labeled as supported, partially available, ambiguous, or unavailable to distinguish visually difficult cases and prioritize manual verification (see Appendix A.4.1 for formal definitions). Clear supported or unavailable cases can be assigned directly, while partial or ambiguous cases are routed for manual verification. These evidence-unit states are then aggregated into a configuration-level sufficiency label. A configuration is sufficient only when the admitted observations collectively support every mandatory evidence unit required by the question. If one or more mandatory units is not labeled supported, the configuration is considered insufficient.

For each question, DriveAnswer constructs a matched strong-weak pair (c^+, c^-), such that $X^+ = X_{q,c^+},\ X^- = X_{q,c^-}$. The *strong* configuration X^+ preserves the required visual evidence to justify the semantic answer, whereas the matched weak configuration X^- removes critical evidence while preserving the same underlying scene fact and question q. Consequently, even though the underlying fact is unchanged, the desired model output is different: under X^+ the model should produce $z_q^\star$, while under X^- it should select *abstention* as the answer. Thus, the benchmark separates knowing the semantic answer from whether the currently available observations justify producing it. For a matched pair (c^+, c^-), success requires both $\hat{z}_{q,c^+} = z_q^\star$ and $\hat{z}_{q,c^-} = \text{ABSTAIN}$. Additional construction checks verifying that matched strong-weak pairs have the same underlying question and world-state target are provided in Appendix A.5.1.

4 METHOD

4.1 BENCHMARK CONSTRUCTION OVERVIEW

We separately construct *scene truth*, *observation-level visual support*, and *question-specific evidence requirements* to create matched strong and weak configurations while preserving the underlying scene fact and question semantics. The construction pipeline proceeds from a query-independent scene representation to question-specific evidence configurations. For each eligible scene fact, we: (i) identify a verified world-state fact, (ii) generate a question targeting that fact, (iii) recover the semantic reference answer independently, (iv) specify the mandatory evidence units required to support the answer, (v) identify the camera observations and grounding targets that instantiate those units, (vi) determine whether each required evidence unit is supported by the admitted observations, (vii) aggregate those support states into a configuration-level sufficiency label, (viii) construct matched strong and weak configurations, and (ix) apply automatic integrity checks followed by targeted human review where needed.

4.1.1 LAYERED REPRESENTATION

DriveAnswer uses four linked representations to keep semantic truth, visual observability, and question-specific evidence requirements distinct.

Layer 1: query-independent scene representation. The first layer describes the observable driving scene independently of any particular benchmark question. It provides the scene-level entities, temporal context, camera observations, and identifiers needed to construct candidate examples without conditioning the representation on a downstream query.

Layer 2: privileged world-state graph. The second layer encodes verified entities, attributes, and relations in the underlying scene. This layer is used to establish semantic scene truth and to identify eligible facts from which questions can be constructed. Graphs are hidden privileged entities.

Layer 3: observation-level evidence candidates. The third layer links relevant world-state entities and relations to the camera observations in which they may be visually supported. It records projection and observation provenance together with cues such as visibility, truncation, occlusion, temporal

continuity, and synchronized cross-view support. These signals identify candidate evidence and difficult observation conditions; they do not determine whether a question is answerable. However, a valid projection does not establish that the queried fact is recoverable from the observation. Thus, world-state verification determines *what is true*, whereas question-specific evidence determines whether that truth can be justified from the observations supplied to the model.

Layer 4: question-specific evidence and matched configurations. The final layer specifies the mandatory evidence units required to answer a particular question. Each required unit is associated with its relevant observations and grounding targets, and each admitted configuration is assigned a support state. A configuration is sufficient only when all mandatory evidence units are supported. Strong configurations retain sufficient evidence, whereas matched weak configurations remove or withhold at least one mandatory unit while preserving the same underlying scene fact and question.

Figure 1 summarizes these four layers and their role in the construction pipeline. Additional schema and implementation details are provided in the Appendix A.3.

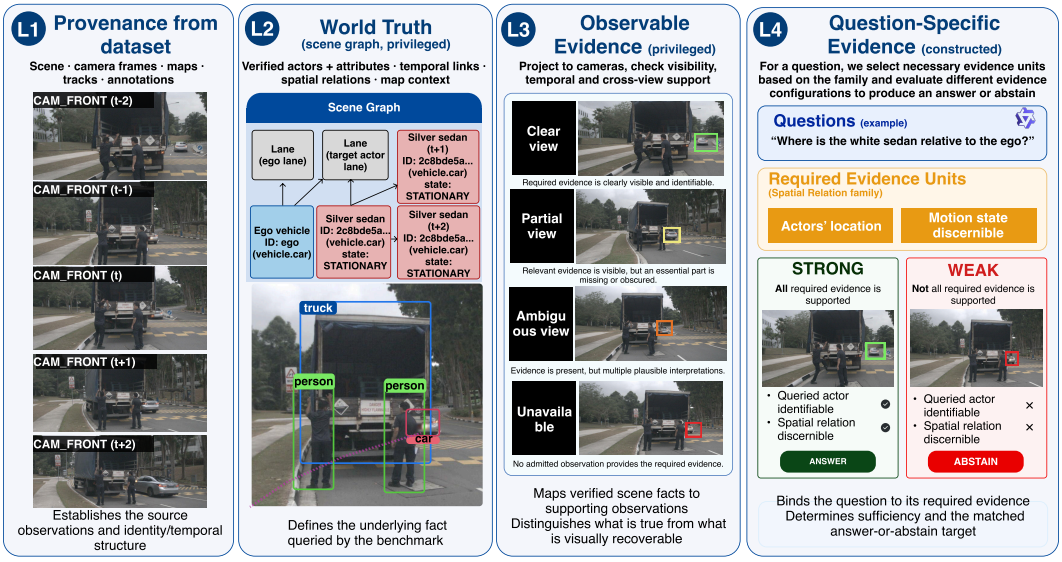


Figure 1: Overview of DriveAnswer. Layers 1–3 encode provenance, privileged world truth, and observation-level evidence. Layer 4 defines questions, required evidence, sufficiency, and answer-or-abstain target. Matched configurations vary the evidence availability while preserving the underlying scene fact and question. Models receive only the question and original unboxed observations; privileged construction records are withheld. Boxes are illustrative only. All images are original nuScenes (Caesar et al., 2020) frames, without blur, degradation, or other image quality modification.

4.1.2 Question and Evidence Construction

For each eligible world-state fact, the benchmark constructs a question targeting that fact, while the semantic reference answer is recovered independently from the verified scene representation. We employ Qwen3-VL-8B (Bai et al., 2025) to assist question generation using a predefined template. Qwen is used only to verbalize questions and referring expressions; reference answers and sufficiency labels are derived independently from verified construction records. Qwen proposes natural-language referring expressions from the construction-only target regions (e.g., object category, appearance, and coarse image position). These generated attributes are not treated as semantic ground truth, but as candidates. A candidate is retained only if the intended target is visually identifiable, the description is unique in the admitted observation, same-category distractors are accounted for, and the wording does not reveal the queried relation or reference answer. Descriptions containing unsupported or unstable attributes are rewritten or rejected. We then specify the mandatory evidence units required to justify that answer. These requirements are family specific and are defined in Appendix A.3.2. These constructions are separate from model evaluation. The benchmarked model receives the original unboxed observation; construction-only regions are never exposed at evaluation time.

DriveAnswer spans seven reasoning families. The six canonical scene-understanding families are *Spatial Relation*, *Map Relation*, *Motion State*, *Temporal Relation*, *Temporal Recovery*, and *Cross-View Recovery*. Together they test whether the model can answer questions with spatial, map, motion, temporal, and multi-view scene properties. *Evidence-Conditioned Decision* extends the same matched-evidence framework to a decision policy and is evaluated separately as a policy-conditioned extension rather than as online driving-policy evaluation. The Evidence-conditioned Decision construction details are provided in Appendix C.

4.1.3 Review and Reconciliation

Automatic Validation. Before an example is admitted to the frozen benchmark, it passes a sequence of automatic consistency checks covering the main construction joins. These checks verify that: (i) the question and reference answer correspond to the same verified world-state fact; (ii) all required evidence units are represented in the question-specific annotation; (iii) grounding targets resolve to valid admitted observations; (iv) strong configurations contain support for every mandatory evidence unit; (v) weak configurations remove support for at least one mandatory unit; (vi) the strong and weak examples preserve the same underlying question and semantic fact; and (vii) no configuration is admitted with unresolved structural or grounding inconsistencies. The audit is deliberately conservative. Failed joins, invalid grounding, or internally inconsistent support records are repaired when the underlying evidence can be verified; otherwise, the affected pair is rejected.

Human Validation. Automatic checks cannot fully resolve cases in which visual support is borderline, multiple interpretations remain plausible, or annotation sources disagree. Such cases are routed to targeted manual review. Review focuses on three questions: (1) whether the reference answer is supported by the verified scene semantics, (2) whether the designated observations actually support each mandatory evidence unit, and (3) whether the resulting strong-weak sufficiency distinction is unambiguous. When annotation components disagree, we reconcile them manually so that a semantic disagreement is resolved at the world-state level; and an observation-support disagreement is resolved by reviewing the admitted visual evidence and the family-specific evidence requirement. Cases that remain genuinely ambiguous after reconciliation are excluded to prevent annotation uncertainty, while difficult but solvable cases are retained. We additionally perform benchmark-level audits over graph construction, projection coverage, evidence reconstruction, support consistency, and sampled human validation (see Appendices A.7 and A.6).

4.2 Controlled Adaptation Study

The primary contribution of this work is the DriveAnswer benchmark. We use a controlled adaptation study with Gemma-4-12B and Qwen3-VL-8B to test how different forms of supervision address failures exposed by the benchmark. Full implementation details are provided in Appendix G.

Training setup. All reported adaptation stages start from the same fresh Gemma base rather than from the previous stage. This staged adaptation allows us to examine the effects of each component on answering and grounding capabilities of the resulting models. For our adaptation, we use LoRA with rank 8 and $\alpha = 16$ adapting the query and value projections of the attention layers, BF16, gradient checkpointing, physical batch size 1, a maximum image side of 384 pixels with ratio preserved, and AdamW. The learning rate for the supervised fine-tuning runs is 3×10^{-5}. A feasibility run fits at approximately 29 GiB of VRAM on an RTX 5090. The training and validation splits contain 14,084 and 3,076 matched questions for seven-family and 13,215 train and 2,974 validation matched questions for the main six-family test, respectively. Stages B-D objectives are applied unchanged to the Qwen adaptation.

Stage A: Zero-shot baseline. Stage A evaluates the unadapted pretrained backbone and serves as the reference point for subsequent stages. It appears as zero-shot in tables and evaluation.

Stage B: Answer + Sufficiency SFT. Each matched question contributes its strong configuration with the target $z_q^\star$ and its weak configuration with target ABSTAIN. The per-pair answer objective is the mean of the two independent cross-entropy terms,

$$\mathcal{L}_{\text{ans+suff}} = \tfrac{1}{2}\left[\mathrm{CE}\big(f(X^+), z_q^\star\big) + \mathrm{CE}\big(f(X^-), \text{ABSTAIN}\big)\right]. \tag{2}$$

The two conditions are supervised jointly as a matched training unit but optimized through independent cross-entropy terms. Stage B introduces no visual-grounding loss nor explicit cross-pair coupling, isolating evidence-sufficiency supervision.

Stage C: + Grounding supervision. Stage C retains Stage B and adds supervised grounding. Grounding tasks are averaged before being added to the objective, preventing questions with more grounding targets from receiving greater training weight. The implemented Stage C objective is

$$\mathcal{L}_{\text{ground+evid}} = \mathcal{L}_{\text{ans+suff}} + \lambda_g \mathcal{L}_{\text{ground}}, \tag{3}$$

where λ_g is 0.5 and $\mathcal{L}_{\text{ground}}$ is the mean supervised loss over the grounding tasks associated with that matched pair.

Stage D: + Explicit matched-pair objective. Stage D retains Stage C and adds a ranking loss that couples the two matched evidence conditions. The pair term encourages the semantic answer under X^+ and ABSTAIN under X^- while also encouraging the corresponding change in relative model support across the matched intervention. The complete objective is

$$\mathcal{L}_{\text{pair-aware}} = \mathcal{L}_{\text{ans+suff}} + \lambda_g \mathcal{L}_{\text{ground}} + \lambda_m \mathcal{L}_{\text{match}}, \tag{4}$$

with $\lambda_m = 0.25$ and $\lambda_g = 0.5$. This stage tests whether direct coupling across strong-weak predictions provides additional benefit. We detail the full training objective of Stage D in Appendix G.1.

5 Experiments

5.1 Evaluation Protocol

The dataset is split by scene into 739 training, 25 validation, and 38 test scenes with no overlap. The released seven-family benchmark contains 2,991 matched questions; our main evaluation and controlled adaptation study use the six scene-understanding families with 2,830 matched questions. Evidence-Conditioned Decision and the 7-family results are reported in Appendices C and G.3.

DriveAnswer reports complementary metrics designed around the matched strong–weak evidence intervention. **Strong Semantic Accuracy (Strong)** measures answer correctness when sufficient evidence is available, whereas **Weak Abstention Accuracy (Weak)** measures whether the model abstains after critical evidence is removed. Our primary evidence-sufficiency metric, **Matched-Pair Success (Pair)**, requires both behaviors for the same question. **Unsupported Commitment Rate (UCR)** measures answering under insufficient evidence, while **Unnecessary Abstention Rate (UAR)** measures abstaining despite sufficient evidence.

Answer behavior alone does not establish whether a model can identify the visual evidence supporting its prediction, highlighting the importance of grounding evaluation. A predicted region is considered correctly localized when its intersection-over-union with the reference region satisfies IoU $\geq t$. Using this threshold, we measure evidence-presence accuracy (**Presence**), the ratio of correcty grounded target at t (**IoU@t**), and joint presence-localization success (**Joint@t**). Because a question may require multiple grounding targets, **All-G@t** requires every target associated with the question to achieve Joint@t success. **Joint Answer–Grounding Success (JAGS@t)** additionally requires Pair Success together with correct grounding of all required evidence. These metrics expose complementary failure modes in semantic answering, evidence sufficiency, and grounding.

5.2 Construction Validity and Sufficiency Controls

Automated checks found no invalid references, duplicate node identifiers, broken temporal links, or inconsistencies between questions and observation configurations. Each retained strong-weak pair preserves the underlying scene fact and semantic question while changing only the available evidence. Structural consistency alone does not establish visual sufficiency, since many queried objects remain visible while the information required to justify the answer, such as motion history, temporal change, cross-view context, or relational support, is removed. Full construction statistics, audit, visibility, and shortcut audits are reported in Appendices B, D and E.

5.3 Zero-Shot Model Performance

We evaluate Gemma-4-12B (Gemma Team et al., 2026), Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025b), Molmo2-ER (Clark et al., 2026), and LLaVA-OneVision-7B (Li et al., 2024) under the same six-family held-out protocol.

Table 2: Zero-shot performance on the six-family test set at $t = 0.50$.

(a) Answer / Abstention

Model	Strong (%) ↑	Weak (%) ↑	Pair (%) ↑	UCR (%) ↓	UAR (%) ↓
Gemma-4-12B	**67.31**	**69.47**	**44.35**	**30.42**	**6.54**
Qwen3-VL-8B	54.81	49.26	17.42	50.74	11.77
InternVL3.5-8B	55.44	44.06	13.71	55.94	13.18
Molmo2-ER	53.64	41.02	14.13	58.98	17.39
LLaVA-OneVision-7B	39.47	29.33	8.62	70.64	6.64

(b) Grounding / Joint Success

Model	Presence (%) ↑	IoU@t (%) ↑	Joint@t (%) ↑	All-G@t (%) ↑	JAGS@t (%) ↑
Gemma-4-12B	**88.98**	61.92	60.43	43.18	**21.55**
Qwen3-VL-8B	84.22	**68.10**	**62.08**	**47.42**	7.49
InternVL3.5-8B	75.77	0.07	15.22	0.07	0.04
Molmo2-ER	83.97	0.35	14.57	0.04	0.00
LLaVA-OneVision-7B	75.59	0.05	15.09	0.00	0.00

Zero-shot models fail to condition answers reliably on evidence availability. Table 2 shows a substantial gap between semantic answering and evidence-sensitive behavior across backbones. Gemma-4-12B achieves the strongest overall zero-shot performance, reaching 67.31% Strong Accuracy but only 44.35% Pair Success. The gap is larger for the remaining models: Qwen3-VL-8B reaches 54.81% Strong Accuracy but only 17.42% Pair Success. Similarly, InternVL3.5-8B, Molmo2-ER, and LLaVA-OneVision-7B also obtain low Pair Success. These failures arise from different behaviors. High UCR values, ranging from 30.42% to 70.64%, reveal frequent answering after required evidence removal. Conversely, UAR values of 13.18% (InternVL3.5-8B) and 17.39% (Molmo2-ER) demonstrate abstaining when sufficient evidence is available.

Grounding and evidence-conditioned answering expose distinct failure modes. Qwen reaches 62.08% Joint@.50 and All-G@.50 of 47.42%, indicating strong target- and question-level grounding. However, this grounding performance does not translate directly into matched evidence-conditioned behavior with low Pair Success and JAGS@.50 of 7.49%. In contrast, Gemma combines lower All-G@.50 (43.18%) with substantially stronger matched-pair behavior (44.35%), yielding the highest zero-shot JAGS@.50 of 21.55%. The remaining models show that target-level presence detection alone is insufficient for complete, question-level grounding and answering since they couple high presence accuracies between 75.59% and 83.97% with near-zero All-G@.50 and JAGS@.50.

Family-level behavior. Table 3 shows that answer accuracy and grounding quality can vary independently across families. Qwen3-VL-8B, for example, grounds Motion State evidence well (All-G@.50=52.21%) but almost never satisfies the matched answer–abstention criterion for the same family (Pair=1.08%), reducing JAGS@.50 to 0.18%. Conversely, Gemma-4-12B maintains stronger coupling between grounding and evidence-conditioned answering, particularly for Spatial Relation, where it achieves 64.63% All-G@.50 and 36.87% JAGS@.50. InternVL3.5-8B, Molmo2-ER, and LLaVA-OneVision-7B often attain moderate-to-high Presence scores, yet their IoU@.50 and question-level All-G@.50 are near zero for most families, further separating evidence presence from precise localization. Temporal Recovery reveals the family-specific discrepancy between answering and grounding for Gemma: Pair Success reaches 93.10%, whereas All-G@.50 is only 20.69%. Hence, a model can exhibit the desired answering behavior without reliably recovering the full visual evidence set, or localize relevant evidence without using evidence appropriately to answer the questions. Taken together, the zero-shot results reveal distinct failure modes across architectures and reasoning families: locating relevant evidence, judging whether it is sufficient, and translating that judgment into answer-abstain behavior. This motivates our controlled adaptation study.

Table 3: Per-family zero-shot performance on the canonical six-family held-out test set at $t = 0.50$. We report semantic answering (Strong), matched evidence-sensitive answering (Pair), target-level grounding (Joint), complete evidence recovery (All-G), and their joint success (JAGS). Full per-family metrics are reported in Appendix F. **All results are expressed as percentages.**

Family	Model	Strong ↑	Pair ↑	Joint@.50 ↑	All-G@.50 ↑	JAGS@.50 ↑
Cross-view	Gemma-4-12B	54.73	53.73	14.18	2.49	1.99
	Qwen3-VL-8B	51.74	51.74	23.38	2.99	1.99
	InternVL3.5-8B	60.70	19.90	0.00	0.00	0.00
	Molmo2-ER	20.40	18.41	0.25	0.00	0.00
	LLaVA-OneVision-7B	73.13	32.34	0.00	0.00	0.00
Map	Gemma-4-12B	54.59	51.89	75.41	54.05	29.73
	Qwen3-VL-8B	43.24	25.95	78.11	57.30	15.14
	InternVL3.5-8B	39.46	7.03	48.65	0.00	0.00
	Molmo2-ER	41.08	4.32	49.19	0.54	0.00
	LLaVA-OneVision-7B	36.76	28.65	49.19	0.00	0.00
Motion	Gemma-4-12B	76.28	33.90	61.09	45.72	18.85
	Qwen3-VL-8B	83.23	1.08	62.98	52.21	0.18
	InternVL3.5-8B	82.96	2.98	0.00	0.00	0.00
	Molmo2-ER	80.25	0.54	0.18	0.00	0.00
	LLaVA-OneVision-7B	46.62	0.09	0.00	0.00	0.00
Spatial	Gemma-4-12B	69.70	51.27	80.09	64.63	36.87
	Qwen3-VL-8B	31.22	27.19	83.13	70.39	19.82
	InternVL3.5-8B	34.56	22.24	45.63	0.23	0.12
	Molmo2-ER	44.59	36.29	42.37	0.00	0.00
	LLaVA-OneVision-7B	26.96	13.71	44.97	0.00	0.00
Temp. Rec.	Gemma-4-12B	96.55	93.10	41.38	20.69	20.69
	Qwen3-VL-8B	82.76	82.76	27.59	3.45	3.45
	InternVL3.5-8B	100.00	13.79	0.00	0.00	0.00
	Molmo2-ER	62.07	44.83	10.34	0.00	0.00
	LLaVA-OneVision-7B	79.31	17.24	3.45	0.00	0.00
Temp. Rel.	Gemma-4-12B	49.09	46.35	47.37	9.82	3.65
	Qwen3-VL-8B	34.02	15.75	45.83	8.90	1.14
	InternVL3.5-8B	28.54	23.97	0.06	0.00	0.00
	Molmo2-ER	24.20	4.79	0.23	0.00	0.00
	LLaVA-OneVision-7B	29.22	0.23	0.00	0.00	0.00

5.4 Controlled Adaptation Results

Evidence-sufficiency supervision primarily changes answer–abstain behavior. Table 4 shows a large increase in matched-pair performance once models are explicitly trained to distinguish sufficient from insufficient evidence. For Gemma-4-12B, Pair Success increases from 44.35% zero-shot to 81.20% at Stage B, while UCR decreases from 30.42% to 0.42%. The corresponding Qwen3-VL-8B change is from 17.42% to 82.86% Pair Success, with UCR decreasing from 50.74% to 1.70%. Later stages change answer-side metrics comparatively little: Gemma reaches 82.05% and 82.44% Pair Success at Stages C and D, while Qwen reaches 81.70% and 82.08%, respectively.

Grounding supervision produces a distinct localization gain. For Gemma, Joint@.50 increases from 59.75% at Stage B to 84.41% at Stage C, while All-G@.50 increases from 49.58% to 73.00% and JAGS@.50 from 42.79% to 62.37%. Qwen shows the same separation more clearly: Stage B already achieves 82.86% Pair Success, but only 59.67% Joint@.50, 43.18% All-G@.50, and 39.01% JAGS@.50. Adding grounding supervision at Stage C increases these to 83.10%, 72.44%, and 62.26%, respectively. Stage D changes these aggregate grounding measures only slightly, reaching 83.11% Joint@.50, 72.72% All-G@.50, and 62.83% JAGS@.50. Thus, answer correctness, evidence sufficiency, and complete visual grounding respond differently to the adaptation objectives. Appendix G presents detailed, per-family results of the adapted models.

Table 4: Controlled adaptation study on the canonical six-family held-out test set. Stages B–D progressively introduce evidence-sufficiency, grounding, and explicit matched-pair supervision.

(a) Evidence-conditioned answering

Backbone	Stage	Strong (%) ↑	Weak (%) ↑	Pair (%) ↑	UCR (%) ↓
Gemma-4-12B	Zero-shot	67.31	69.47	44.35	30.42
Gemma-4-12B	B	81.52	**99.58**	81.20	**0.42**
Gemma-4-12B	C	82.58	99.36	82.05	0.64
Gemma-4-12B	D	82.90	99.40	82.44	0.60
Qwen3-VL-8B	Zero-shot	54.81	49.26	17.42	50.74
Qwen3-VL-8B	B	**84.17**	98.30	**82.86**	1.70
Qwen3-VL-8B	C	82.86	98.52	81.70	1.48
Qwen3-VL-8B	D	83.43	98.20	82.08	1.80

(b) Grounding and joint success

Backbone	Stage	Joint@.50 (%) ↑	All-G@.50 (%) ↑	JAGS@.50 (%) ↑
Gemma-4-12B	Zero-shot	60.43	43.18	21.55
Gemma-4-12B	B	59.75	49.58	42.79
Gemma-4-12B	C	84.41	73.00	62.37
Gemma-4-12B	D	**84.87**	**73.96**	**63.29**
Qwen3-VL-8B	Zero-shot	62.08	47.42	7.49
Qwen3-VL-8B	B	59.67	43.18	39.01
Qwen3-VL-8B	C	83.10	72.44	62.26
Qwen3-VL-8B	D	83.11	72.72	62.83

6 Discussion and Limitations

DriveAnswer exposes a gap between producing a correct semantic answer and determining whether the available observations justify that answer. Across all five zero-shot models, Pair Success is substantially lower than Strong Semantic Accuracy, showing that QA performance can remain high even when models fail to respond to missing evidence. Grounding provides a complementary diagnostic: correctly localizing relevant content does not guarantee translation to correct answer-abstain behavior. For both Gemma and Qwen, evidence-sufficiency supervision accounts for a large gain in matched-pair behavior, grounding supervision produces a large localization improvement, while explicit coupling adds only family-specific improvements. The cross-backbone comparison therefore supports the broader conclusion that semantic answering, evidence sufficiency, and grounding are related but distinct capabilities that are best understood through joint evaluation.

Limitations and Future Work. DriveAnswer is currently constructed from nuScenes and therefore inherits its sensors, geographic coverage, annotation ontology, and temporal sampling. Future work will extend the benchmark to additional domains. Our adaptation study involves two VLM architectures that exhibit a similar trend, providing initial evidence that the observed improvements are not backbone-specific, but broader evaluation across model scales and architectures would strengthen this conclusion. Finally, evidence sufficiency is inherently question-dependent. Thus, the limited number of families also confines the benchmark's scope to the predefined tasks. Further work could include more diverse tasks and edge case tasks to ensure better generalization.

7 Conclusion

We introduced DriveAnswer, a matched-evidence benchmark for evaluating reliable driving reasoning beyond answer correctness by jointly considering semantic accuracy, evidence sufficiency, and visual grounding. Its matched strong-weak construction tests whether models can answer when evidence is sufficient, abstain when it is not, and ground supported predictions in the relevant observations. Our results show that these capabilities can diverge substantially and benefit from being evaluated together. More broadly, reliable multimodal reasoning requires not only finding relevant evidence, but recognizing when that evidence is sufficient to justify an answer.

Reproducibility Statement

We provide the information needed to reproduce DriveAnswer construction, training, and evaluation throughout the paper and supplementary material. Benchmark construction, matched-evidence interventions, and split definitions are described in Section 4.1 and Appendices A-B. Training objectives and implementation details for the staged Gemma and Qwen adaptations are given in Section 4.2 and Appendix G.1. Evaluation metrics and statistical procedures, and per-family results are documented in Section 5 and Appendix D, G.2, G.3 and E.

AI Use Statement

Qwen3-VL-8B was used as part of the benchmark construction pipeline to verbalize questions and referring expressions, as described in Section 4.1.2. Its outputs were reviewed and verified as part of the pipeline, and the authors take responsibility for the final work.

References

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.

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.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12943. IEEE, 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.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2728–2742, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10284–10296, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

Yue Li, Meng Tian, Zhenyu Lin, Jiangtong Zhu, Dechang Zhu, Haiqiang Liu, Yueyi Zhang, Zhiwei Xiong, and Xinhai Zhao. Fine-grained evaluation of large vision-language models in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9431–9442. IEEE, 2025.

Weixin Liang, Yanhao Jiang, and Zixuan Liu. Graghvqa: Language-guided graph neural networks for graph-based visual question answering. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pp. 79–86, 2021.

Dacheng Liao, Mengshi Qi, Peng Shu, Zhining Zhang, Yuxin Lin, Liang Liu, and Huadong Ma. Robodrivevlm: A novel benchmark and baseline towards robust vision-language models for autonomous driving. *arXiv preprint arXiv:2512.01300*, 2025.

Nishanth Madhusudhan, Vikas Yadav, and Alexandre Lacoste. Knowing when not to answer: Evaluating abstention in multimodal reasoning systems. *arXiv preprint arXiv:2604.14799*, 2026.

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.

Xianhui Meng, Yuchen Zhang, Zhijian Huang, Zheng Lu, Ziling Ji, Yaoyao Yin, Hongyuan Zhang, Guangfeng Jiang, Yandan Lin, Long Chen, et al. Is your vlm for autonomous driving safety-ready? a comprehensive benchmark for evaluating external and in-cabin risks. *arXiv e-prints*, pp. arXiv–2511, 2025.

Nicholas Moratelli, Christopher Davis, Leonardo FR Ribeiro, Bill Byrne, and Gonzalo Iglesias. Benchmarking deflection and hallucination in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28348–28370, 2026.

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.

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 multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3129–3144, 2023.

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.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.767.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4626–4635. IEEE, 2025.

Hao Vo, Khoa Vo, Phu Loc Nguyen, Sieu Tran, Duc Minh Nguyen, Ngo Xuan Cuong, Gladys Gawugah, Sreevenkata Anjani Tishita Godavarthi, Chase Rainwater, Nghi D. Q. Bui, Anh Nguyen, Duy Minh Ho Nguyen, and Ngan Le. DRIVESPATIAL: A benchmark for spatiotemporal intelligence in vlms for autonomous driving. *arXiv preprint arXiv:2605.23176*, 2026.

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, 2025a.

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.

Yujin Wang, Yutong Zheng, Wenxian Fan, Tianyi Wang, Hongqing Chu, Daxin Tian, Bingzhao Gao, Jianqiang Wang, and Hong Chen. ScenePilot-Bench: A large-scale dataset and benchmark for evaluation of vision-language models in autonomous driving. *arXiv preprint arXiv:2601.19582*, 2026.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2022.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214. IEEE, 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6585–6597. IEEE, 2025.

Yanxu Zhu, Shitong Duan, Xiangxu Zhang, Jitao Sang, Peng Zhang, Tun Lu, Xiao Zhou, Jing Yao, Xiaoyuan Yi, and Xing Xie. Mohobench: Assessing honesty of multimodal large language models via unanswerable visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29205–29213, 2026.

A Benchmark Construction and Annotation Protocol

A.1 Frozen benchmark population and split

DriveAnswer contains 20,151 unique grounded questions drawn from 802 scenes and divided into seven reasoning families. Scene-level splitting prevents questions derived from the same underlying scene from appearing in more than one split. The frozen split contains 739 training scenes, 25 validation scenes, and 38 held-out test scenes. No question identifier maps to more than one scene.

Each question is evaluated under two matched observation conditions. The strong configuration preserves the observations required to justify the verified semantic reference answer. The weak configuration changes only the available evidence while preserving the underlying scene fact, queried

entities, and question; its intended behavior is therefore ABSTAIN. Across the frozen manifest this yields 40,302 condition instances for 20,151 unique questions.

The seven question families contain 6,098 spatial-relation questions, 1,451 map-relation questions, 6,676 motion-state questions, 2,354 temporal-relation questions, 380 temporal-recovery questions, 2,060 cross-view-recovery questions, and 1,132 evidence-conditioned decision questions.

A.2 END-TO-END CONSTRUCTION FLOW

Each DriveAnswer example is constructed through the same sequence: (1) identify an eligible verified world-state fact from the query-independent scene representation; (2) generate a question without exposing its reference answer; (3) derive the semantic reference independently from privileged construction records; (4) specify the question-specific mandatory evidence units; (5) identify and verify the observations and grounding associated with those units; (6) assign evidence-unit support states from the admitted observations; (7) aggregate those states into configuration-level sufficiency; (8) construct matched strong and weak observation configurations while preserving the underlying scene fact and question; and (9) apply automatic integrity checks and targeted human review before the example is admitted to the frozen benchmark.

The remainder of this section describes these stages in detail. The evaluated model receives only the question, original unboxed images, and camera–time identifiers; all construction annotations remain privileged.

A.3 QUESTION AND EVIDENCE ANNOTATION PROTOCOL

We next detail construction flow as summarized above: how questions are generated, semantic references are recovered independently, required evidence is specified, grounding is verified, and observation-level support is converted into configuration-level sufficiency. DriveAnswer separates semantic truth, visual evidence, grounding, and sufficiency rather than deriving all labels from a single annotation step. The protocol is organized around five logical construction components: question generation, reference-answer recovery, required-evidence specification, observation-support assessment, and verification. These roles may be implemented within the same construction pipeline, but information is kept separated so that one stage cannot determine all downstream labels.

A.3.1 SEPARATED CONSTRUCTION ROLES

Question generation. Questions are generated only from eligible verified scene records. The semantic reference answer is not exposed to the language-based question generator. Qwen3-VL-8B is used only to verbalize benchmark questions and referring expressions from precomputed construction records. The prompt provides the reasoning family, target entities, admissible descriptive attributes, and family-specific generation constraints. It instructs the model to preserve the supplied scene fact, avoid introducing unsupported information, use an unambiguous referring expression, and follow the required question format. Few-shot demonstrations provide a structured input containing the reasoning family, target entity information, and admissible descriptive attributes, followed by an example natural-language question satisfying the family-specific format. The demonstrations are used only to illustrate the desired wording and output structure; they do not provide benchmark labels for the generated examples. Generated referring expressions are treated as candidates rather than semantic ground truth. They are retained only when the intended target is visually identifiable, the description is unique in the admitted observation, same-category distractors are accounted for, and the wording does not leak the answer. Unsupported or unstable descriptions are rewritten or rejected. Reference answers, grounding targets, evidence requirements, and strong-weak sufficiency labels are derived independently from the benchmark construction records.

Reference-answer recovery. The semantic answer is recovered independently from privileged construction information. For the six scene-understanding families, the reference is derived from verified Layer-2 world-state facts or family-specific recovery records. For Evidence-Conditioned Decision, the reference action is produced by the frozen retrospective decision policy described in Appendix C. The reference answer is therefore fixed before strong and weak observation configurations are compared.

A.3.2 FAMILY-SPECIFIC EVIDENCE REQUIREMENTS

Table 5 makes explicit the distinction between privileged world truth and the visual evidence that supports a model-facing answer. Layer 2 facts determine the semantic reference answer, whereas Layer 3 observations determine where that fact may be visually supported. Layer 4 binds these components to a question-specific evidence requirement and answer-or-abstain target.

Table 5: Question-family evidence definitions. The strong condition preserves observations required to support the verified semantic answer. The weak condition preserves the underlying world-state fact while withholding the visual evidence needed to justify that answer.

Family	**World-state quantity**	**Evidence required in strong**	**Matched weak intervention**
Spatial relation	Relative spatial configuration of queried entities	A camera observation in which the relevant entities can be identified and their relationship established	Remove or make insufficient the observation that establishes the queried relationship
Map relation	Relation between a queried entity and mapped scene structure	Visual support connecting the queried entity to the relevant lane, roadway, crossing, or other map structure	Remove the support required to establish the entity–map relation
Motion state	Whether an object is moving or stationary	Multiple temporally separated observations preserving object identity	Retain an anchor observation but remove the temporal evidence needed to establish motion
Temporal relation	How the relation between two entities changes over time	Temporally separated observations supporting the relational change	Remove the observations required to establish that change
Temporal recovery	Whether missing evidence at the anchor can be recovered at a nearby time	A nearby observation of the same referent that restores the missing evidence	Retain only the insufficient anchor observation
Cross-view recovery	Whether another synchronized view restores missing evidence	A synchronized camera in which the same referent or relation becomes identifiable	Remove the supporting synchronized view while retaining the anchor
Evidence-conditioned decision	Reference action under the frozen conservative temporal decision policy	Multi-camera temporal evidence sufficient to establish relevant actor behavior and scene evolution	Retain the synchronized anchor scene while removing the temporal evolution required by the reference policy

A.3.3 GROUNDING VERIFICATION

Grounding is verified before evidence support is assigned. Each referential evidence unit is linked to the observation in which its visual support is expected and, where applicable, to the corresponding spatial region. An invalid grounding case is repaired or rejected rather than relabeled as insufficient evidence. This distinction prevents annotation or projection errors from being counted as legitimate weak-evidence examples.

For questions requiring multiple evidence targets, each target is verified individually. Question-level sufficiency is not inferred from successful grounding of only one target when multiple mandatory units are required.

Table 6: Evidence-unit support states used by DriveAnswer.

State	Interpretation
SUPPORTED	The admitted observations clearly establish the required evidence unit and its grounding is valid.
PARTIAL	Relevant visual evidence is present, but an essential component needed to establish the required fact is missing.
AMBIGUOUS	Relevant evidence is present, but the admitted observations permit more than one reasonable interpretation.
UNAVAILABLE	No admitted observation establishes the required evidence unit.

A.3.4 OBSERVATION-SUPPORT ASSESSMENT

Support is assessed only from the observations admitted by the model-facing configuration. Reviewers inspect the same unboxed observations available to the evaluated model. Privileged world truth may be consulted to verify identity, chronology, or a construction error, but it cannot substitute for visual evidence when deciding whether the admitted observations support the answer.

A.3.5 LAYER 4 SUPPORT RECORD

Layer 4 binds the semantic question to its required evidence, grounding verification, support state, and model-facing target behavior. A representative evidence-unit record for illustration is shown below:

```
{
  "question_id": "...",
  "configuration_id": "...",
  "evidence_unit_id": "...",
  "grounding_status": "VALID | BAD_GROUNDING",
  "support_state":
      "SUPPORTED | PARTIAL | AMBIGUOUS | UNAVAILABLE",
  "supporting_camera": "CAM_FRONT | ... | null",
  "supporting_sample": "... | null",
  "evidence_mechanism":
      "DIRECT | CROSS_VIEW | TEMPORAL | NONE",
  "confidence": "<value in [0,1]>",
}
```

Identifiers, camera names, observation references, temporal ordering, and label vocabularies are validated automatically. A referenced image must correspond to the camera and sample recorded in the support record. Malformed records, unresolved identifiers, invalid temporal links, and invalid grounding are treated as construction failures rather than silently converted into insufficient evidence. Table 6 summarizes the evidence-unit support states. The evidence_mechanism field records how support is obtained. DIRECT indicates that the required evidence is available in the designated observation; CROSS_VIEW indicates recovery through a synchronized alternative camera; TEMPORAL indicates recovery from a different admitted time step; and NONE indicates that no admitted observation establishes the required evidence. The fields illustrate the information retained for audit and construction; evaluated models do not receive these records.

A.4 SUPPORT AND SUFFICIENCY LABELS

Support is assigned first at the evidence-unit level and then aggregated into a configuration-level sufficiency label. This distinction is important because a single question may require multiple pieces of evidence.



(a) Strong configuration contains the observations required to justify the answer. A correct answer to the question is expected.



(b) Weak configuration does not contain enough observations required to justify the answer, thus ABSTAIN is the preferred response, even when the queried object remains visible.

Figure 2: World truth visibility does not guarantee scene observability. An actor may have high visibility ratio in the metadata yet remains unobservable in a scene.

A.4.1 From Evidence Units to Configuration Sufficiency

Let R_q denote the mandatory evidence units for question q. A configuration is labeled SUFFICIENT only when every mandatory evidence unit is SUPPORTED:

$$\text{Sufficient}(q, c) = 1\left[\forall e \in R_q,\ \text{support}(e, c) = \textsc{Supported}\right].$$

If any mandatory unit is PARTIAL, AMBIGUOUS, or UNAVAILABLE, the observations do not justify producing the semantic answer. The model-facing target is therefore

$$y_{q,c} = \begin{cases} z_q^\star, & \text{if } c \text{ is sufficient,} \\ \texttt{ABSTAIN}, & \text{otherwise.} \end{cases}$$

Internally, configuration records may preserve the reason for insufficiency (e.g., ambiguous versus unavailable) for auditing. At evaluation time, all non-sufficient configurations map to the same required behavioral target, Abstain.

A.5 Strong and Weak Configuration Construction

For each retained question, DriveAnswer constructs a matched pair (c^+, c^-). The strong configuration c^+ must preserve every mandatory evidence unit. The weak configuration c^- preserves the same question and underlying semantic fact while removing or making insufficient at least one mandatory evidence unit (see Figure 2a).

Critically, weak examples are not defined simply by hiding the queried object. For many questions, the referent remains visible in both configurations. Instead, the intervention may remove temporal history, relational context, map support, a synchronized alternative view, or another piece of evidence required to justify the answer (see Figure 2b).

For recovery families, the anchor observation is required to be insufficient. A cross-view or temporal recovery example is accepted only if the additional admitted observation restores evidence that is genuinely absent from the anchor. Cases in which the anchor already determines the answer are not counted as recovery examples.

A.5.1 Matched-Evidence Verification

A candidate matched pair is retained only if all applicable checks pass:

1. **World-state invariance.** The semantic reference answer is unchanged across strong and weak configurations.
2. **Question invariance.** The semantic question and queried entities are identical across the pair.

3. **Identity consistency.** Persistent object identity is preserved across cameras and time steps used by the evidence record.
4. **Observation validity.** Every referenced observation resolves to an existing image, and the recorded camera identity and sample correspond to the actual image path.
5. **Temporal consistency.** Temporal evidence is chronologically ordered and belongs to the intended scene and track.
6. **Strong sufficiency.** Every mandatory evidence unit is supported in the strong configuration.
7. **Weak insufficiency.** At least one mandatory evidence unit becomes partial, ambiguous, or unavailable in the weak configuration.
8. **Criticality of removed evidence.** Removing the designated critical evidence changes configuration sufficiency.
9. **Non-critical-context control.** Removing context that is not part of the mandatory evidence set must not, by itself, change the semantic reference answer or create an artificial sufficiency change.
10. **Recovery validity.** For cross-view and temporal recovery, the anchor must lack the required evidence and the admitted synchronized or temporal observation must restore it.
11. **Grounding validity.** Evidence loss caused by an annotation, projection, identity, or box error is repaired or rejected rather than treated as a valid weak intervention.

These checks ensure that the strong–weak intervention changes the *availability of question-specific evidence* rather than changing the underlying scene fact, question, or referent.

Matched pair integrity. The frozen construction enforces question-level invariance across strong and weak conditions. The semantic reference answer is derived from the same underlying world state in both configurations, queried entity identity is unchanged, and temporal observations remain chronologically consistent. The intervention therefore acts on evidence availability rather than on the underlying semantic fact.

Across the original six reasoning families, 19,019 questions contain the target-bearing Layer 2/Layer 3 records used by the reconstruction audit. The 1,132 evidence-conditioned decision questions use a separate policy-relevant-actor grounding construction and are therefore audited separately in Appendix C.

A.6 Automatic Validation and Review Routing

Construction combines deterministic checks with targeted human review. Automatic validation verifies identifier resolution, scene membership, camera consistency, temporal ordering, duplicate identifiers, grounding references, matched-question identity, reference-answer invariance, and expected strong/weak targets.

Cases are routed for additional review when the evidence state is ambiguous, confidence is low, grounding cannot be verified automatically, or the family-specific construction rule identifies a difficult case. Human review assesses sufficiency from the same unboxed model-facing observations rather than from privileged scene annotations. Privileged information is used only to diagnose construction errors or verify identity.

Ambiguous reviewed cases are retained only when a defensible support judgment can be established under the benchmark definition. Otherwise, they are removed from the benchmark rather than forced into a supported or unsupported label.

A.7 Human Validation

We perform an additional stratified validation of 120 questions from the six original scene-understanding families, with 20 questions sampled from each family. For each configuration, the reviewer asked whether the question could be answered confidently from the admitted observations and whether the supporting evidence could be explained in the reviewer's own words. If the reviewer

could not confidently determine and justify the answer from the available observations, the configuration was not judged sufficient. The reviewer assesses four properties independently: whether the strong configuration contains sufficient evidence, whether the weak configuration is insufficient, whether the queried referent is clear, and whether the example admits an unintended alternative interpretation.

The reviewer agrees that the strong configuration is sufficient for 117/120 examples (97.5%) and that the weak configuration is insufficient for 117/120 examples (97.5%). Referent clarity is accepted for 111/120 examples (92.5%), and absence of unintended ambiguity for 115/120 examples (95.8%). These are marginal criterion counts and are not interpreted as a 97.5% matched-pair acceptance rate.

The Decision family is reviewed separately because its semantic reference is policy-conditioned rather than a direct scene fact. A stratified set of 60 held-out Decision questions contains all 10 PROCEED examples and 25 each from SLOW and STOP. The frozen reference action agrees with manual assessment in 57/60 cases (95.0%): 10/10 for PROCEED, 22/25 for SLOW, and 25/25 for STOP. These reviews are reported as construction-quality diagnostics rather than as model-performance measurements as in Table 7.

Table 7: Human-review diagnostics across the original six families and the Decision-family label audit.

Criterion	**Agreement**
Strong condition sufficient	117/120 (97.5%)
Weak condition insufficient	117/120 (97.5%)
Referent clear	111/120 (92.5%)
No unintended ambiguity	115/120 (95.8%)
Decision reference action, overall	57/60 (95.0%)
PROCEED	10/10 (100.0%)
SLOW	22/25 (88.0%)
STOP	25/25 (100.0%)

B GRAPH AND EVIDENCE VALIDATION

B.1 GRAPH CONSTRUCTION AND RECONSTRUCTION AUDITS

The final graph collection contains 34,149 scene graphs with 7,875,985 nodes and 89,533,730 typed edges. Automated integrity checks found no broken structural references. Geometric projection succeeds for 99.8361% of eligible projection attempts. Temporal construction uses an exact five-frame neighborhood centered on the anchor observation when available.

We additionally reconstruct the privileged fact associated with each target-bearing question independently from its observation-level support record. Among 19,019 target-bearing questions, 18,960 reconstruct directly, corresponding to 99.690%. The remaining 59 normalization exceptions occur in the spatial-relation family and are retained as explicit normalization exceptions rather than silently altering the underlying relation labels. The construction audit verifies that every retained strong configuration contains support for all mandatory evidence units and that every matched weak configuration removes support for at least one mandatory unit. These 100% rates therefore measure compliance with the frozen construction rules rather than independent perceptual agreement. The support rates verify that the benchmark satisfies its strong-weak construction rules; they are not independent estimates of human perceptual sufficiency, which is evaluated separately in Appendix A.7. Tables 8-11 summarize the graph reconstruction, matched-pair integrity, and observation-support audits.

B.2 OBSERVATION-SUPPORT AUDIT

Projection of an annotated target into a camera is not treated as proof of perceptual sufficiency. The benchmark separately records observation-level support and question-specific evidence requirements. All strong configurations preserve the family-specific support required by construction under the

Table 8: Full graph construction audit.

Audit quantity	Result
Scene graphs	34,149
Graph nodes	7,875,985
Graph edges	89,533,730
Structural integrity errors	0
Projection success	99.8361%
Strong Layer 2 answer preserved	100%
Weak Abstain target	100%
Temporal neighborhood	5 frames

Table 9: Layer 2 to Layer 3 reconstruction audit.

Quantity	Result
Target-bearing questions	19,019
Directly reconstructed	18,960
Reconstruction rate	99.690%
Normalization exceptions	59

benchmark construction protocol, whereas the matched weak configuration removes that required support.

For the 19,019 target-bearing questions, all required targets are visually present under the strong configuration. Under weak configurations, 60.308% retain all target objects as visible while 39.692% make at least one target unavailable. This distinction is important: a weak configuration may leave the queried objects visibly present while removing the temporal, cross-view, map, or relational evidence necessary to justify the answer.

Table 10: Matched-pair integrity audit.

Check	Pass rate
Strong semantic target preserved	100.0%
Weak target requires abstention	100.0%
Strong reference matches world truth	100.0%
Weak reference matches abstention target	100.0%
Strong/weak observations resolve	100.0%

Table 11: Construction-level target visibility and required-evidence audit.

Quantity	Result
Target-bearing questions	19,019
Strong: targets visible	100.000%
Weak: all targets visible	60.308%
Weak: target unavailable	39.692%
Strong: required support present by construction	100.000%
Weak: ≥ 1 mandatory support unit removed by construction	100.000%

B.3 Visibility Triage Rather Than Visibility Truth

The visibility procedure was enforced because geometric projection and source visibility metadata are useful but incomplete. The augmentation therefore combines several cues such as target size, boundary truncation, source visibility, estimated occlusion, foreground overlap/crowding, and blur when measurable into a review-routing record. We freeze the routing thresholds only after calibration

on labeled cases. The exact score is not released as a truth label because two observations with similar risk can differ in whether the evidence needed by a particular question is actually identifiable. This distinction is notable for occlusion. Partial occlusion does not automatically imply abstention. A reviewer or support expert instead asks whether the required entity or relation can be established from the admitted observations. The triage stage only decides which cases deserve closer inspection.

B.4 VISIBILITY-ONLY DIAGNOSTIC

We present the visibility-only diagnostic as an oracle-support sanity check, not as a learned baseline as detailed in Table 12. The diagnostic uses privileged construction annotations and therefore must not be interpreted as ordinary model performance. A permissive target-visibility rule obtains 37.21% matched-pair success, while a stricter observability rule reaches 43.89%. These results show that target visibility is informative but insufficient for determining question-specific evidence sufficiency: even privileged visibility rules fail on more than half of the matched pairs. Thus, visibility alone cannot substitute for reasoning over the temporal, cross-view, relational, or map evidence required by the question.

Table 12: Visibility diagnostic. These oracle-support rules use privileged observation annotations and are not model baselines.

Rule	**Strong (%)** ↑	**Weak (%)** ↑	**Pair (%)** ↑	**UCR (%)** ↓	**UAR (%)** ↓
Permissive visibility	100.00	37.21	37.21	62.79	0.00
Strict observability	98.55	45.34	43.89	54.66	1.45

C EVIDENCE-CONDITIONED DECISION FAMILY

C.1 DECISION CONSTRUCTION

The evidence-conditioned decision family contains 1,132 questions: 869 training, 102 validation, and 161 held-out test questions. Decision labels are not treated as direct physical world-state facts. Instead, Proceed, Slow, and Stop are produced via human-designed rules. This policy operates on verified actor trajectories and temporal conflict signals and uses fixed construction thresholds: an 8 m stop distance, 20 m slow distance, 12 m close-caution distance, 25 m clearance threshold, 15 m separating-clearance threshold, and a 1 m relative-change threshold. The presence of leading and following and applicable traffic rules is also accounted for. Complicated and ambiguous construction cases such as driving in pedestrian-dense areas or driving near emergency vehicles are flagged for manual reviews. Because each example contains observations on both sides of the anchor, this family evaluates *retrospective temporal-window reasoning*. It must not be interpreted as online closed-loop driving or as a policy that acts before future observations become available.

C.2 DECISION DISTRIBUTION AND EVIDENCE STATISTICS

Table 13: Decision-family label distribution.

Split	**Proceed**	**Slow**	**Stop**	**Total**
Train	61	690	118	869
Validation	0	96	6	102
Test	10	124	27	161

The 161 Decision test questions contain 521 unique policy-relevant actor targets, an average of 3.236 actors per question. Further statistics are reported in Tables 13 and 14. The number ranges from zero to 18. Eight test questions contain no policy-relevant actor and 153 contain at least one. For grounding evaluation, actor-observation localization instances are enumerated separately from unique policy-relevant actor identities. The Decision grounding evaluator therefore contains 1,058 target instances over the same 161 test questions. We keep these denominators separate: 521 describes unique actors used by the reference policy, whereas 1,058 counts actor-observation grounding targets.

Table 14: Decision-family evidence statistics.

Quantity	Result
Mean relevant actors, train	3.474
Mean relevant actors, validation	2.441
Mean relevant actors, test	3.236
Approaching actor trends	2,095
Stable actor trends	1,463
Separating actor trends	231
Proceed examples with no relevant actor	60

C.3 DECISION SHORTCUT CONTROLS

The held-out Decision label distribution is imbalanced: 124 SLOW, 27 STOP, and 10 PROCEED. An always-SLOW predictor reaches 77.02% accuracy, but only 29.01% macro-F1 and 33.33% balanced accuracy. Indicating monotonic policy are not suffice for determining the correct driving decision. We additionally evaluate a count-only baseline using the number of policy-relevant actors. For every actor count observed during training, the baseline predicts the majority training label associated with that count. The learned rule maps zero actors to PROCEED and every observed positive actor count to SLOW. The count-only model obtains 81.99% accuracy, 59.47% macro-F1, and 60.00% balanced accuracy (see Table 15). This association is partly expected: clear scenes with no relevant actors tend to support PROCEED, while scenes with relevant actors require the ego vehicle to act more cautiously. More severe cases requiring STOP are correspondingly rarer. However, actor count alone does not determine the action: the count-only baseline predicts none of the 27 STOP questions correctly. The number of actors is treated as privileged construction information that is not provided to the model at evaluation; thus this experiment is a distributional diagnostic and not a model-facing baseline.

Table 15: Decision-family shortcut baselines.

Baseline	Accuracy (%) ↑	Macro-F1 (%) ↑	Balanced accuracy (%) ↑
Always SLOW	77.02	29.01	33.33
Actor-count only	81.99	59.47	60.00

D EVALUATION AND STATISTICAL RELIABILITY

D.1 EVALUATION METRICS

For each matched question pair, Strong Semantic Accuracy measures whether the strong-condition answer equals the semantic reference. Weak Abstention Accuracy measures whether the weak-condition response is Abstain. Matched-Pair Success requires both events for the same question. Unsupported Commitment Rate (UCR) measures parsed non-Abstain commitments under weak evidence. Unnecessary Abstention Rate (UAR) measures abstention on answerable strong inputs.

D.2 ALL-G AND JAGS

For a question with multiple required grounding targets, All-G@t equals one iff every required target achieves Joint@t success. Joint Answer–Grounding Success is then defined as

$$\mathrm{JAGS@}t = \mathbb{1}\left[\mathrm{PairSuccess} \wedge \mathrm{AllG@}t\right].$$

Because answering and grounding are elicited through separate calls, JAGS measures co-occurring answer, abstention, and localization success. It is not interpreted as evidence that the predicted boxes were the causal internal explanation of the semantic answer.

Table 16: Selected results on the canonical six-family held-out test set with 95% scene-clustered bootstrap intervals. All systems are evaluated on the same 2,830 matched questions. $t = 0.50$.

System	Strong (%) ↑	Pair (%) ↑	JAGS@t (%) ↑
Gemma-C	82.58 [78.80, 86.16]	82.05 [78.14, 85.70]	62.37 [55.43, 69.39]
Gemma-D	82.90 [79.26, 86.29]	82.44 [78.66, 85.93]	63.29 [56.41, 70.23]
Qwen-C	82.86 [78.94, 86.62]	81.70 [77.63, 85.59]	62.26 [54.62, 69.94]
Qwen-D	83.43 [79.39, 87.16]	82.08 [77.77, 86.05]	62.83 [55.06, 70.50]

D.3 Scene-clustered bootstrap

D.3.1 6-family Bootstrap Results

We estimate uncertainty with a scene-clustered bootstrap over the six-family test set. We choose scene-clustered bootstrapping because matched strong-weak questions and grounding targets may originate from the same scene and hence are not statistically independent. Moreover, metrics such as All-G and JAGS are question-level conjunctions, which requires bootstrapping of scenes rather than parametric variance tests. We therefore estimate uncertainty by resampling held-out scenes rather than individual questions or targets. The six-family scene-clustered bootstrap confirms that the main adaptation patterns are not driven by a small number of held-out scenes (see Table 16). For Gemma, Stage C reaches 82.58% Strong Accuracy (95% CI: [78.80, 86.16]), 82.05% Pair Success ([78.14, 85.70]), and 62.37% JAGS@.50 ([55.43, 69.39]); Stage D is similar at 82.90% ([79.26, 86.29]), 82.44% ([78.66, 85.93]), and 63.29% ([56.41, 70.23]), respectively. Qwen shows the same stability between the final two stages: Stage C obtains 82.86% Strong Accuracy ([78.94, 86.62]), 81.70% Pair Success ([77.63, 85.59]), and 62.26% JAGS@.50 ([54.62, 69.94]), compared with 83.43% ([79.39, 87.16]), 82.08% ([77.77, 86.05]), and 62.83% ([55.06, 70.50]) at Stage D. The corresponding grounding intervals likewise overlap strongly: Qwen Joint@.50 is 83.10% [78.72, 87.43] at Stage C and 83.11% [78.76, 87.49] at Stage D.

Table 17: Stage C minus Stage D paired differences on the canonical six-family test set. Values are percentage points with 95% scene-clustered intervals.

Metric	Gemma C–D	Qwen C–D
Strong (%) ↑	-0.32 [-1.25, +0.56]	-0.57 [-1.84, +0.50]
Pair (%) ↑	-0.39 [-1.32, +0.50]	-0.39 [-1.70, +0.73]
UCR (%) ↓	+0.04 [-0.15, +0.25]	-0.32 [-0.67, +0.00]
Joint@.50 (%) ↑	-0.45 [-1.02, +0.08]	-0.02 [-0.38, +0.34]
All-G@.50 (%) ↑	-0.95 [-2.00, +0.00]	-0.28 [-0.89, +0.32]
JAGS@.50 (%) ↑	-0.92 [-2.09, +0.18]	-0.57 [-1.65, +0.38]

The paired analysis in Table 17 provides a stricter comparison because both stages are evaluated on exactly the same questions and bootstrap scene draws. For Gemma, the C–D differences are -0.32 points for Strong Accuracy (95% CI: [-1.25, +0.56]), -0.39 for Pair Success ([-1.32, +0.50]), -0.45 for Joint@.50 ([-1.02, +0.08]), and -0.92 for JAGS@.50 ([-2.09, +0.18]). Qwen is similarly close: the C–D differences are -0.57 for Strong Accuracy ([-1.84, +0.50]), -0.39 for Pair Success ([-1.70, +0.73]), -0.02 for Joint@.50 ([-0.38, +0.34]), and -0.57 for JAGS@.50 ([-1.65, +0.38]). Most intervals contain zero, so the six-family results do not support a consistent aggregate improvement from Stage C to Stage D.

The scene-clustered analysis reinforces the separation between answer behavior and grounding. For both backbones, Stage B→C produces large gains in grounding and JAGS while changes in answer-side metrics are comparatively smaller; subsequent Stage C→D changes are small.

Table 18: Gemma no-image text+metadata control.

Strong (%) ↑	**Weak (%)** ↑	**Pair (%)** ↑	**UCR (%)** ↓	**UAR (%)** ↓
3.31	99.36	3.08	0.47	96.42

E NO-IMAGE+METADATA SHORTCUT CONTROLS

We evaluate a zero-shot Gemma-4-12B under the 2,991-pair test protocol after removing both image placeholders and image pixels and report the results in Table 18. The question text, allowed label set, and camera and time observation metadata are left unchanged. We therefore refer to this setting as the *no-image text and metadata baseline*. Without visual input, Gemma strongly defaults to abstention: Strong Semantic Accuracy falls to 3.31%, Weak Abstention Accuracy is 99.36%, Matched-Pair Success is 3.08%, UCR is 0.47%, and UAR reaches 96.42%. Of 2,991 strong examples, 2,884 produce Abstain. The family breakdown is also informative. Strong accuracy is 11.94% for cross-view recovery, 0.00% for map relation, 1.71% for motion state, 4.72% for spatial relation, 51.72% for temporal recovery, 0.00% for temporal relation, and 0.00% for Decision. The unusually high temporal-recovery value suggests residual task or language priors in that small family and is treated as a limitation rather than evidence of visual reasoning. This control indicates that the benchmark as presented to a zero-shot model is not trivially solvable from question wording and camera and metadata alone.

F ZERO-SHOT BASELINES

F.1 ZERO-SHOT ANSWERING PERFORMANCE PER FAMILY

Table 19 reveals that aggregate zero-shot results conceal substantial differences across reasoning families. Recovery tasks are comparatively favorable for some models: Gemma reaches 93.10% Pair Success on Temporal Recovery, while Qwen reaches 82.76%. In contrast, Motion State illustrates a different failure mode: abstention is rare among all models, reducing Pair Success to 1.08% for Qwen, 2.98% for InternVL, 0.54% for Molmo, and 0.09% for LLaVA.

Decision is the clearest shared zero-shot weakness. Gemma, Qwen, InternVL, and Molmo obtain zero matched-pair success, while LLaVA reaches only 0.62%. Map, Spatial, and Temporal Relation families show intermediate and model-dependent behavior. These differences indicate that the overall zero-shot failure is not attributable to a single reasoning family.

F.2 ZERO-SHOT GROUNDING PERFORMANCE PER FAMILY

The family-level results reveal distinct grounding failure modes across reasoning types as detailed in Table 20. Spatial Relation and Map Relation are favorable for Gemma and Qwen, both of which achieve substantially higher target-level Joint@0.5 and question-level All-G@0.5 than the remaining zero-shot models. Temporal Recovery is also relatively successful for Gemma at the target level, although complete-evidence recovery remains considerably lower. Cross-view Recovery and Decision are substantially harder since target-level grounding does not reliably translate into complete question-level evidence recovery (Qwen reaches 23.38% for Joint@0.5 but only 2.99% for All-G@0.5). Decision is similarly challenging, with high presence accuracy among many models but weak localization and JAGS, indicating that detecting relevant actors is substantially easier than localizing the complete evidence needed for meaningful answering. The weaker grounding models exhibit this discrepancy among the majority of reasoning tasks. Together, the zero-shot models support the joint reporting of presence detection, target localization, complete evidence recovery, and matched evidence-answering because they simultaneously measure different aspects of models' performance.

Table 19: Per-family zero-shot answering under the seven-family held-out protocol. Pair requires a correct strong-condition answer and correct abstention on the matched weak condition. **Results are expressed as percentages.**

Family	Model	Strong ↑	Weak ↑	Pair ↑	UCR ↓	UAR ↓
Cross-view recovery	Gemma-4-12B	54.73	99.00	**53.73**	1.00	45.27
	Qwen3-VL-8B	51.74	**100.00**	51.74	**0.00**	48.26
	InternVL3.5-8B	60.70	45.27	19.90	54.73	39.30
	Molmo2-ER	20.40	97.01	18.41	2.99	79.60
	LLaVA-OneVision-7B	**73.13**	53.23	32.34	46.77	**26.87**
Decision	Gemma-4-12B	18.63	0.00	0.00	100.00	**0.00**
	Qwen3-VL-8B	13.66	0.00	0.00	100.00	**0.00**
	InternVL3.5-8B	**63.35**	0.00	0.00	100.00	0.62
	Molmo2-ER	6.21	0.00	0.00	100.00	**0.00**
	LLaVA-OneVision-7B	34.78	**2.48**	**0.62**	**97.52**	1.86
Map relation	Gemma-4-12B	**54.59**	**95.14**	**51.89**	**4.86**	5.41
	Qwen3-VL-8B	43.24	68.65	25.95	31.35	0.54
	InternVL3.5-8B	39.46	31.35	7.03	68.65	**0.00**
	Molmo2-ER	41.08	9.73	4.32	90.27	**0.00**
	LLaVA-OneVision-7B	36.76	82.16	28.65	17.84	14.05
Motion state	Gemma-4-12B	76.28	**44.72**	**33.90**	**55.28**	**0.00**
	Qwen3-VL-8B	**83.23**	1.71	1.08	98.29	0.09
	InternVL3.5-8B	82.96	4.87	2.98	95.13	1.08
	Molmo2-ER	80.25	1.53	0.54	98.47	0.09
	LLaVA-OneVision-7B	46.62	0.63	0.09	99.37	0.45
Spatial relation	Gemma-4-12B	**69.70**	73.85	**51.27**	26.15	**4.95**
	Qwen3-VL-8B	31.22	**92.17**	27.19	**7.83**	18.09
	InternVL3.5-8B	34.56	73.39	22.24	26.61	24.19
	Molmo2-ER	44.59	83.18	36.29	16.82	26.61
	LLaVA-OneVision-7B	26.96	63.82	13.71	36.18	11.18
Temporal recovery	Gemma-4-12B	96.55	96.55	**93.10**	3.45	3.45
	Qwen3-VL-8B	82.76	**100.00**	82.76	**0.00**	17.24
	InternVL3.5-8B	**100.00**	13.79	13.79	86.21	**0.00**
	Molmo2-ER	62.07	82.76	44.83	17.24	37.93
	LLaVA-OneVision-7B	79.31	24.14	17.24	75.86	20.69
Temporal relation	Gemma-4-12B	**49.09**	**97.26**	**46.35**	**2.74**	9.13
	Qwen3-VL-8B	34.02	49.77	15.75	50.23	16.44
	InternVL3.5-8B	28.54	92.01	23.97	7.99	16.44
	Molmo2-ER	24.20	42.24	4.79	57.76	20.32
	LLaVA-OneVision-7B	29.22	0.68	0.23	99.32	**0.00**

G Adapted Models

G.1 Exact Stage-D matched-pair objective

For a matched strong–weak pair (X^+, X^-), let y denote the semantic answer and let $a = \mathrm{Abstain}$. We define the support score for target u as

$$S(X, u) = -\,\mathrm{CE}(f(X), u)\,,$$

Table 20: Per-family zero-shot grounding results on the held-out test set. Models are grouped within each question family to expose family-specific differences. Joint@.50 evaluates target-level grounding; All-G@.50 requires every target associated with a question to be grounded; JAGS@.50 additionally requires matched-pair QA success. We report $t = 0.50$. **All values are expressed as percentages.** Cross is Cross-view Recovery; Map Rel. is Map Relation; Temp. Rec. is Temporal Recovery; Temp. Rel. is Temporal Relation. Crossview N=201; Map Relation N=185, Decision N=161; Spatial Relation N=868, Motion State N=1109, Temporal Recovery N=29, Temporal Relation N=438.

Family	Model	Presence ↑	IoU@t ↑	Joint@t ↑	All-G@t ↑	JAGS@t ↑
Cross.	Gemma-4-12B	**74.38**	19.06	14.18	2.49	**1.99**
	Qwen3-VL-8B	58.71	**39.83**	**23.38**	**2.99**	**1.99**
	InternVL3.5-8B	38.31	0.00	0.00	0.00	0.00
	Molmo2-ER	73.63	0.34	0.25	0.00	0.00
	LLaVA-OneVision-7B	56.22	0.00	0.00	0.00	0.00
Map Rel.	Gemma-4-12B	**94.05**	61.67	75.41	54.05	**29.73**
	Qwen3-VL-8B	89.46	**72.73**	**78.11**	**57.30**	15.14
	InternVL3.5-8B	83.51	0.00	48.65	0.00	0.00
	Molmo2-ER	89.19	0.67	49.19	0.54	0.00
	LLaVA-OneVision-7B	88.11	0.00	49.19	0.00	0.00
Deci sion	Gemma-4-12B	**17.57**	9.74	2.48	0.00	52.74
	Qwen3-VL-8B	**86.48**	13.99	**12.67**	**14.29**	0.00
	InternVL3.5-8B	46.88	0.00	1.04	1.86	0.00
	Molmo2-ER	94.90	0.20	0.76	0.00	0.00
	LLaVA-OneVision-7B	13.99	0.00	1.51	4.97	**0.62**
Spatial Rel.	Gemma-4-12B	**91.70**	76.72	80.09	64.63	**36.87**
	Qwen3-VL-8B	89.99	**84.97**	**83.13**	**70.39**	19.82
	InternVL3.5-8B	87.56	0.26	45.63	0.23	0.12
	Molmo2-ER	79.20	0.00	42.37	0.00	0.00
	LLaVA-OneVision-7B	88.33	0.00	44.97	0.00	0.00
Motion State	Gemma-4-12B	**88.64**	68.92	61.09	45.72	**18.85**
	Qwen3-VL-8B	85.71	**73.49**	**62.98**	**52.21**	0.18
	InternVL3.5-8B	75.16	0.00	0.00	0.00	0.00
	Molmo2-ER	85.03	0.21	0.18	0.00	0.00
	LLaVA-OneVision-7B	75.07	0.00	0.00	0.00	0.00
Temp. Rec.	Gemma-4-12B	77.59	**53.33**	**41.38**	**20.69**	**20.69**
	Qwen3-VL-8B	62.07	44.44	27.59	3.45	3.45
	InternVL3.5-8B	48.28	0.00	0.00	0.00	0.00
	Molmo2-ER	**81.03**	12.77	10.34	0.00	0.00
	LLaVA-OneVision-7B	75.86	4.55	3.45	0.00	0.00
Temp. Rel.	Gemma-4-12B	**89.27**	53.07	**47.37**	**9.82**	**3.65**
	Qwen3-VL-8B	81.85	**56.00**	45.83	8.90	1.14
	InternVL3.5-8B	72.26	0.08	0.06	0.00	0.00
	Molmo2-ER	88.93	0.26	0.23	0.00	0.00
	LLaVA-OneVision-7B	64.90	0.00	0.00	0.00	0.00

so that larger $S(X, u)$ indicates greater model support for u. With margin $m = 0.5$, the matched-pair objective imposes four hinge constraints:

$$h_{\text{strong}} = \max\big(0,\, m - S(X^+, y) + S(X^+, a)\big)\,, \tag{5}$$

$$h_{\text{weak}} = \max\big(0,\, m - S(X^-, a) + S(X^-, y)\big)\,, \tag{6}$$

$$h_{\text{sem}} = \max\big(0,\, m - S(X^+, y) + S(X^-, y)\big)\,, \tag{7}$$

$$h_{\text{abs}} = \max\big(0,\, m - S(X^-, a) + S(X^+, a)\big)\,. \tag{8}$$

The first two terms enforce the desired prediction within each evidence condition: the strong configuration should prefer the semantic answer over abstention, whereas the weak configuration should prefer abstention over the semantic answer. The latter two terms enforce the desired change across the matched intervention: semantic-answer support should be higher under strong evidence, while abstention support should be higher under weak evidence.

The matched-pair loss is

$$\mathcal{L}_{\text{match}} = \frac{1}{4}\left(h_{\text{strong}} + h_{\text{weak}} + h_{\text{sem}} + h_{\text{abs}}\right).$$

The complete Stage-D objective is

$$\mathcal{L}_{\text{pair-aware}} = \mathcal{L}_{\text{ans+suff}} + \lambda_g \mathcal{L}_{\text{ground}} + \lambda_m \mathcal{L}_{\text{match}}, \tag{9}$$

with $\lambda_g = 0.5$ and $\lambda_m = 0.25$. Stage D therefore tests whether direct coupling across matched strong–weak predictions provides additional benefit beyond sufficiency-aware answer supervision and referential grounding supervision.

G.2 Overall Gemma and Qwen Performance on 6-family questions

Tables 21, 22, 23, and 24 report the complete six-family per-family results for Gemma and Qwen, respectively.

Gemma-4-12B per-family behavior. Gemma shows the clearest Stage B→C grounding gains in Motion State, Spatial Relation, and Temporal Relation. Motion State Joint@.50 rises from 60.46% to 88.59%, with JAGS@.50 increasing from 44.91% to 73.22%. Spatial Relation improves from 79.98% to 91.81% Joint@.50 and from 66.82% to 82.72% JAGS@.50, while Temporal Relation rises from 46.35% to 78.54% Joint@.50 and from 8.90% to 28.08% JAGS@.50. Cross-view Recovery remains the hardest family despite improvement in target-level grounding: Joint@.50 increases from 11.94% at Stage B to 51.49% at Stage C and 56.47% at Stage D, but All-G@.50 remains only 14.43% and 21.39% at the latter two stages. Temporal Recovery shows a similar pattern at smaller scale, reaching only 34.48% All-G@.50 at Stage D. Across most families, Stage D changes the Stage-C result only modestly, consistent with the aggregate C–D comparison. Results are shown in Tables 21 and 22

Qwen3-VL-8B per-family behavior. Qwen exhibits the same broad separation between selective answering and grounding quality. At Stage B, Pair Success is already high for several families, including 100.00% for Cross-view Recovery, 89.09% for Motion State, and 90.90% for Spatial Relation, while complete grounding is substantially lower. Cross-view Recovery is the strongest example: Stage-B JAGS@.50 is only 1.00%, increasing to 26.37% at both Stages C and D. Motion State improves from 37.96% JAGS@.50 at Stage B to 72.59% at Stage C and 73.94% at Stage D, while Spatial Relation rises from 68.32% to 81.91% and 82.49%. Temporal Relation also improves substantially in grounding, with Joint@.50 increasing from 46.23% at Stage B to 76.60% at Stage C, although JAGS@.50 remains lower at 23.29%. Map Relation shows a smaller gain, and Temporal Recovery remains difficult because of the small held-out population. As with Gemma, most family-level changes from Stage C to Stage D are small relative to the Stage B→C grounding gains.

Overall, the per-family results show that while some families gain stronger JAGS with training, some failure persisted, suggesting challenging families can improve in answering and grounding separately yet still struggle to satisfy both requirements simultaneously.

Table 21: Per-family Gemma-4-12B answering results on the canonical six-family held-out test set. Pair denotes matched-pair QA success. **All results are expressed as percentages**. Cross. is Cross-view Recovery; Map Rel. is Map Relation; Temp. Rec. is Temporal Recovery; Temp. Rel. is Temporal Relation.

Family	Stage	Strong ↑	Weak ↑	Pair ↑	UCR ↓	UAR ↓
Cross-view Recovery	B	100.00	100.00	100.00	0.00	0.00
	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	100.00	100.00	0.00	0.00
Map Relation	B	50.81	98.38	50.27	1.62	10.27
	C	50.27	98.92	49.73	1.08	10.27
	D	49.19	98.38	48.65	1.62	8.65
Motion State	B	84.76	100.00	84.76	0.00	0.09
	C	86.65	99.73	86.38	0.27	0.18
	D	86.65	100.00	86.65	0.00	0.18
Spatial Relation	B	93.55	98.96	92.63	1.04	5.88
	C	93.32	98.50	92.05	1.50	5.65
	D	93.43	98.39	92.05	1.61	5.07
Temporal Recovery	B	100.00	100.00	100.00	0.00	0.00
	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	100.00	100.00	0.00	0.00
Temporal Relation	B	52.74	100.00	52.74	0.00	1.60
	C	55.48	100.00	55.48	0.00	1.14
	D	57.76	100.00	57.76	0.00	0.91
Overall	B	81.52	**99.58**	81.20	**0.42**	2.76
	C	82.58	99.36	82.05	0.64	2.65
	D	**82.90**	99.40	**82.44**	0.60	**2.44**

G.3 Overall Gemma and Qwen Performance on 7-family set

The following results are reported from separate adaptation training with the Decision family included and should not be interpreted as equivalent to the results reported in the six-family study.

Sufficiency and Grounding Supervision Drive Most Adaptation Gains. The adaptation results in Tables 25 and 28 show that the largest change from zero-shot behavior comes from supervising evidence-conditioned answering. Both adapted backbones achieve high abstention accuracy and Pair Success while preserving answering accuracy under supported condition. Here, Gemma-C reaches 81.58% Strong Accuracy, 98.93% Weak Abstention Accuracy, and 80.68% Pair Success. Qwen-C shows the same overall pattern at a slightly higher QA level, reaching 83.95% Strong Accuracy, 98.53% Weak Abstention Accuracy, and 82.78% Pair Success. Thus, evidence-sensitive answering is not specific to a single backbone.

Grounding is evaluated through a prediction of presence and, when found, observation index and normalized box. Construction boxes are hidden from the model and used only for scoring. On the test set, the Gemma-C model reaches 96.82% presence accuracy, 76.67% localization accuracy at IoU ≥ 0.5, and 77.16% joint presence-localization success. Stage D changes these aggregate grounding results minimally, reaching 97.03% presence accuracy, 76.76% IoU ≥ 0.5, and 77.43% Joint@0.5. Similarly, Qwen-C reaches 96.45% presence accuracy, 74.01% localization accuracy at IoU ≥ 0.5, and 74.71% joint presence–localization success. Stage D is similar, with 96.49% presence accuracy, 73.70% localization accuracy at IoU ≥ 0.5, and 74.46% joint presence–localization success. Thus, these results show that grounding supervision substantially improves the adapted models' localization capabilities and that the additional Stage D pair objective does not produce a uniform aggregate localization improvement beyond explicit grounding supervision. Despite their differences in their zero-shot behavior, the two backbones converge to similar evidence-supported performance after

Table 22: Per-family Gemma-4-12B grounding and evidence-supported answering results on the canonical six-family held-out test set. Joint@.50 is target-level joint presence–localization success; All-G@.50 requires all grounding targets for a question to succeed; JAGS@.50 additionally requires Pair Success. We report $t = 0.50$.

Family	Stage	Joint@.50 (%) ↑	All-G@.50 (%) ↑	JAGS@.50 (%) ↑
Cross-view Recovery	B	11.94	15.42	15.42
	C	51.49	14.43	14.43
	D	56.47	21.39	21.39
Map Relation	B	77.03	58.38	30.27
	C	90.54	82.16	42.16
	D	90.54	82.16	40.54
Motion State	B	60.46	53.56	44.91
	C	88.59	83.77	73.22
	D	89.04	84.67	73.94
Spatial Relation	B	79.98	69.01	66.82
	C	91.81	84.22	82.72
	D	91.81	84.45	82.26
Temporal Recovery	B	27.59	24.14	24.14
	C	60.34	27.59	27.59
	D	63.79	34.48	34.48
Temporal Relation	B	46.35	14.61	8.90
	C	78.54	49.54	28.08
	D	78.42	49.32	29.68
Overall	B	59.75	49.58	42.79
	C	84.41	73.00	62.37
	D	**84.87**	**73.96**	**63.29**

adaptation. Qwen retains slightly stronger strong-condition understanding, whereas Gemma attains slightly stronger grounding and question-level evidence completion.

Explicit matched-pair objective Does Not Produce Uniform Improvements. The explicit matched-pair ranking objective in Stage D has a family-dependent effect. For Gemma, Stage D slightly increases Joint@0.5 from 77.16% to 77.43% and All-G@0.5 from 72.58% to 73.19%, but Strong Accuracy falls from 81.58% to 78.87%, Pair Success from 80.68% to 78.13%, and JAGS@0.5 from 61.79% to 59.41%. Qwen shows an even smaller C-D difference: Pair Success changes from 82.78% to 82.25%, Joint@0.5 from 74.71% to 74.45%, and JAGS@0.5 from 61.12% to 60.62%. These results suggest that explicit pair coupling is not a universally additive improvement once sufficiency and grounding supervision are already present; its effect is small and can vary by backbone and question family. Appendix D.3 presents paired bootstrap uncertainty. Detailed per-family results for Gemma and Qwen are reported below in Appendix G.3.2.

G.3.1 Per-family Gemma Performance

Adapted Model Performance Varies by Reasoning Family Tables 25 and 26 show the differences in performance of all models across the different families. For both backbones, spatial and motion questions are among the strongest evidence-supported categories. Gemma-C reaches 81.91% JAGS@0.5 on spatial relation and 75.02% on motion state, while Qwen reaches 83.64% and 75.38%, respectively as seen in Table 28. In contrast, Decision remains substantially harder: JAGS@0.5 is only 12.42% for Gemma C and 13.66% for Qwen-C, despite much higher Pair Success. This gap suggests that producing appropriate answer-abstain behavior is easier than grounding all of the policy-relevant actors required to support a driving decision. Map Relation reveals another challenge: JAGS@0.5 near 38–39% for both backbones, while temporal relation remains more difficult at 31.28% for Gemma-C and 24.43% for Qwen-C. The family-level comparison thus reveals the differences and challenges that each family poses for the adapted models.

Table 23: Per-family Qwen3-VL-8B answering results on the canonical six-family held-out test set. Pair denotes matched-pair QA success. **Results are expressed as percentages.** Cross. is Cross-view Recovery; Map Rel. is Map Relation; Temp. Rec. is Temporal Recovery; Temp. Rel. is Temporal Relation.

Family	Stage	Strong (%) ↑	Weak (%) ↑	Pair (%) ↑	UCR (%) ↓	UAR (%) ↓
Cross-view Recovery	B	100.00	100.00	100.00	0.00	0.00
	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	99.00	99.00	1.00	0.00
Map Relation	B	58.38	96.76	56.76	3.24	5.95
	C	58.92	96.22	56.76	3.78	5.95
	D	57.84	95.14	55.68	4.86	4.86
Motion State	B	89.09	100.00	89.09	0.00	0.00
	C	86.74	100.00	86.74	0.00	0.00
	D	87.92	100.00	87.92	0.00	0.00
Spatial Relation	B	94.82	95.16	90.90	4.84	3.69
	C	94.59	95.97	91.24	4.03	3.69
	D	94.82	95.39	91.13	4.61	3.69
Temporal Recovery	B	100.00	100.00	100.00	0.00	0.00
	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	100.00	100.00	0.00	0.00
Temporal Relation	B	53.20	100.00	53.20	0.00	0.91
	C	50.91	100.00	50.91	0.00	0.00
	D	51.60	100.00	51.60	0.00	0.00
Overall	B	**84.17**	98.30	**82.86**	1.70	1.66
	C	82.86	**98.52**	81.70	**1.48**	1.52
	D	83.43	98.20	82.08	1.80	**1.45**

G.3.2 Per-family Qwen performance

Qwen exhibits the same broad pattern as Gemma: the effect of Stage D matched-pair objective is family-dependent rather than uniform improvements. Cross-view Recovery and Temporal Recovery remain high in matched-pair QA, while Spatial Relation and Temporal Relation remain substantially harder, with lower Pair Success. Table 27 illustrates these detailed results.

Grounding shows a different pattern as seen in Table 28. Qwen achieves high target-level localization on Spatial Relation, Motion State, Map Relation, but complete-evidence recovery is much weaker for Cross-view Recovery and Decision. This gap between target-level grounding and question-level All-G/JAGS indicates that Qwen can often localize individual targets without consistently recovering the full evidence set required for a matched question. For Stage C, JAGS@.50 ranges from 13.66% on Decision and 25.87% on Cross-View Recovery to 75.38% on Motion State and 83.64% on Spatial Relation, highlighting substantial variation in how reliably Qwen combines correct answering with complete evidence recovery.

Stage D produces only modest changes overall with Spatial Relation remaining strong while Cross-view Recovery and Temporal Recovery experiencing mixed results. This result reinforces the finding that explicit matched-pair coupling does not yield a monotonic improvement across different models (Gemma and Qwen).

G.4 All-G as a function of target count

All-G is an all-or-nothing question-level conjunction and therefore becomes increasingly difficult as the number of required grounding targets grows. The test set contains 2,429 questions with two grounding targets and 506 questions with four. Only 56 questions contain six or more grounding targets. Across all systems, All-G@.50 decreases substantially as target count increases. For example,

Table 24: Per-family Qwen3-VL-8B grounding and evidence-supported answering results on the canonical six-family held-out test set. Joint@.50 is target-level joint presence–localization success; All-G@.50 requires all grounding targets for a question to succeed; JAGS@.50 additionally requires Pair Success. We report $t = 0.50$.

Family	Stage	Joint@.50 (%) ↑	All-G@.50 (%) ↑	JAGS@.50 (%) ↑
Cross-view Recovery	B	12.44	1.00	1.00
	C	57.71	26.37	26.37
	D	57.96	26.37	26.37
Map Relation	B	77.30	55.68	34.59
	C	78.11	63.78	40.00
	D	78.11	63.78	38.92
Motion State	B	57.48	41.93	37.96
	C	87.15	83.59	72.59
	D	87.02	83.59	73.94
Spatial Relation	B	83.30	69.82	68.32
	C	91.37	84.10	81.91
	D	91.59	84.68	82.49
Temporal Recovery	B	27.59	13.79	13.79
	C	74.14	58.62	58.62
	D	75.86	58.62	58.62
Temporal Relation	B	46.23	9.59	4.57
	C	76.60	46.80	23.29
	D	76.48	47.49	22.83
Overall	B	59.67	43.18	39.01
	C	83.10	72.44	62.26
	D	**83.11**	**72.72**	**62.83**

Table 25: Per-family Gemma-4-12B answering results on the full seven-family held-out test set. Pair denotes matched-pair QA success.

Family	Stage	Strong (%) ↑	Weak (%) ↑	Pair (%) ↑
Cross-view	C	100.00	99.50	99.50
	D	100.00	100.00	100.00
Decision	C	63.35	100.00	63.35
	D	67.70	98.14	66.46
Map Relation	C	54.05	98.38	53.51
	D	52.43	98.38	51.89
Motion State	C	87.20	99.37	86.56
	D	78.09	99.82	77.91
Spatial Relation	C	94.12	97.58	92.05
	D	94.70	97.58	92.74
Temporal Recover	C	100.00	100.00	100.00
	D	100.00	100.00	100.00
Temporal Relation	C	51.14	100.00	51.14
	D	53.65	100.00	53.65
Overall	C	**81.58**	98.93	**80.68**
	D	78.87	**99.03**	78.13

Table 26: Per-family Gemma-4-12B grounding and evidence-supported answering results on the full seven-family held-out test set. Joint@.50 is target-level joint presence–localization success; All-G@.50 requires all grounding targets associated with a question to succeed; JAGS@.50 additionally requires matched-pair QA success. We report $t = 0.50$.

Family	**Stage**	**Joint@.50 (%)** ↑	**All-G@.50 (%)** ↑	**JAGS@.50 (%)** ↑
Cross-view Recovery	C	59.20	29.35	29.35
	D	60.20	29.85	29.85
Decision	C	26.94	26.09	12.42
	D	26.94	27.33	13.04
Map Relation	C	85.95	74.59	39.46
	D	85.68	72.97	37.84
Motion State	C	89.59	85.93	75.02
	D	90.44	87.29	67.54
Spatial Relation	C	90.82	83.29	81.91
	D	91.81	85.37	84.10
Temporal Recovery	C	75.86	55.17	55.17
	D	62.07	34.48	34.48
Temporal Relation	C	79.97	54.79	31.28
	D	79.34	52.74	31.28
Overall	C	77.16	72.58	**61.79**
	D	**77.43**	**73.19**	59.41

Table 27: Per-family QA results for adapted Qwen3-VL-8B on the seven-family held-out test set. Cross. is Cross-view Recovery; Decision is Evidence-conditioned Decision; Map is Map Relation; Motion is Motion State; Spatial is Spatial Relation; Temp. Rec. is Temporal Recovery; Temp. Rel. is Temporal Relation.

Family	**Stage**	**Strong** (%) ↑	**Weak** (%) ↑	**Pair** (%) ↑	**UCR** (%) ↓	**UAR** (%) ↓
Cross.	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	100.00	100.00	0.00	0.00
Decision	C	75.78	100.00	75.78	0.00	0.00
	D	76.40	100.00	76.40	0.00	0.00
Map	C	57.84	95.68	55.68	4.32	5.41
	D	56.22	95.14	53.51	4.86	4.32
Motion	C	89.45	100.00	89.45	0.00	0.00
	D	88.37	100.00	88.37	0.00	0.00
Spatial	C	95.05	95.85	91.47	4.15	3.57
	D	95.39	95.51	91.47	4.49	3.23
Temp. Rec.	C	100.00	100.00	100.00	0.00	0.00
	D	100.00	100.00	100.00	0.00	0.00
Temp. Rel.	C	53.65	100.00	53.65	0.00	0.68
	D	53.42	100.00	53.42	0.00	0.00
Overall	C	**83.95**	**98.53**	**82.78**	**1.47**	1.47
	D	83.55	98.40	82.25	1.60	**1.20**

Gemma Stage C falls from 79.09% with two targets to 49.41% with four, while Qwen Stage C falls from 77.73% to 40.91%. The same pattern is substantially more severe in the zero-shot models (see Table 29). No evaluated system achieves All-G@.50 on the 56 questions requiring six or more targets.

Table 28: Per-family grounding and evidence-supported answering for Qwen3-VL-8B on the seven-family held-out test set. Joint@.50 is target-level joint presence–localization success; All-G@.50 requires all grounding targets associated with a question to succeed; JAGS@.50 additionally requires matched-pair QA success. We report $t = 0.50$. Cross. is Cross-view Recovery; Decision is Evidence-conditioned Decision; Map is Map Relation; Motion is Motion State; Spatial is Spatial Relation; Temp. Rec. is Temporal Recovery; Temp. Rel. is Temporal Relation. **All results are expressed as percentage.**

Family	**Stage**	**Presence** ↑	**IoU@t** ↑	**Joint@t** ↑	**All-G@t** ↑	**JAGS@t** ↑
Cross.	C	87.56	67.05	58.71	25.87	25.87
	D	88.06	66.38	58.46	26.87	26.87
Decision	C	98.77	21.99	22.31	25.47	13.66
	D	98.77	22.78	23.16	24.22	13.04
Map Rel.	C	84.86	85.00	77.57	62.70	37.84
	D	85.68	82.78	77.30	62.70	38.92
Motion state	C	98.92	88.15	87.20	83.41	75.38
	D	99.10	87.67	86.88	83.32	74.21
Spatial Rel.	C	95.58	92.92	92.09	85.83	83.64
	D	94.80	92.42	91.10	84.22	82.26
Temp. Rec.	C	94.83	76.36	72.41	51.72	51.72
	D	94.83	76.36	72.41	55.17	55.17
Temp. Rel.	C	97.37	77.78	75.74	45.89	24.43
	D	97.83	77.36	75.68	46.35	25.80
Overall	C	96.45	**74.01**	**74.71**	**70.04**	**61.12**
	D	**96.49**	73.70	74.46	69.64	60.62

All-G is lower for questions with more targets, although target count also varies with reasoning families and difficulty.

Table 29: All-G@t by number of grounding targets. $t = 0.50$. The two largest buckets contain 2,429 and 506 questions, respectively.

System	**2 targets (%)** ↑	**4 targets (%)** ↑
Gemma-C	79.09	49.41
Gemma-D	80.16	47.83
Qwen-C	77.73	40.91
Qwen-D	77.15	41.30
Gemma zero-shot	48.70	8.50
Qwen zero-shot	54.43	8.50
InternVL zero-shot	0.21	0.00
Molmo zero-shot	0.04	0.00
LLaVA zero-shot	0.33	0.00

H Qualitative Examples and Failure Modes

This section provides qualitative examples for six scene-understanding families. **All examples use the original nuScenes (Caesar et al., 2020) camera frames; DriveAnswer introduces no blur, degradation, or other image-quality modification.**

QUESTION. What is the relationship between the silver minivan and the ego?

STRONG: In the strong configuration, the minivan is **visible** to the model.

WEAK: In the weak configuration, the silver minivan's identity is **not** visible to the model.



Figure 3: **Spatial Relation example.** The strong configuration preserves the visual support needed to establish the target's relation to the ego, whereas the weak configuration withholds critical target evidence while preserving the underlying scene fact.

QUESTION. What is the relationship between the silver minivan and the walkway?

STRONG: In the strong configuration, the minivan **AND** the walkway are **visible** to the model.

WEAK: In the weak configuration, the silver minivan's identity is **not** visible.



Figure 4: **Map Relation example.** The strong configuration jointly supports the queried actor and its relation to mapped scene structure, whereas the weak configuration removes critical visual support required to establish that relation.

H.1 Spatial and Map Reasoning

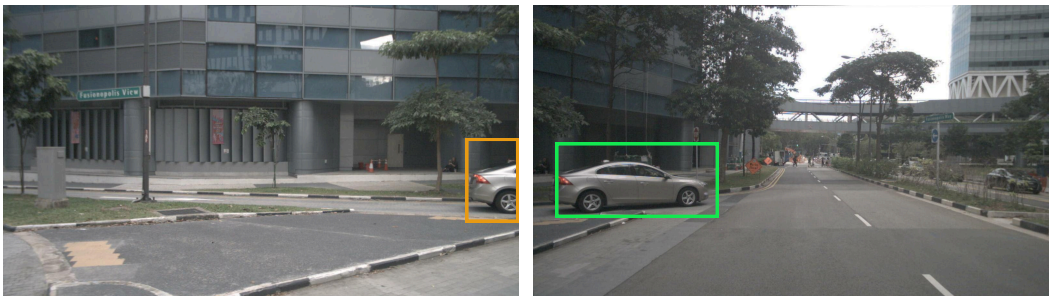
Spatial and map examples illustrate that preserving an object in view does not necessarily preserve the evidence required to establish its relationship to another entity or to mapped scene structure. These examples as illustrated in Figures 3 and 4 are useful for distinguishing target visibility from relation sufficiency.

These examples illustrate two distinct recover mechanisms. Cross-view Recovery restores evidence through a synchronized alternative camera, whereas Temporal Recovery restores evidence through neighboring time steps. in both cases, the anchor observation alone is insufficient for question answering, so successful behavior requires using the additional admitted evidence defined by the each family.

The temporal examples further show why object visibility alone is insufficient. The queried actor may remain identifiable in the weak configuration, while the sequence no longer support answers about its motion or the evolution of its relation to another actor.

QUESTION. *Can the silver sedan's identity be recovered from another camera's view?*

A) STRONG: In the strong configuration, the silver sedan's **identity** is **recoverable** from the cross-view camera.



B) WEAK: In the weak configuration, the silver sedan's **identity** is **not** recoverable from the cross-view camera.



Figure 5: **Cross-view Recovery example.** The anchor observation alone is insufficient; the strong configuration provides synchronized camera view from which the queried actor can be recovered, whereas the weak sequence does not.

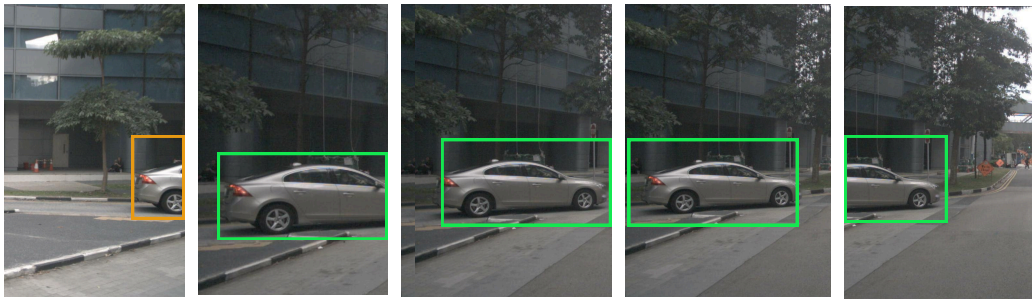
H.2 Cross-view and Temporal Recovery

Recovery examples require that the designated anchor observation be insufficient by itself. Cross-view recovery adds a synchronized camera that restores the missing evidence, while temporal recovery adds a nearby time step. Examples in which the anchor already determines the answer are treated as controls rather than as genuine recovery cases as seen in Figures 5 and 6.

These recovery families test whether the model can use additional admitted observations to reconstruct evidence that is unavailable at the anchor frame. Specifically, Cross-view Recovery restores the missing evidence through synchronized alternative camera(s), whereas Temporal Recovery restores it through consecutive timesteps. The key distinction is that the semantic question and underlying scene are held fixed while the observation set changes. These families primarily focuses on whether the model can make use of additional observations when the evidence required by the question is unavailable at the anchor. A successful model should therefore answer only when the added cross-view or temporal observation restores the evidence needed for the queried inference, rather than relying on the anchor frame alone. These cases separate object visibility at a single observation from the broader question of whether the admitted observation set is sufficient to support the requested inference.

QUESTION. Can the silver sedan's identity be recovered from the temporal evidence?

A) STRONG: In the strong configuration, the silver sedan's identity is **recoverable** from the temporal sequence.



B) WEAK: In the weak configuration, the silver sedan's identity is **not** recoverable from the temporal sequence.

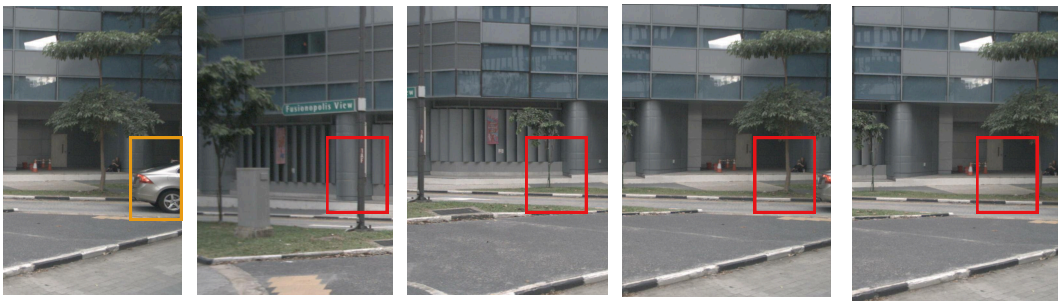


Figure 6: **Temporal Recovery example.** The anchor observation alone is insufficient; the strong configuration provides neighboring temporal observations that recover the missing evidence, whereas the weak sequence does not.

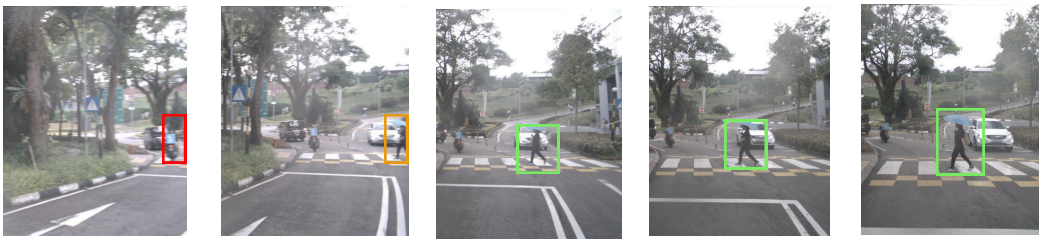
H.3 Motion State and Temporal Reasoning

Motion-state and temporal-relation examples use temporally separated observations of the same scene. A single clear anchor image may be sufficient to recognize an object while remaining insufficient to infer whether it is moving or how its relation to another actor changes over time. In qualitative visualizations in Figures 7 and 8, the anchor frame may be displayed at full emphasis and neighboring temporal frames at reduced visual emphasis to make the temporal evidence structure explicit.

Motion state and temporal relation use temporally separated observations of the same scene. A single clear anchor image may be sufficient to recognize an actor while remaining insufficient to infer whether it is moving or how its relation to another actor changes over time. These temporal families are intentionally different from recovery in that the relevant actor may remain visible throughout the sequence, while the requested property cannot be established from appearance alone. Motion State requires evidence of how an actor changes across time, and Temporal Relation requires evidence of how the relation between actors evolve. The weak configurations therefore preserve partial visual content while removing enough temporal support to make the requested inference cannot be reliably answered. This distinction is important because a model can correctly identify the objects involved while still lacking the evidence needed to infer motion or temporal change. These examples therefore isolate a different failure mode from pure detection or presence errors. The benchmark asks whether the model conditions its answer on the temporal evidence actually available, rather than whether it can recognize the entities appearing in individual frames.

QUESTION. Is the pedestrian stationary or moving throughout the sequence?

STRONG: In the strong configuration, the minivan is **pedestrian's motion becomes visible** to the model with temporal frames.



WEAK: In the weak configuration, the **pedestrian's motion is not recoverable** even with temporal evidence.

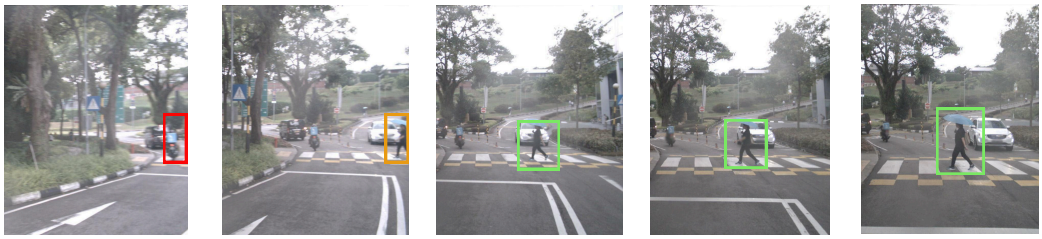


Figure 7: **Motion State example.** The strong sequence provides sufficient temporal evidence to infer the queried actor's motion, whereas the weak sequence preserves partial observations without enough temporal support to determine its motion state.

Taken together, the qualitative examples show that insufficient evidence does not necessarily correspond to an absent object or an empty observation. Across the six families, the strong-weak intervention primarily changes the evidence required for the queried inference rather than simply removing the required object. The examples therefore illustrate the benchmarks' central distinction between observing relevant content and having sufficient evidence to justify an answer. Conversely, the strong condition adds only the evidence needed to support the same underlying semantic question. These examples visualize the core benchmark distinction between observing relevant scene content and having sufficient evidence to support the requested inference.

QUESTION. *What is the distance between the ego and the pedestrian throughout the sequence?*

STRONG: In the *strong* configuration, the pedestrian's changing relation to the ego is **observable** across the temporal sequence.



WEAK: In the *weak* configuration, the **pedestrian's changing location is not recoverable** even with temporal evidence.



Figure 8: **Temporal Relation example.** The strong sequence reveals how the queried relation changes over time, whereas the weak sequence lacks sufficient temporal evidence to establish that change.