# MOSAICDRIVE: TASK-AWARE MULTI-CAMERA REASONING FOR AUTONOMOUS DRIVING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Tool-augmented driving agents typically acquire additional camera views when they are uncertain about what they have already seen. We identify a structural limitation of this design: the confidence that triggers retrieval is computed from evidence the agent already holds, and so carries no information about regions it has never observed. Acquiring more evidence also introduces a second difficulty: a newly retrieved observation can contradict or override a correct prior judgment, so broader acquisition is not monotonically beneficial. We therefore treat acquisition and verification as parts of a single decision. We present **MosaicDrive**, a driving agent framework that derives camera and tool requirements from the question, requires the selected evidence before an answer is produced, and resolves the resulting observations through **Arbiter**, a source-aware verifier that checks claims against available evidence, issues targeted re-queries, and preserves conflicts that cannot be settled. We evaluate on MOSICQA, a multi-camera driving question-answering resource with executed tool traces and scene-disjoint splits, under a protocol that separates coverage from selectivity, correction from abstention, and answer quality from acquisition cost. Controlled camera-policy comparisons and evidence-corruption tests isolate the contribution of task-conditioned acquisition and of explicit verification.

## 1 INTRODUCTION

A driver considering a lane change to the right looks right before moving, whether or not the forward scene appears clear. Vision-language models now support such reasoning in driving agents (Tian et al., 2025; Sima et al., 2024), and driving question-answering benchmarks extend it to surrounding views (Qian et al., 2024; Park et al., 2025; Yu et al., 2025a). Figure 1 shows why surrounding views matter: the front camera shows a scooter rider ahead on the right, while a side camera reveals a postal vehicle alongside. What an agent can conclude is bounded by what it chose to look at, so the acquisition decision precedes, and partly determines, the reasoning that follows.

An agent that has seen only the forward scene is asked whether the adjacent lane is clear, and reports high confidence. Tool-augmented agents make acquisition explicit and condition it on uncertainty: Agent-Driver accesses perception and prediction modules through a tool library (Mao et al., 2024), and DriveAgent-R1 begins from a front view and retrieves more evidence when its reasoning indicates a need (Zheng et al., 2026a). But the triggering signal is computed from evidence the agent already holds, so it is silent on regions never observed: low confidence about what is seen is informative, while an unseen region produces neither confidence nor doubt, and driving VLMs given degraded or missing views still return plausible answers from general knowledge (Xie et al., 2025). Adjusting a retrieval threshold cannot recover this. Task identity, by contrast, is available before acquisition begins, which makes it an admissible basis for coverage requirements where confidence is not.

A side view is retrieved, misread, and overturns a judgment that was already correct; DriveAgent-R1 reports exactly this case and names "resolving informational conflicts" as open work (Zheng et al., 2026a). Acquiring more evidence is therefore not monotonically beneficial: each observation risks a perception error, and a policy that trusts the most recent acquisition converts it directly into an answer error. Views may also be complementary rather than contradictory, or concern different objects, so a verifier must distinguish these cases rather than defaulting to the newest observation



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

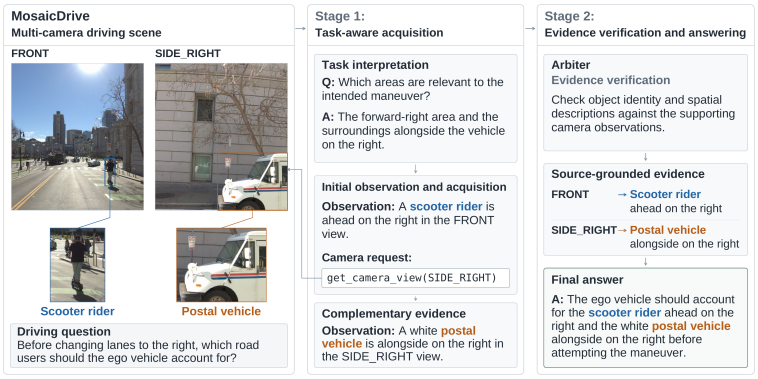


Figure 1: Illustrative MosaicDrive workflow. Stage 1 turns a maneuver-related question into evidence requirements and retrieves a complementary camera view; Stage 2 organizes observations by source and shows Arbiter checking claims before answering. The example does not establish an exhaustive set of road users or the feasibility of a right lane change.

or the prior answer. Acquisition policy and trust policy are consequently not separable: prior work treats them apart, with active perception acquiring without arbitrating among sources and conflict-aware aggregation arbitrating over evidence it did not choose to collect (Liu et al., 2026). Coupling them is what allows broader acquisition to be safe rather than merely broader.

We introduce **MosaicDrive**, a framework that joins task-conditioned acquisition with explicit evidence verification for driving question answering. The acquisition stage derives camera and tool requirements from the question and requires the selected evidence before an answer is produced, so coverage is set by the task rather than by the agent's confidence in what it has already seen. Observations retain their source, so each claim traces to the camera or tool that produced it. The verification stage uses **Arbiter** to check acquired claims, issue targeted re-queries, and preserve conflicts that cannot be settled, under externally established referent identity. Our scope is static multi-camera observation; a single frame does not establish object motion or maneuver safety. We evaluate under a controlled protocol comparing front-only input, front-first adaptive retrieval, all-camera input, and MosaicDrive with and without Arbiter on a shared backbone and tool environment, reporting answer quality and acquisition cost jointly. Our main contributions are as follows.

- We identify a structural limitation of confidence-gated acquisition in multi-camera driving: the trigger signal is computed from acquired evidence and is uninformative about unobserved regions. We show that this is a property of the signal rather than an artifact of threshold choice, and that task identity is an admissible alternative available before acquisition.
- We show that acquisition and verification form a single decision, since additional evidence carries a risk of incorrect override, and introduce Arbiter, which resolves disputed claims through targeted re-query under externally established referent identity.
- We construct MOSICQA and a controlled evaluation protocol that separates coverage from selectivity, correction from abstention, and answer quality from acquisition cost, and we report an explicit implementation status for every component.

## 2 Related Work

**Multi-view driving question answering.** NuScenes-QA derives questions from annotation scene graphs, NuPlanQA emphasizes multi-view spatial reasoning, and DriveLM connects perception, prediction, and planning through graph-structured questions (Qian et al., 2024; Park et al., 2025; Sima et al., 2024); WaymoQA targets safety-critical situations and evaluates surrounding coverage (Yu et al., 2025a). These benchmarks supply the views to the model rather than requiring it to decide which views it needs. Driving VLMs likewise consume a fixed view set (Gopalkrishnan et al., 2024; Wang et al., 2026) and answer from textual priors when views are degraded or absent (Xie et al., 2025) (non-driving view selection: Appendix G).

**Acquisition triggers in tool-augmented agents.** Agent-Driver reaches perception and prediction

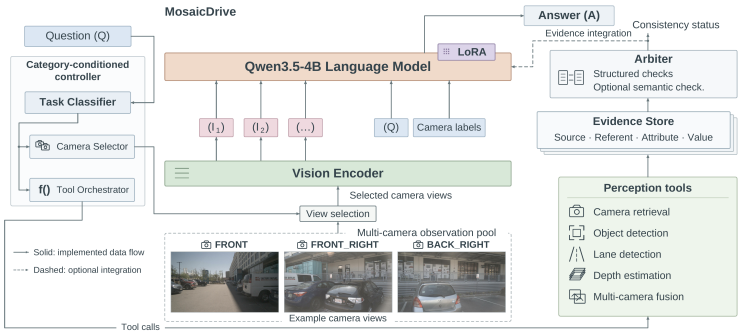


Figure 2: MosaicDrive architecture. A task-conditioned controller selects camera views and perception tools. Selected images are encoded for the LoRA adapted language model, while tool outputs populate an evidence store for Arbiter checks.

Table 1: Implementation status. Trained: learned parameters (designed components carry none, n/a). Data: val, MOSICQA validation (921); trace, executed traces; corr., corruption and real-conflict subsets; SFT/RL, the 2,000-record pools of Section 3.4; acq., training split of the acquisition benchmark (Section 4.3). Stage 3 is trained and evaluated as an offline policy; its live reinforcement-learning phase did not learn to call tools.

| | Task clf. | Cam. sel. | Tool orch. | Evid. store | Arb. T1 | Arb. T2 | Re-query | Ans. SFT | Stg. 2 | Stg. 3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Implemented | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Trained | n/a | n/a | n/a | n/a | n/a | n/a | n/a | ✓ | ✓ | ✓ |
| Evaluated | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Data | val | val | val | trace | corr. | corr. | corr. | SFT | RL | RL |

modules through a tool library (Mao et al., 2024), AgentThink trains tool use with demonstrations and reinforcement learning (Qian et al., 2025), and MTRDrive pairs a dynamic toolkit with retrieved experience (Luo et al., 2025). DriveAgent-R1 is closest to our setting: it starts from a front view and learns when to retrieve more evidence, with a per-call penalty in its reward (Zheng et al., 2026a). In each case the decision to acquire is conditioned on the agent's current reasoning state (Appendix G). That models cannot express calibrated uncertainty about inputs they never received is documented broadly (Kadavath et al., 2022; Whitehead et al., 2022; Zhang et al., 2025), but its consequence for driving questions whose evidence lies outside the initial view has not been measured. MosaicDrive instead derives coverage requirements from the question, before any observation.

**Verification as part of acquisition.** InfoCoordiBridge aligns heterogeneous perception outputs into a conflict-aware summary (Liu et al., 2026), and DriveJudge selects deterministic evaluation functions through contextual reasoning (Sun et al., 2026). Both arbitrate over evidence collected independently of the arbitration mechanism, while active-perception systems acquire without specifying how a disputed observation is treated, as in the failure case of Section 1. Separating the two understates acquisition cost, since it omits the risk that a retrieved observation is wrong and trusted; MosaicDrive couples them, so a disputed claim leads to a further query or an unresolved outcome. Table 20 (Appendix G) compares these systems and reviews the broader knowledge-conflict literature.

## 3 MosaicDrive

MosaicDrive combines task-conditioned evidence acquisition, a source-aware evidence store, and a vision-language answer model (Figure 2). Acquisition mappings determine which cameras and tools are requested, while LoRA (Hu et al., 2022) adapts the model that answers from the selected observations. Because the framework combines designed components with trained ones, Table 1 states the status of each, and every claim in Section 4 is scoped to that table.

### 3.1 Problem Formulation

Given a question $q$ and camera array $\mathcal{C}$, the agent issues requests $a_k$, receives observations $o_k$, and produces an answer $y$ from the resulting history. The design objective balances answer quality against acquisition cost:

$$\max_{\pi} \mathbb{E}_{\pi}\left[R_{\text{ans}}(y, y^*) - \lambda \sum_{k=1}^{K} c(a_k)\right], \qquad \sum_{k=1}^{K} c(a_k) \leq B, \tag{1}$$

with $y^*$ a reference answer, $c$ an acquisition cost, and $B$ the interaction budget. The training stages approach this objective incrementally: Stage 1 addresses the answer term alone; Stage 2 fixes the acquisition condition and optimizes answer quality within it; and Stage 3 optimizes both terms, offline over cached reader outcomes and in a live episode of executable camera and tool actions, with $c$ entering as an explicit penalty (Equation 6). The prior of Section 3.2 bounds $\sum_k c(a_k)$ by construction. The evaluated input scope is static multi-camera imagery, not temporal motion or trajectory prediction.

### 3.2 Task-Conditioned Acquisition Prior

Acquisition is conditioned on task identity rather than on the agent's confidence in evidence it already holds. A keyword-rule classifier maps the question to one of seven categories (lane change, blind spot, intersection, surrounding awareness, parking, emergency response, weather or edge case), and fixed category-to-camera and category-to-tool mappings determine the requests:

$$z = f(q), \qquad \mathcal{C}_z = g(z), \qquad \mathcal{T}_z = h(z). \tag{2}$$

Every question invokes the controller before answer synthesis.

**Camera selection.** The prior applies a fixed category-to-camera mapping, translated to the camera names of each sensor domain. It has no trainable parameters and ignores scene content: two questions of one category receive the same cameras. It asserts that task identity alone, recoverable from the question, already determines a useful acquisition set, and that requiring that set before answering is preferable to confidence-triggered retrieval. It does not assert minimality or sufficiency, which Section 4.4 tests against annotations and by counterfactual removal.

**Tool selection.** The tool prior comes from per-category frequencies in training-split annotations. Operations run through the runtime interface, retaining arguments, responses, and failures; unavailable temporal measurements are excluded rather than replaced by single-frame guesses; and executed responses populate the evidence store before answer synthesis.

### 3.3 Evidence Verification and Answer Generation

**Evidence representation.** An observation is a claim $e = (s, r, f, v, p, t)$, where $s$ identifies the source, $r$ the referent, $f$ the attribute, $v$ its value, $p$ the reported confidence, and $t$ the timestamp. Unstructured descriptions are retained as text, and comparable claims require aligned referents, units, and coordinate conventions.

**Referent resolution.** Two claims are comparable only if they concern the same entity: a red vehicle from one camera and a white vehicle from another is a contradiction only under a shared referent, and is otherwise complementary evidence about two objects. A pair therefore enters conflict checking only when linked by one of three admissible sources: same source re-query, where identity follows from the request; declared identity, where the tool emits a shared track or instance identifier; or geometric association, where calibrated projection places the referents within a disclosed tolerance. Unlinked pairs are routed to a complementary evidence bucket and never reported as conflicts. Association is a separate decision from resolution: a pair wrongly linked and then resolved is an association failure, not a verification success.

**Arbiter.** For numeric claims sharing a referent and field, the structured check is

$$d_f = \max_i v_i - \min_i v_i, \qquad b_f = \mathbb{I}[d_f > \tau_f], \tag{3}$$

with $\tau_f$ an attribute specific tolerance; categorical claims are compared for equality, and a semantic tier checks free text consistency where structured comparison does not apply. On detecting disagreement, the module issues a targeted re-query under an explicit budget, falling back to heuristic down

weighting or an unresolved flag when the re-query does not settle it. A consistency status records that retained claims survived the checks applied; it is not a certificate of factual correctness.

**Answer model** encodes the attached views and generates an answer conditioned on the question, camera labels, and claims retained after verification. Tolerances are conservative fixed values, not fitted (Appendix D).

### 3.4 MOSICQA AND MULTI-SOURCE DATA CONSTRUCTION

MOSICQA connects driving questions to the camera observations and tool interactions used to answer them, and is built on nuScenes imagery and annotations (Caesar et al., 2020).

**Records and coverage.** A candidate record $d_i = (s_i, t_i, q_i, z_i, \widetilde{y}_i, \widetilde{\mathcal{C}}_i, \widetilde{\mathcal{P}}_i)$ pairs a source scene and sample with a question, task category, candidate answer, and proposed camera and tool requirements; the tilde marks annotation proposals rather than verified evidence. The 5,643 records across the seven categories were generated from driving scenes with GPT-4o (OpenAI et al., 2024) and partitioned into 3,924 training, 921 validation, and 798 test records with disjoint source scenes and samples (checked on scene identifiers, since legacy QA identifiers can repeat). These counts describe the candidate pool, not verified answers, and questions requiring motion or future outcomes are excluded unless reformulated and reviewed.

**Execution, review, and pools.** A request becomes an observation only after execution, and the trace retains arguments, observations, source camera references, and failures. Questions are reframed to remove sensor-name instructions, each factual claim is linked to images and source annotations, and source-image integrity, recorded execution, and answer review are tracked separately. The training collection combines selected MOSICQA candidates with AgentThink (Qian et al., 2025) and WaymoQA (Yu et al., 2025a) records into separate 2,000-record SFT and RL pools, each half nuScenes and half Waymo, with no source-scene overlap (Appendix C).

### 3.5 PROGRESSIVE TRAINING

Figure 4 (Appendix E) summarizes the stages. The stages are separated so each contribution can be isolated (Section 4.6); we use LoRA on Qwen3.5-4B.

**Stage 1: evidence-conditioned answer SFT.** With attached images $\mathcal{I}^{\text{att}}$ and synthetic target $\widehat{y}$,

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\mathcal{D}_{\text{SFT}}}\left[\frac{1}{|\widehat{y}|}\sum_t \log p_\theta(\widehat{y}_t \mid q, \mathcal{I}^{\text{att}}, \widehat{y}_{<t})\right]. \tag{4}$$

Only answer tokens contribute; the stage trains answer generation, not camera or tool selection, and its targets are evidence-conditioned synthetic answers rather than verified ground truth.

**Stage 2: forced-condition policy gradient.** Answers are sampled under front-view and task-selected conditions using cached tool observations, isolating the evidence condition from variation in tool responses. For condition $m$, $G$ samples receive a sentence-embedding similarity reward (Reimers & Gurevych, 2019) against a generated reference:

$$r_{m,j} = \cos\left(e(y_{m,j}), e(\widehat{y}^{\text{ref}})\right), \qquad \widehat{a}_{m,j} = r_{m,j} - \tfrac{1}{G}\textstyle\sum_{\ell=1}^{G} r_{m,\ell}. \tag{5}$$

Token cross-entropy is weighted by these within-condition advantages; the implementation omits the clipped importance ratio and reference-policy KL terms of GRPO (Shao et al., 2024) and is a group-centered policy-gradient stage. The reward measures phrasing agreement with a generated reference, not visual correctness, so we score answers with the rubric judge of Section 4.1 and report the rank correlation in Section 4.6; a judge-reward variant adds the grade (Appendix E).

**Stage 3: learned acquisition.** Stage 3 learns which cameras to acquire at a stated cost. Its reward instantiates Equation 1:

$$R = R_{\text{ans}}(y, y^*) - \lambda_c \textstyle\sum_k c(a_k), \tag{6}$$

where $R_{\text{ans}}$ is exact-match correctness on a benchmark whose answers come from annotations, and $c$ assigns separate per-view and per-tool costs (0.05 per extra camera and 0.02 per tool call). We train an offline policy that names up to two extra cameras from the front image and the question, rewarded from a table of cached reader outcomes over all camera-subset choices, and a live policy that emits camera and tool actions over up to three steps. A penalty on unresolved conflicts (Zhou et al., 2026) is not used. DriveAgent-R1 also penalizes each tool call (Zheng et al., 2026a); the difference here is the separation of view and tool costs. Penalizing calls can be gamed by suppressing acquisition altogether (Ma et al., 2026), which the coverage prior of Section 3.2 is intended to prevent. Varying $\lambda_c$ traces the accuracy against acquisition cost frontier of Figure 3.

Table 2: **Camera-policy comparison on 1,719 held-out questions.** Rubric grade (%, 95% scene-level interval); Views and Calls per question (SFT-v2 run).

| Policy | SFT-v2 | Untuned | RL prototype | Views | Calls |
|---|---|---|---|---|---|
| Front-only | 28.0 [25.8, 30.1] | 12.1 [10.3, 14.0] | 28.1 [26.1, 30.1] | 1.0 | 2.7 |
| All-camera | 25.8 [23.6, 27.9] | 20.4 [18.7, 22.1] | 28.8 [26.6, 31.0] | 6.0 | 13.7 |
| Front-first adaptive | 23.7 [21.5, 25.7] | 15.6 [14.0, 17.4] | 26.9 [24.7, 29.2] | 3.1 | 7.6 |
| Matched-budget random | 24.5 [22.4, 26.7] | 15.4 [13.8, 16.9] | 29.7 [27.4, 31.9] | 3.6 | 8.4 |
| Query-all tools (selected cameras) | 23.4 [21.3, 25.5] | 18.6 [17.0, 20.2] | 23.4 [21.3, 25.5] | 3.6 | 33.1 |
| Category-mapped, recency (no Arbiter) | 25.4 [22.9, 27.9] | 18.0 [16.1, 20.0] | 27.9 [25.9, 30.0] | 3.6 | 7.7 |
| **Category-mapped + Arbiter** | 24.8 [22.4, 27.3] | 17.5 [15.6, 19.5] | 28.3 [26.3, 30.4] | 3.6 | 7.7 |

# 4 EXPERIMENTS

We ask whether task-conditioned acquisition improves the accuracy–cost trade-off over front-only and all-camera input, whether acquisition can be learned, whether Arbiter improves conflict handling, and how far the answer metrics can be trusted. Every number below comes from a completed run.

## 4.1 EXPERIMENTAL SETUP

**Data and splits.** Answer quality is measured on all 1,719 held-out MOSICQA questions (921 validation and 798 test records from 74 scenes, none used in training); selection quality uses the 921 validation records. Intervals resample *scenes* (74 for MOSICQA, 142 for the acquisition benchmark), since questions from one scene share images.

**Serving and metrics.** All models run locally behind vLLM with greedy decoding: the answer model (Qwen3.5-4B, untuned or LoRA-adapted), a fixed tool model (Qwen2.5-VL-7B) producing all tool evidence once per question, a Qwen2.5-VL-72B (AWQ) reader, and a local judge (gpt-oss-20b). Answer quality is a reference-based rubric grade (1, 0.5, 0) with embedding similarity as a secondary score. Neither is human-reviewed: references are generated (GPT-4o) and unreviewed, and 72% of answers are graded incorrect (Section 4.6), so acquisition is also tested on a benchmark graded by exact match against nuScenes annotations (Section 4.3). No human driving baseline is reported, since no human evaluation of MOSICQA answers was collected.

## 4.2 CAMERA ACQUISITION POLICY COMPARISON

This comparison separates coverage from verification under one backbone, question set, prompt, decoding setting, and tool implementation. Baselines are front-only input, all-camera input, matched-budget random (as many cameras and tools as the prior), query-all-tools (selected cameras, every tool), and front-first adaptive retrieval in the style of DriveAgent-R1 (Zheng et al., 2026a), where the answer model names the extra views it needs; the category-mapped prior is run with and without Arbiter. Each policy uses three answer models (untuned Qwen3.5-4B, SFT-v2, and the RL prototype), since conclusions differ by training.

For the task-trained models the policies are not separated and the front image alone is as good or better: with SFT-v2, mapped scores $-3.2$ against front-only (paired interval $[-5.1, -1.3]$) and $+0.3$ against matched-budget random ($[-1.7, +2.2]$), and with the RL prototype all differences span zero. The untuned model, which uses the evidence, shows the expected ordering (mapped $+5.3$ over front-only, $[+3.4, +7.2]$, and $+2.0$ over random). What the prior delivers is cost: 3.6 of six cameras and 7.7 tool calls against 13.7 for all-camera, while query-all-tools spends 4.3$\times$ the calls for no gain. Front-first adaptive selection scores no higher, and removing Arbiter changes the grade by at most 0.6 because no natural conflict is detected (Section 4.5). Since task-trained models' grades depend little on images (Section 4.6), this benchmark has limited power to show that a view helps, so acquisition is tested next on one that can.

Table 3: **Acquisition policies on the ground-truth benchmark** (852 test questions, exact-match accuracy %, 95% scene-level interval). *Adaptive*: the 72B reader names up to two extra cameras. *Learned*: offline policy at cost 0.05. *Agentic RL*: live policy after RL.

| Policy | Accuracy | Views |
|---|---|---|
| Front-only | 57.2 [55.3, 59.2] | 1.0 |
| Category-mapped (fixed) | 59.5 [57.5, 61.7] | 4.0 |
| Matched-budget random | 59.4 [57.0, 61.9] | 4.0 |
| All-camera | 61.4 [59.2, 63.6] | 6.0 |
| Adaptive (72B reader chooses) | 67.3 [64.7, 69.8] | 1.9 |
| **Learned,(ours)** | 64.8 [62.3, 67.1] | 1.5 |
| Agentic RL (live, 948 q) | 57.3 [55.3, 59.3] | 1.0 |

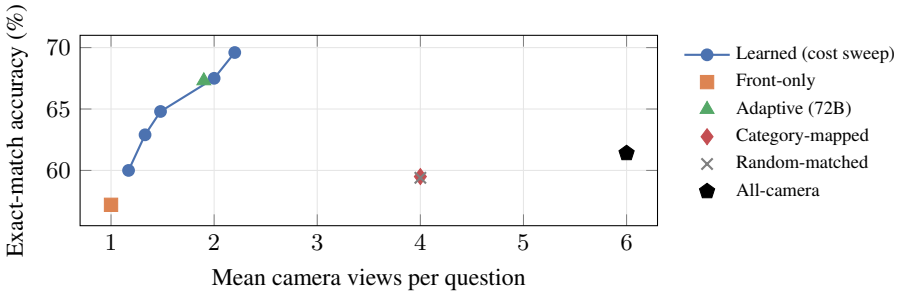


Figure 3: **Accuracy against acquisition cost.** The learned policy traces a curve as its per-view cost rises from 0 to 0.2 (right to left).

### 4.3 Ground-Truth Acquisition Benchmark and Learned Acquisition

Grades against generated references cannot show whether acquiring the *right* camera matters, so we built 1,800 yes/no questions whose answers come from nuScenes 3D annotations, graded by exact match with no judge (Appendix F). Each question names a region of the ego frame, never a camera, and the region maps to the camera covering it. The questions come from 300 scenes absent from every MOSICQA split, split by scene into 948 training and 852 test questions; the reader is Qwen2.5-VL-72B seeing only the acquired images.

Acquisition matters, and the fixed prior does not target it. The reader's recall is 33.0% when the relevant camera was acquired and 7.3% when it was not. The fixed category mapping is $+2.3$ over front-only ($[+0.7, +4.1]$) but indistinguishable from random cameras at the same budget ($+0.1$, $[-2.5, +2.5]$), because categories do not identify the asked region. A policy that reads the question does: the adaptive reader reaches 67.3% at 1.9 views, $+10.1$ over front-only and $+5.9$ over all-camera, though largely by language matching, since questions state the region.

**Learned acquisition (Stage 3).** A Qwen3.5-4B policy names up to two extra cameras from the front image and the question, with reward $\mathbf{1}[\text{correct}] - 0.05 \times$ views looked up from cached reader outcomes over all 16 camera subsets (offline, no tools). Policy gradient and a per-question warm start collapsed to never acquiring; labels from the best mean reward per asked region produced a policy that fits (Appendix F). It reaches 64.8% at 1.5 views: $+7.6$ over front-only ($[+5.9, +9.5]$), $+5.3$ over the fixed mapping, and $+3.4$ over all-camera with a quarter of its views, and raising the cost from 0 to 0.2 moves it monotonically from 2.20 views and 69.6% to 1.17 views and 60.0% (Figure 3). Because questions about a region share a label, it is a learned region-to-camera rule, not a visually grounded policy.

**Tool evidence and live agentic acquisition.** Adding detector output as text to the front and oracle-region cameras raises accuracy from 66.8% to 86.4% ($+19.6$, $[+16.4, +22.8]$), so tool evidence is informative for a reader that uses it. A live policy over `CAMERA`, `TOOL`, and `ANSWER` actions never learned to call the tool under reinforcement learning (0% held-out tool use in 200- and 948-

Table 4: **Camera and tool selection quality** (%), against legacy MOSICQA annotations on the 921 validation records. Query-all is shown as a recall ceiling and excluded from best-result marking.

| | Camera selection | | | Tool selection | | |
|---|---|---|---|---|---|---|
| **Variant** | P | R | F1 | P | R | F1 |
| Random baseline | 44.5 | 44.5 | 44.5 | 23.0 | 23.0 | 23.0 |
| *Query-all (ceiling)* | 43.6 | 100.0 | 59.9 | 23.9 | 98.7 | 37.9 |
| **Fixed mapping (ours)** | **75.3** | **96.7** | **83.5** | **71.7** | **85.0** | **75.7** |

Table 5: **Conflict detection and resolution by Arbiter** (%), on the 14,898 corrupted pairs. Resolved, Flagged, and Missed sum to 100.

| | Detection | | | Outcome | | |
|---|---|---|---|---|---|---|
| **Variant** | P | R | F1 | Resolved ↑ | Flagged | Missed ↓ |
| Recency-trusting (baseline) | – | – | – | 0.3 | 0.0 | 99.7 |
| Tier 1 only | 100.0 | 99.7 | 99.9 | 1.5 | 95.8 | 2.8 |
| **Full Arbiter (ours)** | **100.0** | **99.7** | **99.9** | **79.6** | 0.0 | **20.4** |

question runs), which we attribute to the sparse end-of-episode reward and report as a negative result (Appendix F).

### 4.4 Component-Level Selection Results

Table 4 reports selection against the annotated sets on the 921 validation records; query-all attains maximal recall but the lowest precision. The keyword classifier is correct on 98.3% of the records it classifies at 94.0% coverage (92.4% overall); its rules were tuned on these questions, so this is a consistency check (Appendix B). Removing selected cameras and adding unselected ones on all 1,719 questions shows the selected set beats front-only by 0.122 and that all six cameras add nothing beyond it, while annotation membership predicts necessity poorly (Appendix F).

### 4.5 Conflict Detection and Resolution

**Controlled corruption.** On 29,819 pairs from 20 held-out scenes (14,898 with a 10 m synthetic corruption built from oracle geometry), a policy can resolve correctly (within 3 m), flag unresolved, or commit wrongly (Table 5). The re-query is a disclosed simulation that fails with probability 0.2, so the full policy's resolution rate is about one minus that by construction (Appendix D).

Recency almost never lands within tolerance, reproducing the failure DriveAgent-R1 reports (Zheng et al., 2026a); structured checks abstain on 95.8%, down-weighting alone is 98.2% incorrect, and re-querying turns detection into resolution.

**Real cross-camera conflicts.** The end-to-end pipeline detected no conflict in 67,041 runs because referents are camera-scoped, so we built a real set from nuScenes annotations: objects visible in two adjacent cameras, oracle identity, and independent detector runs on each (403 matched pairs; Appendix C). The cameras disagree by more than 3 m on 171 (42.4%); there recency is right 32.2% of the time and wrong 67.8%, while Arbiter abstains on 90.1% and is wrong only 1.2% of the time it commits (Table 18, Appendix F). Live re-query is a no-op with a deterministic tool, so its benefit is shown only in simulation.

### 4.6 Ablations and Metric Validity

**Architecture and grounding controls** (full tables in Appendix F). Removing Arbiter changes the SFT-v2 grade by $-0.6$ ($[-1.6, +0.5]$), and the oracle category is indistinguishable from the predicted one. Six raw images without perception tools outscore the tool pipeline (31.7 against 24.8). Removing all images raises the SFT-v2 grade by $2.4$ ($[+0.8, +3.8]$) and another scene's images cost $0.5$ ($[-2.2, +1.1]$), because questions already state the decisive facts and small-VLM tool evidence can contradict them; the grade does not measure visual grounding for task-trained models.

**Reward and metric validity.** Over 66,904 answers, the Spearman correlation between embedding similarity and the rubric grade is 0.166 (95% interval 0.133–0.197), and similarity ranks the better answer higher in only 60.8% of 352,458 pairs; the grade falls with length ($\rho = -0.18$), and a second judge agrees on 79.0% ($\kappa = 0.59$). The RL prototype's $+3.4$ over SFT-v2 becomes $-1.1$ ($[-4.1, +2.1]$) under a length control, so we do not claim it improves correctness. Likewise, task-trained models' higher grades than zero-shot VLMs against GPT-4o-written references are not a capability ranking (Appendix F).

## 5 Conclusion and Limitations

MosaicDrive couples task-conditioned evidence planning with source-aware evidence verification for driving QA, addressing a structural limit of confidence-gated acquisition: the trigger is silent about regions never observed. The controlled evaluation supports three findings. Task-conditioned selection agrees with annotated camera and tool sets far better than random or query-all selection (camera F1 83.5% against 44.5% and 59.9%) while using 60% of the cameras and 56% of the tool calls of all-camera input. Acquisition can be learned at a stated cost: an offline policy reaches 64.8% at 1.5 views on an exact-match benchmark, $+7.6$ over the front camera and $+3.4$ over all six cameras with a quarter of the views, and a live object detector raises accuracy from 66.8% to 86.4%. Arbiter turns detection into resolution: under injected corruption it resolves 79.6% of conflicts against 0.3% for recency, and on real cross-camera disagreements it is wrong 1.2% of the time it commits, against 67.8%. The evidence has limits. The fixed category prior does not beat random cameras on the exact-match benchmark, the learned policy is a region-to-camera rule rather than a visually grounded one, and live reinforcement learning did not learn to call tools. Arbiter's resolution rate depends on a simulated re-query, real-study identity is oracle-provided, and MOSICQA answers are graded by proxies. Static imagery does not establish motion or maneuver safety.

## References

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. DRAGged into conflicts: Detecting and addressing conflicting sources in search-augmented LLMs. *arXiv preprint arXiv:2506.08500*, 2025.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic reinforced policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Akshay Gopalkrishnan, Ross Greer, and Mohan Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *arXiv preprint arXiv:2403.19838*, 2024.

Jack Hong, Chenxiao Zhao, ChengLIn Zhu, Weiheng Lu, Guohai Xu, and XingYu. Deepeyesv2: Toward agentic multimodal model. In *The Fourteenth International Conference on Learning Representations*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, Xiaoye Qu, Dongrui Liu, Lizhen Cui, and Yuntao Du. Benchmarking multimodal knowledge conflict for large multimodal models. *arXiv preprint arXiv:2505.19509*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Shuo Liu, Lei Shi, Haowen Liu, Jing Xu, Yufei Gao, and Yucheng Shi. Information coordination as a bridge: A neuro-symbolic architecture for reliable autonomous driving scene understanding. *arXiv preprint arXiv:2605.04475*, 2026.

Ziang Luo, Kangan Qian, Jiahua Wang, Yuechen Luo, Jinyu Miao, Zheng Fu, Yunlong Wang, Sicong Jiang, Zilin Huang, Yifei Hu, et al. Mtrdrive: Memory-tool synergistic reasoning for robust autonomous driving in corner cases. *arXiv preprint arXiv:2509.20843*, 2025.

Da Ma, Ziyue Yang, Hongshen Xu, Haotian Fang, Kai Yu, and Lu Chen. Empowering LLM tool invocation with tool-call reward model. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *First Conference on Language Modeling*, 2024.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*, 2025.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, et al. GPT-4o System Card, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8066–8076, October 2025.

Thinh Pham, Nguyen Phan Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. SealQA: Raising the bar for reasoning in search-augmented language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, Yining Shi, He Zhe Lim, Li Liu, Tianbao Zhou, Hongyi Wang, Huang Yu, Yifei Hu, Guang Li, Guang Chen, Hao Ye, Lijun Sun, and Diange Yang. AgentThink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10663–10682, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.564.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. NuScenes-QA: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024. doi: 10.1609/aaai.v38i5.28253.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger Luo, Ping, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, 2024. doi: 10.1007/978-3-031-72943-0_15.

Xinglong Sun, Kevin Xie, Jenny Schmalfuss, Despoina Paschalidou, Xiuming Zhang, Sanja Fidler, Kashyap Chitta, and Jose M Alvarez. Drivejudge: Rethinking autonomous driving evaluation with vision-language models. *arXiv preprint arXiv:2606.17362*, 2026.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726. PMLR, 06–09 Nov 2025.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Jie Wang, Guang Li, Zhijian Huang, Chenxu Dang, Hangjun Ye, Yahong Han, and Long Chen. VGGDrive: Empowering Vision-Language Models with Cross-View Geometric Grounding for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2602.20794.

Ziyue Wang, Chi Chen, Fuwen Luo, Yurui Dong, Yuanchi Zhang, Yuzhuang Xu, Xiaolong Wang, Peng Li, and Yang Liu. ActiView: Evaluating active perception ability for multimodal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7605–7633, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.376.

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 (ECCV)*, 2022.

Chen Henry Wu, Neil Kale, and Aditi Raghunathan. Why foundation models struggle with cross-modal context. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6585–6597, 2025. doi: 10.1109/ICCV51701.2025.00621.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

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

Sicheng Yu, CHENGKAI JIN, Huanyu Wang, Zhenghao Chen, Sheng Jin, ZHONGRONG ZUO, XU XIAOLEI, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Framevoyager: Learning to query frames for video large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, XianPeng Lang, and Hang Zhao. Driveagent-r1: Advancing VLM-based autonomous driving with active perception and hybrid thinking. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Xinyu Zhou, Chang Jin, Carsten Eickhoff, Zhijiang Guo, and Seyed Ali Bahrainian. When silence is golden: Can LLMs learn to abstain in temporal QA and beyond? In *The Fourteenth International Conference on Learning Representations*, 2026.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. ACTIVE-o3: Empowering MLLMs with Active Perception via Pure Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2505.21457.

# A USE OF LARGE LANGUAGE MODELS

**Models inside the evaluated system.** Several models are components of the pipeline rather than tools used to write the paper, and each is named where it is used. GPT-4o (OpenAI et al., 2024) proposed the questions, candidate answers, and camera/tool annotations of the legacy MOSICQA collection (Section 3.4). Qwen3.5-4B, adapted with LoRA (Hu et al., 2022), is the answer model. The perception tools of Appendix B are backed by a vision-language model reached through an OpenAI-compatible endpoint, and gpt-4o-mini is the model evaluated in the tool-access pilot of Appendix F. A locally served gpt-oss-20b provides the judge grade in the judge-reward variant of Stage 2 (Appendix E), and the sentence-embedding model all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) provides the similarity reward. A large language model assistant was used for editing the manuscript text.

# B PERCEPTION TOOLKIT AND TASK-CONDITIONED MAPPINGS

**Tool registry.** The runtime exposes ten tools (Table 6). Every tool takes a camera name from the six nuScenes channels (`CAM_FRONT`, `CAM_FRONT_LEFT`, `CAM_FRONT_RIGHT`, `CAM_BACK`, `CAM_BACK_LEFT`, `CAM_BACK_RIGHT`) and a scene or sample token that resolves to real imagery. `cross_camera_verify` is registered but is a disclosed stub that returns `not_implemented` rather than a fabricated agreement, and it appears in no task mapping.

Table 6: **Tool registry.** Purpose and principal arguments of the ten registered tools.

| Tool | Purpose | Principal arguments |
|---|---|---|
| `get_camera_view` | Returns the raw image path for a camera | camera, temporal |
| `detect_objects` | Detects objects of requested types with distance and relative position | camera, types |
| `detect_lanes` | Lane count, ego lane, and availability of left and right lanes | camera |
| `estimate_depth` | Distance to an object or image region with a coarse confidence | camera, region |
| `estimate_velocity` | Object velocity estimate, single-frame or over a window | camera, object_id, window |
| `fuse_cameras` | Combined summary across several views | cameras |
| `detect_traffic_signs` | Detects and interprets traffic signs and signals | camera |
| `visual_qa` | Free-form question about a camera image | camera, question |
| `query_memory` | Past observations from the agent's memory | query |
| `cross_camera_verify` | Intended cross-view consistency check (stub) | cameras |

**Call format and output schema.** Requests are function calls with named arguments, for example

```
detect_objects(camera="CAM_BACK_LEFT", types=["vehicle",
"pedestrian", "cyclist"], scene_token=tok)
estimate_depth(camera="CAM_FRONT", region="center",
scene_token=tok)
fuse_cameras(cameras=["CAM_FRONT", "CAM_FRONT_LEFT"],
scene_token=tok)
```

and each response is stored as one or more claims $e = (s, r, f, v, p, t)$ (Section 3.3). `detect_objects` returns a list of records with a `type`, a `distance_m`, and a coarse relative position; in the real-conflict study (Appendix D) the reported distances take only eleven distinct values (3, 5, 8, 9, 10, 12, 15, 20, 25, 30, and 40 m), so tool distances are coarse by construction.

**Task-conditioned mappings.** Table 7 lists the camera and tool sets applied to each of the seven categories. The camera mapping is a fixed, hand-specified table with no trainable parameters; it is the same table used by the deterministic-camera baseline of the original evaluation harness. The tool mapping is derived from data: a tool is included for a category when it appears in at least

30% of that category's training-split tool annotations (373 to 697 training records per category), and `get_camera_view` is included for every category because looking precedes any other operation. Questions the classifier does not recognize fall back to all six cameras and `get_camera_view` only.

Table 7: **Fixed camera and tool mappings by category.** F: front, FL: front left, FR: front right, B: back, BL: back left, BR: back right.

| Category | Cameras | Tools |
|---|---|---|
| Lane Change Safety | F, FL, FR | `get_camera_view`, `estimate_velocity` |
| Blind Spot Detection | B, BL, BR | `get_camera_view`, `detect_objects`, `fuse_cameras`, `estimate_velocity` |
| Intersection Navigation | F, FL, FR | `get_camera_view`, `detect_objects`, `detect_traffic_signs` |
| 360° Situational Awareness | all six | `get_camera_view`, `fuse_cameras` |
| Parking Maneuvers | B, BL, BR | `get_camera_view`, `estimate_depth`, `fuse_cameras` |
| Weather/Edge Cases | F, FL, FR | `get_camera_view`, `detect_lanes`, `estimate_depth` |
| Emergency Response | F, B, FR | `get_camera_view`, `detect_objects`, `fuse_cameras` |

**Task classifier.** The classifier is an ordered list of keyword rules over the lower-cased question; the first category with a matching keyword wins and an unmatched question returns no category. Rule order matters where vocabularies overlap. Lane-change questions often mention a blind spot as a checked sub-step, so distinctive lane-change phrases (*target lane*, *gap in the*, *gap size*, *trailing vehicle*, *merge into*) are tested before the blind-spot rules, and the ambiguous phrase *lane change* is used only as a last fallback. The keyword lists and rule order were derived from per-category word frequencies on the 921 validation questions themselves, and a later extension added phrases seen in WaymoQA and AgentThink templates. The figures below are therefore measured on the data the rules were tuned on, and are best read as a consistency check of the rule set rather than as held-out accuracy. On those questions the classifier predicts a category for 866 (coverage 94.0%) and is correct on 98.3% of them (92.4% over all 921, counting an unmatched question as wrong). The 15 errors fall into five (true → predicted) pairs: Blind Spot Detection → Lane Change Safety (6), Weather/Edge Cases → Intersection Navigation (4), 360° Situational Awareness → Intersection Navigation (3), and 360° Situational Awareness → Weather/Edge Cases (1) or Lane Change Safety (1).

## C MOSICQA Construction Details

**Splits.** Table 8 gives the per-category composition of the three partitions. Source scenes are disjoint across partitions (169, 37, and 37 scenes, 243 in total), and overlap is checked on scene identifiers rather than on the legacy QA identifiers, which can repeat across partitions.

Table 8: **MOSICQA composition.** Records per category in each partition.

| Category | Train | Val | Test | Total |
|---|---|---|---|---|
| Lane Change Safety | 587 | 120 | 155 | 862 |
| Blind Spot Detection | 632 | 141 | 136 | 909 |
| Intersection Navigation | 583 | 115 | 121 | 819 |
| 360° Situational Awareness | 697 | 132 | 139 | 968 |
| Parking Maneuvers | 569 | 150 | 106 | 825 |
| Weather/Edge Cases | 373 | 146 | 61 | 580 |
| Emergency Response | 483 | 117 | 80 | 680 |
| Total records | 3,924 | 921 | 798 | 5,643 |
| Source scenes | 169 | 37 | 37 | 243 |

**Question reframing and evidence review.** An acquisition benchmark must not name the desired cameras in the question. The protocol removes explicit sensor-name instructions while retaining driving intent and spatial language, and removes unsupported numerical distances or object assertions. Evidence review then links each factual claim to images and, where applicable, source annotations; calibration and projection identify candidate supporting views, but projection into an image does not establish visibility under occlusion. Source-image integrity, recorded execution, and answer review are tracked separately, since a record with valid images and successful tool calls can still carry an incorrect answer. Unsupported claims are corrected or removed rather than accepted through agreement. Questions requiring motion or future outcomes are excluded from the static subset unless reformulated and reviewed.

**Executed records.** An annotated request becomes an observation only after execution, $o_{i,k} = \mathrm{Execute}(a_{i,k}; \mathcal{I}_{s_i,t_i})$, with the trace retaining arguments, observations, source camera references, and failures. An answer generator then receives the collected claims and up to four selected images, yielding an evidence-conditioned synthetic answer; conditioning on executed evidence provides traceability but does not certify the answer. The 4,000-record SFT and RL total counts unique source-linked instances, not epochs, tool calls, or rollouts. Selected AgentThink records come from its released test pool and are designated training-exposed, so overlapping scenes are ineligible for held-out claims.

**Training pools.** The two training pools are assembled from three sources and allocated by domain so that each pool holds 1,000 nuScenes-domain and 1,000 Waymo records (Table 9). At most six questions are drawn from any single frame. The SFT and RL pools share no source scene. MOSICQA records enter the pools only from training-split scenes: none of the scenes behind the SFT or RL pool records appears in the validation or test partitions, and the AgentThink records likewise fall outside both. The SFT pool draws on 92 MOSICQA and 50 AgentThink scene groups, and the RL pool on 74 and 44.

Table 9: **Training pool composition.** Records by source dataset.

| Pool | MOSICQA | AgentThink | WaymoQA | Total |
|---|---|---|---|---|
| SFT ($\mathcal{D}_{\mathrm{SFT}}$) | 638 | 362 | 1,000 | 2,000 |
| RL ($\mathcal{D}_{\mathrm{RL}}$) | 616 | 384 | 1,000 | 2,000 |

## D ARBITER: POLICY, PROTOCOLS, AND EXTENDED RESULTS

**Tier 1 check and resolution policy.** For claims sharing a referent and field, disagreement is declared when the spread of numeric values exceeds a field-specific tolerance, and when categorical values are not all equal. Tolerances are 3 m for `distance_m`, 10 km/h for `speed_kmh`, and exact match for `num_lanes` and for any numeric field without a listed tolerance. They are conservative starting values, not fitted. On a detected disagreement the policy proceeds in strict order. (1) If a re-query budget and callback are supplied, one fresh claim replaces the most recent claim and the pair is re-checked; agreement is recorded as *re-queried accepted*. (2) Otherwise numeric claims are blended by their mean and categorical claims resolve to the claim with the higher reported confidence, breaking ties by recency; both are marked *down-weighted*, so a categorical down-weight is never indistinguishable from agreement. (3) If a numeric spread still exceeds an outer band of three times the tolerance (9 m for distance), the conflict is flagged *unresolved* and no value is returned. Tier 2 handles free-text claims through an injected judge callback and reports every multi-claim case as unresolved when none is supplied. The studies below contain structured distance claims only, so they exercise Tier 1 and the re-query step.

**Synthetic corruption protocol.** For every real agent at every keyframe of 20 held-out nuScenes scenes, a first claim is the oracle distance from the ego vehicle and a second claim is either the same value (a clean pair) or the value shifted by 10 m along the agent's bearing with a random sign and a 0.5 m floor (a corrupted pair), each with probability one half. This yields 29,819 pairs (14,921 clean, 14,898 corrupted). No tool is called: the claims are computed from annotation geometry, so the study measures the resolution logic under known ground truth. The recency policy accepts the second claim. A resolution is scored correct when its value lies within 3 m of the true distance. The re-query is simulated: with probability 0.8 it returns the true distance plus Gaussian noise of

standard deviation 1.5 m, and with probability 0.2 it returns the true distance shifted by 8 m. A failed re-query therefore leaves a 8 m disagreement that falls inside the 9 m outer band, is blended to a value 4 m from the truth, and is scored wrong, so the Resolved rate of the full policy is close to one minus the modelled failure rate by construction (79.6% observed). It should be read as a property of the resolution logic under this assumption, not as a measured success rate of live re-querying.

**Four-way outcome breakdown.** Table 10 separates the outcomes that Table 5 groups. Detection is identical for the two Arbiter variants (TP 14,855, FP 0, FN 43, TN 14,921) because only the resolution step differs. The 43 undetected corrupted pairs are those whose corrupted value stays within the 3 m tolerance; they are also the only pairs on which the recency policy is correct. In Table 5, *Missed* is the sum of the undetected and the wrongly resolved columns.

Table 10: **Outcome breakdown on the 14,898 corrupted pairs.** Counts, with the share of corrupted pairs in parentheses.

| Policy | Undetected | Flagged unresolved | Resolved correct | Resolved wrong |
|---|---|---|---|---|
| Recency-trusting | – | – | 43 (0.3%) | 14,855 (99.7%) |
| Tier 1 only | 43 (0.3%) | 14,265 (95.8%) | 220 (1.5%) | 370 (2.5%) |
| Full Arbiter | 43 (0.3%) | 0 (0.0%) | 11,857 (79.6%) | 2,998 (20.1%) |

**Worked examples.** Table 11 shows the first three corrupted pairs from the hand-inspection sample of the full policy. In the first two the fresh claim agrees with the retained one and the answer is within 3 m of the truth while the recency value is 10 m off. The third is a down-weighted resolution: the blended value is 4.0 m from the truth and is scored wrong, the failure mode described above.

Table 11: **Hand-inspected corrupted pairs** (distances in meters).

| Truth | Recency value | Arbiter value | Resolution | Arbiter error |
|---|---|---|---|---|
| 37.79 | 47.79 | 37.53 | re-queried accepted | 0.26 |
| 20.93 | 30.93 | 19.94 | re-queried accepted | 0.99 |
| 45.72 | 35.72 | 49.72 | down-weighted | 4.00 |

**Real cross-camera disagreement.** Table 18 scores the matched items of the real conflict subset (Appendix C) against the annotated distance, comparing the recency policy with Tier 1 and no re-query. Where the cameras disagree, recency is correct on 32.2% and wrong on 67.8%; Tier 1 abstains on 90.1% (154 items) and is wrong on 1.2% (2 items). It commits on the remaining 17 items, resolving 15 correctly and 2 incorrectly. Where the cameras agree the two policies return the same value and are identical. A live re-query on the same static image changes nothing, because the tool is deterministic at temperature 0 and returns the same detection. That a majority of agreeing pairs are still wrong reflects detector error shared by both views, which no cross-view consistency check can reveal.

## E Training Details and Answer Prompt

**Stage 1: answer SFT.** The base model is Qwen3.5-4B. LoRA uses rank 16, scaling $\alpha = 32$, and dropout 0.05 on the attention, MLP, and linear-attention projection layers, giving 38,993,920 trainable parameters, 0.85% of the 4.58B total. The run uses the 2,000 records of $\mathcal{D}_{\mathrm{SFT}}$ for one epoch with gradient accumulation of 8, for 250 optimizer steps. The running mean of the training loss falls from 0.90 at the first logged step to 0.50 at the end of the epoch, which took about 6.5 hours on one GPU. Loss is reported only as an optimization diagnostic; whether the adapter improves answers is a separate downstream question.

**Stage 2: forced-condition policy gradient.** For each prompt and each of the two forced evidence conditions, front view only and tool-enabled, $G = 4$ answers are sampled with up to 200 new tokens. Tools are not re-executed inside the loop: the tool-enabled condition receives the images and the real evidence claims gathered when the record was generated, and the front-only condition receives the front image and only the claims that came from the front camera. Both conditions are scored against the same reference. Advantages are the rewards centered within each condition's group of four,

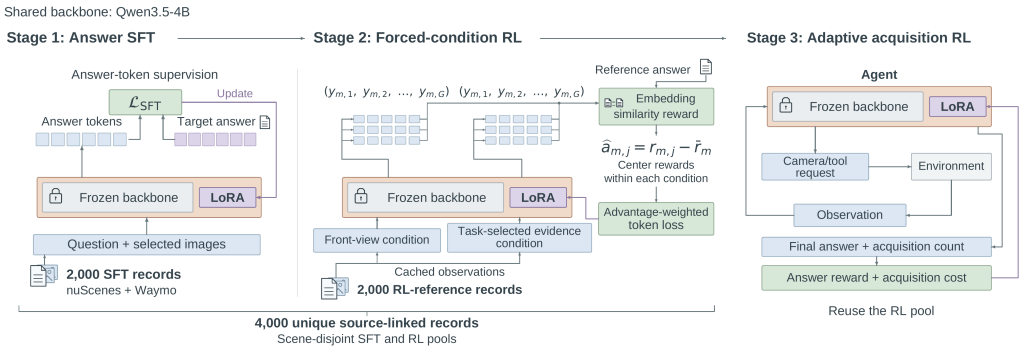


Figure 4: MosaicDrive training framework. Stage 1 adapts answer generation on 2,000 generated records. Stage 2 is a forced condition policy gradient stage on cached observations, with embedding similarity rewards centered within each condition. Stage 3 learns acquisition at a stated cost, offline from cached reader outcomes and live over camera and tool actions, on the training split of the ground-truth acquisition benchmark. The SFT and RL pools contain 4,000 unique source-linked records with disjoint scenes.

rollouts with zero advantage are skipped, and the update weights the token cross-entropy of each sampled answer by its advantage; the clipped importance ratio and reference-policy KL term of GRPO are not implemented. Two reward definitions were run on the RL pool. The similarity variant scores an answer by its cosine similarity to the generated reference under all-MiniLM-L6-v2. The judge variant scores an answer by the grade of a locally served gpt-oss-20b judge plus 0.25 times that similarity, after dropping references that hedge (1,951 of 2,000 references kept), and was run on 100 prompts in about 3.9 hours. Mean training reward under the judge variant was 0.611 in the front-only condition and 0.621 in the tool-enabled condition. Because similarity to a generated reference rewards phrasing rather than visual correctness, the judge variant was introduced to tie the reward to a correctness grade.

**Stage 3: adaptive acquisition.** The implemented policy acts in episodes of at most three actions. At each step it sees the question and everything acquired so far and emits one line: `CAMERA <cam>` opens another camera, `TOOL detect_objects <cam>` runs the detector on an opened camera and adds its output as evidence, and `ANSWER` stops, after which a fixed reader (Qwen2.5-VL-72B) answers from the images and evidence acquired; an episode that reaches the action limit is forced to answer. The reward is exact-match correctness minus 0.05 per additional camera, minus 0.02 per tool call, and minus 0.2 for an unparsable action; no penalty on unresolved conflicts is used. Training is a supervised warm start on a region rule, whose label is the subset of cameras with the best mean reward for each asked region, followed by group-relative policy gradient with group size 4 on live episodes. It is run on a ground-truth-graded acquisition benchmark of 1,800 yes/no questions (948 training, 852 test) whose answers come from nuScenes 3D annotations and depend on which camera is opened. Each question names a region of the ego frame, never a camera, so the relevant camera is the one covering that region and grading is exact match with no judge.

**Compute.** All runs use NVIDIA RTX 6000 Ada GPUs (48 GB). Models are served with vLLM behind OpenAI-compatible endpoints, and adapters are selected by name at the endpoint.

**Answer prompt.** One prompt builder is shared by SFT, RL rollouts, and every evaluation harness, which removes any train-to-inference prompt mismatch by construction. With images and evidence attached the prompt is:

```
You are answering a driving-scene question using the camera images
attached and the collected evidence below.

Camera legend (in the order attached):
{camera legend}

Question: {question}
```

```
Collected evidence:
- {referent} | {field}={value}
...

Give a concise, direct answer. Commit to what the images and
evidence support; say something is unavailable only if it truly
cannot be determined from them.
```

The image-only arms omit the evidence block and the phrase referring to it, and the text-only control omits the images and the camera legend. Evidence lines are deduplicated to one line per referent and field, keeping the last value when no conflict was detected and the Arbiter-resolved value otherwise.

# F Extended Results

This section gives the full tables behind the summaries of Section 4. Every interval is a 95% interval from resampling scenes, and every number comes from a completed run.

**Camera-policy comparison, full statistics.** For the task-trained models the policies are not separated, and the front image alone is as good or better. With SFT-v2 the mapped policy scores $-3.2$ against front-only (paired scene-level interval $[-5.1, -1.3]$), $-1.0$ against all-camera ($[-2.4, +0.4]$), and $+0.3$ against matched-budget random ($[-1.7, +2.2]$); with the RL prototype the differences are $+0.2$, $-0.6$, and $-1.1$, all spanning zero. The untuned model, which uses the evidence, shows the expected ordering: mapped is $+5.3$ over front-only ($[+3.4, +7.2]$) and $+2.0$ over matched-budget random ($[0.0, +4.2]$), though $-2.9$ below all-camera. What the prior delivers is cost: 3.6 of six cameras (60%) and 7.7 tool calls against 13.7 for all-camera (56%), while query-all-tools spends $4.3\times$ the calls for no gain. Front-first adaptive selection uses about as many views as the mapping and scores no higher. Removing Arbiter changes the grade by at most 0.6 points because the deployed pipeline detects no natural conflict (Section 4.5). Since the grade of task-trained models depends little on their images (Section 4.6), this benchmark has limited power to show that a particular view helps, so acquisition is tested next on one that can.

**Acquisition benchmark construction.** Grades against generated references cannot show whether acquiring the *right* camera matters, so we built yes/no questions whose answers come from nuScenes 3D annotations and are graded by exact match, with no judge. Each question names a region of the ego frame, never a camera ("Is there a pedestrian to the rear left of the ego vehicle within 25 meters?"), and the region maps to the camera that covers it. Positives are annotated objects of the asked class in the region, at least 60% visible with lidar support; negatives have none in the region or within a $10^\circ$ margin. The 1,800 questions come from 300 scenes absent from every MOSICQA split and training pool, are balanced between Yes and No, and are split by scene into 948 training and 852 test questions (142 scenes). The reader is Qwen2.5-VL-72B seeing only the acquired images, so the camera choice is the only thing that varies. Positives are annotated objects of the asked class in the region and range, at least 60% visible with lidar support and at least 90 pixels tall in the region's camera; negatives are questions where no object of that class lies in the region or within a $10^\circ$ margin of its boundary within 25 m, so a "No" is never a hidden miss. Each sampled frame contributes one positive and one negative, and the split is by scene (an earlier split by frame leaked scenes between splits and was discarded).

**Answering models under identical inputs.** Table 12 compares answering models under the same images, camera labels, tool evidence, prompt, and greedy decoding on all 1,719 held-out questions. Tool evidence is produced once per question by one fixed tool model and reused for every answerer, so differences reflect how each model uses the same evidence. Each model runs in three arms (front image only; the category-selected cameras without tools; the same cameras with tools), and the effect of tools is the third minus the second arm. Zero-shot models receive no task training, whereas MosaicDrive's answer models are trained on generated task data, so the task-trained rows are not a like-for-like ranking. Closed-source models (GPT-4.1, GPT-5, Gemini 1.5, Claude 3.5) were not evaluated because the available API accounts had no usable credit, and LLaVA-OneVision and InternVL2.5 could not be loaded in our serving stack, so no result here compares with them. Task-trained models score 24–33%, above zero-shot Qwen2.5-VL-72B (8.6–20.5) and the untuned 4B (11.2–19.1), but adapters trained on GPT-4o-style targets are graded against GPT-4o-written references, so style agreement contributes. Tool evidence helps both zero-shot Qwen models ($+4.8$

for the 72B model, $+5.9$ for the untuned 4B) and the 3B model ($+5.5$), and lowers the task-trained ones ($-2.1$ for SFT-v2, $-8.5$ for the RL prototype); for the previous-generation Qwen2-VL-72B the effect is $-1.5$ ($[-3.1, +0.2]$). In a separate 40-sample pilot with gpt-4o-mini on held-out records, embedding similarity rose from 56.4% without tools to 80.9% with tools, a 43% relative gain. Scale predicts less than expected: the 3B model beats the 72B model front-only (13.0 against 8.6), whereas the 72B model leads once cameras are added.

Table 12: **Answering models under identical inputs** (rubric grade, %, 1,719 questions, 74 scenes). $\Delta$ tools is the paired difference (cameras and tools) minus (cameras, no tools).

| Model | Training | Front | Cams | +tools | $\Delta$tools |
|---|---|---|---|---|---|
| *General VLMs (zero-shot)* | | | | | |
| Qwen2.5-VL-3B | zero-shot | 13.0 [11.4, 14.6] | 11.9 [10.5, 13.3] | 17.3 [15.8, 18.8] | $+5.5$ [$+3.6$, $+7.3$] |
| Qwen2-VL-72B (AWQ, prev. gen.) | zero-shot | 15.6 [13.9, 17.5] | 17.4 [15.5, 19.2] | 16.0 [14.6, 17.3] | $-1.5$ [$-3.1$, $+0.2$] |
| Qwen2.5-VL-72B (AWQ) | zero-shot | 8.6 [7.1, 10.1] | 15.7 [14.0, 17.4] | 20.5 [19.0, 22.1] | $+4.8$ [$+3.0$, $+6.7$] |
| Qwen3.5-4B (untuned) | zero-shot | 19.1 [17.4, 20.9] | 11.2 [9.9, 12.4] | 17.1 [15.3, 19.0] | $+5.9$ [$+4.2$, $+7.7$] |
| *Same backbone, task-trained* | | | | | |
| Qwen3.5-4B + SFT-v2 (ours) | task-trained | 25.4 [23.2, 27.4] | 26.4 [24.1, 28.7] | 24.3 [22.1, 26.6] | $-2.1$ [$-4.1$, $+0.0$] |
| Qwen3.5-4B + SFT-v2 + RL | task-trained | 33.4 [31.2, 35.6] | 34.9 [32.5, 37.3] | 26.4 [24.4, 28.4] | $-8.5$ [$-10.6$, $-6.3$] |

**Architecture ablation.** Table 13 removes or replaces one component at a time, with questions and decoding fixed, for all three answer models. Removing Arbiter changes the grade by $-0.6$ points for SFT-v2 (paired interval $[-1.6, +0.5]$) because no conflict was detected in any of the 67,041 question–variant runs, so the evidence passed to the answer model is identical. The oracle category is indistinguishable from the predicted one (0.0, $[-1.3, +1.3]$). Replacing the mapped policy with six raw images and no perception tools scores higher for SFT-v2 ($+6.8$, $[+4.8, +8.8]$) and the RL prototype ($+8.1$, $[+6.2, +10.1]$) and is indistinguishable for the untuned model ($+1.0$, $[-0.9, +2.8]$).

Table 13: **Architecture ablation on all 1,719 held-out questions.** Rubric grade (%) for the SFT-v2, untuned, and RL answer models; Cam. F1 and Tool F1 against legacy annotations (%); Views and Calls per question. "w/o task classifier" falls back to all six cameras with only the camera-view tool.

| Variant | SFT-v2 | Untuned | RL | Cam. F1 | Tool F1 | Views / Calls |
|---|---|---|---|---|---|---|
| **Full (predicted category)** | 24.8 [22.4, 27.3] | 17.5 [15.6, 19.5] | 28.3 [26.3, 30.4] | 81.5 | 74.3 | 3.6 / 7.7 |
| Oracle category | 24.7 [22.4, 27.0] | 18.5 [16.6, 20.6] | 28.2 [26.2, 30.2] | 83.1 | 75.2 | 3.5 / 7.8 |
| w/o task classifier | 31.7 [28.8, 34.5] | 18.5 [16.7, 20.4] | 36.4 [33.9, 38.8] | 60.0 | 52.7 | 6.0 / 6.0 |
| w/o Arbiter (recency) | 25.4 [22.9, 27.9] | 18.0 [16.1, 20.0] | 27.9 [25.9, 30.0] | 81.5 | 74.3 | 3.6 / 7.7 |
| w/o Arbiter Tier 2 | 25.0 [22.8, 27.3] | 18.1 [16.4, 19.9] | 28.1 [26.1, 30.1] | 81.5 | 74.3 | 3.6 / 7.7 |
| Query-all tools | 23.4 [21.3, 25.5] | 18.6 [17.0, 20.2] | 23.4 [21.3, 25.5] | 81.5 | 37.4 | 3.6 / 33.1 |
| Matched-budget random | 24.5 [22.4, 26.7] | 15.4 [13.8, 16.9] | 29.7 [27.4, 31.9] | 49.9 | 23.7 | 3.6 / 8.4 |

**Training stages.** Table 14 compares the untuned model, SFT-v2, and the RL prototype under identical inference. SFT-v2 improves on the untuned model by $+7.3$ points with the mapped policy ($[+4.7, +9.9]$) and raises similarity to the references from 0.779 to 0.840 while shortening answers from 969 to 703 characters, largely style agreement. The RL prototype adds $+3.4$ ($[+1.6, +5.3]$) with the mapped policy, but its reward is the judge's grade, its answers are 36% shorter (453 characters), the judge penalizes length (Spearman $-0.18$), and similarity fell from 0.840 to 0.806. In a regression controlling for answer length the RL effect is $-1.1$ ($[-4.1, +2.1]$), so the gain is explained by length. The run neither collapsed nor hedged (no empty answer in 724 sampled rollouts), but we cannot say it improved correctness without a human-reviewed check. Training loss is not evidence of downstream accuracy.

**Reward and metric validity, second judge.** On all 66,904 scored answers from the three architecture ablations the Spearman correlation between similarity and the rubric grade is 0.166 (95%

Table 14: **Training stages** on all 1,719 held-out questions. Rubric grade (%) under three acquisition policies; mean answer length in characters and embedding similarity (mapped policy).

| Model | Mapped | Front-only | All-camera | Length | Sim. |
|---|---|---|---|---|---|
| Untuned Qwen3.5-4B | 17.5 [15.6, 19.5] | 12.1 [10.3, 14.0] | 20.4 [18.7, 22.1] | 969 | 0.779 |
| SFT-v2 | 24.8 [22.4, 27.3] | 28.0 [25.8, 30.1] | 25.8 [23.6, 27.9] | 703 | 0.840 |
| **SFT-v2 + RL prototype** | 28.3 [26.3, 30.4] | 28.1 [26.1, 30.1] | 28.8 [26.6, 31.0] | 453 | 0.806 |

interval 0.133–0.197), and similarity ranks the better-graded answer higher in 60.8% of 352,458 same-question pairs. The judge grades 72% of answers incorrect, 8.5% partial, and 19.5% correct against the generated references. We regraded 3,000 random answers with Qwen2.5-VL-72B as a text-only grader: the judges agree on 79.0% (Cohen's $\kappa = 0.59$, Spearman 0.76), the second is more lenient (mean grade 0.274 against 0.244), and per-variant mean grades correlate at Spearman 0.83 over 39 (model, variant) cells. The main contrasts survive the change of judge (tool evidence raises the untuned model's grade by 6.5 points under both and lowers the RL prototype's under both), though each cell holds only 61–97 answers and small contrasts change sign. Both scores are automatic and neither is human-reviewed; a reviewed subset is required before any claim of correctness.

**Counterfactual camera necessity.** For each question with selected set $S$ we answer from images alone (fixed 72B reader, no tool evidence) with $S$, with $S$ minus each camera, and with $S$ plus each unselected camera, and measure the change in rubric grade (Table 15). The selected set is supported over front-only ($+0.122$, but only 17% of questions see a positive effect), all six cameras do no better than the selected set, and removing a selected camera has a small effect whether or not it is annotated (0.069 and 0.052). An added, non-annotated camera has essentially no effect ($+0.002$) while an added, annotated one moves the grade by 0.046, so annotation agreement is at best a weak proxy for necessity. The reader sees images without tool evidence, so this speaks to image necessity, not to the full pipeline.

Table 15: **Counterfactual camera necessity** on all 1,719 questions. Mean change in rubric grade (0–1 scale) with a scene-level bootstrap interval; positive means the camera helps. $n$ counts camera effects; "Annotated" means listed in the legacy `required_cameras`; "Helps" is the share of effects that are strictly positive.

| Effect | $n$ | Mean | 95% CI | Helps (%) |
|---|---|---|---|---|
| Remove a selected, annotated camera | 3933 | $+0.069$ | $[+0.054, +0.084]$ | 13 |
| Remove a selected, non-annotated camera | 1542 | $+0.052$ | $[+0.031, +0.072]$ | 11 |
| Add an unselected, annotated camera | 577 | $+0.046$ | $[+0.016, +0.078]$ | 12 |
| Add an unselected, non-annotated camera | 4241 | $+0.002$ | $[-0.010, +0.014]$ | 6 |
| All six cameras vs. selected set | 1716 | $+0.005$ | $[-0.009, +0.019]$ | 8 |
| Selected set vs. front-only | 1714 | $+0.122$ | $[+0.102, +0.141]$ | 17 |

**Cross-dataset evaluation.** NuPlan is never used in training and uses an eight-camera rig. We take 1,000 template questions over 63 scenes (question text and scene pointers only) and run the same pipeline, with the category map translated to nuPlan cameras. Because nuPlan has no answer key, Qwen2.5-VL-72B answers the same images and evidence and its answer is the reference, so the score is agreement with a stronger model, not correctness, and the evidence arm is favored since the reference saw the evidence. The keyword classifier recognizes only 30% of these questions. Waymo-only training does at least as well as nuScenes-only ($+2.2$ on held-out nuScenes, $[+0.4, +4.2]$; $+3.0$ on nuPlan, $[+0.5, +5.6]$), the mixed adapter matches Waymo-only, and it has twice the training records, so the comparison does not isolate domain diversity. The untuned model agrees more with the 72B reference than the mixed adapter on nuPlan ($+3.3$, $[+0.2, +6.2]$), whereas the adapters beat it by 5.6 to 7.8 points on nuScenes.

**Learned acquisition details.** Table 17 gives the cost sweep of Figure 3. At zero cost the policy picks the physically relevant camera only 80% of the time (100% at higher cost), because a small mean-reward edge for an unrelated camera on the "ahead" region is enough when acquisition is free; the policy chases reward statistics, not scene understanding. Of three training attempts, group-

Table 16: **Training-domain ablation.** Rubric grade (%) for the category-mapped policy on held-out nuScenes (1,719 questions) and on nuPlan (1,000 questions, graded against a 72B-VLM reference). The last column is the paired difference from the mixed adapter on nuPlan with evidence. Single-domain adapters use 1,000 training records and the mixed adapter 2,000.

| Training data | nuScenes (held-out) | nuPlan, evidence | nuPlan, no evidence | nuPlan − mixed |
|---|---|---|---|---|
| Untuned Qwen3.5-4B | 17.5 [15.6, 19.5] | 52.1 [49.5, 54.7] | 35.3 [32.7, 38.1] | +3.3 [+0.2, +6.2] |
| nuScenes only | 23.1 [20.9, 25.3] | 46.0 [43.4, 48.4] | 39.9 [37.1, 42.8] | −2.8 [−5.2, −0.5] |
| Waymo only | 25.3 [23.3, 27.4] | 48.9 [46.4, 51.6] | 43.0 [40.4, 45.7] | +0.2 [−2.1, +2.6] |
| **Mixed (ours)** | 24.8 [22.4, 27.3] | 48.8 [46.4, 51.1] | 38.0 [35.7, 40.5] | – |

relative policy gradient collapsed (about 81% of late rollouts answered "none" even with injected exploration), a warm start on each question's reward-maximizing subset collapsed (4.7% of samples requested a camera, since 78% of those labels are "none"), and the region-mean labels succeeded (labels are "none" for ahead, ahead-left, and behind and the region's own camera for ahead-right, rear-left, and rear-right, with mean gains of $+0.05$, $+0.13$, and $+0.19$). The live agentic policy reaches 58.9% (200-question RL run, $-5.9$ against its warm start, $[-7.7, -4.1]$) and 57.3% (948-question run, $-7.5$, $[-9.4, -5.6]$); in the smaller run only 26 of 800 rollouts called the tool, with rates of 3.5%, 5.0%, 3.0%, and 1.5% over four windows of 50 questions. With a 4B policy and reward only at the end of a one-to-three-step episode, credit for a tool call is weak and rare, so updates reinforce answering immediately. A denser reward, such as crediting tool calls from the measured evidence gain, is the more promising direction.

Table 17: **Cost weight of the learned acquisition policy** on the ground-truth benchmark. Acquire (%) is the share of test questions with any extra camera.

| Cost per view | Accuracy | Views | Acquire (%) |
|---|---|---|---|
| 0 | 69.6 [67.0, 72.1] | 2.20 | 100 |
| 0.02 | 67.5 [64.9, 70.1] | 2.00 | 100 |
| 0.05 | 64.8 [62.3, 67.1] | 1.48 | 48 |
| 0.1 | 62.9 [60.6, 65.4] | 1.33 | 33 |
| 0.2 | 60.0 [57.7, 62.1] | 1.17 | 17 |

**Real cross-camera conflicts.** Table 18 scores the 403 matched pairs of the real conflict set (Appendix C); on agreeing pairs Arbiter and recency are identical by construction.

Table 18: **Real cross-camera pairs** (oracle identity). Correct / unresolved / incorrect (%) against the annotated distance with a 3 m tolerance; Arbiter here is Tier 1 without re-query.

| Split | Recency | Structured Arbiter | $n$ |
|---|---|---|---|
| All matched pairs | 39.2 / 0.0 / 60.8 | 29.3 / 38.2 / 32.5 | 403 |
| Real disagreement | 32.2 / 0.0 / 67.8 | 8.8 / 90.1 / 1.2 | 171 |
| Cameras agree | 44.4 / 0.0 / 55.6 | 44.4 / 0.0 / 55.6 | 232 |

**Arbiter sensitivity.** Table 19 varies one factor at a time around the base setting (10 m corruption, 20% re-query failure, 3 m tolerance). The re-query-first outcome tracks the re-query reliability almost one for one, is insensitive to corruption magnitude, and its correct rate peaks at tolerances of 2 to 3 m (with a tighter tolerance the failed re-queries are flagged unresolved instead of committed wrong); down-weighting and structured checks alone succeed only when the corruption is small relative to the tolerance.

## G Extended Related Work

This appendix extends Section 2 with literature outside the driving domain that informs our acquisition and verification design but is not itself compared against in Table 20.

Table 19: **Sensitivity of resolution outcomes** (%). Re-query-first is Correct / Unresolved / Incorrect; other columns show the Correct rate. Base row in bold. Correct counts the 0.3% of pairs whose corruption stays within tolerance, so it is 0.3 points above Resolved in Table 5.

| Factor | Value | Re-query-first | Down-weight | Structured | Recency |
|---|---|---|---|---|---|
| Corruption (m) | 5 | 79.9 / 0.0 / 20.1 | 100.0 | 100.0 | 0.3 |
| | **10** | **79.9 / 0.0 / 20.1** | **1.8** | **1.8** | **0.3** |
| | 15 | 79.9 / 0.0 / 20.1 | 1.8 | 1.8 | 0.3 |
| | 20 | 79.9 / 0.0 / 20.1 | 1.8 | 1.8 | 0.3 |
| Re-query failure | 0.0 | 100.0 / 0.0 / 0.0 | 1.8 | 1.8 | 0.3 |
| | 0.1 | 89.9 / 0.0 / 10.1 | 1.8 | 1.8 | 0.3 |
| | **0.2** | **79.9 / 0.0 / 20.1** | **1.8** | **1.8** | **0.3** |
| | 0.4 | 59.9 / 0.0 / 40.1 | 1.8 | 1.8 | 0.3 |
| | 0.6 | 40.8 / 0.0 / 59.2 | 1.8 | 1.8 | 0.3 |
| Tolerance (m) | 1 | 75.7 / 24.3 / 0.0 | 1.8 | 0.3 | 0.3 |
| | 2 | 79.9 / 20.1 / 0.0 | 1.8 | 1.8 | 0.3 |
| | **3** | **79.9 / 0.0 / 20.1** | **1.8** | **1.8** | **0.3** |
| | 5 | 75.3 / 0.0 / 24.7 | 1.8 | 0.9 | 0.3 |

Table 20: Acquisition trigger, agent capabilities, and training across representative driving work. Dataset entries and agent frameworks are grouped separately.

| Work | Trigger | Acquisition | | Training | |
|---|---|---|---|---|---|
| | | View acquisition | Conflict handling | SFT | RL |
| *QA benchmarks* | | | | | |
| NuScenes-QA (Qian et al., 2024) | Sup. | ✓ | × | × | × |
| NuPlanQA (Park et al., 2025) | Sup. | ✓ | × | × | × |
| DriveLM (Sima et al., 2024) | Sup. | ✓ | × | ✓ | × |
| WaymoQA (Yu et al., 2025a) | Sup. | ✓ | × | ✓ | × |
| *Agent frameworks* | | | | | |
| Agent-Driver (Mao et al., 2024) | Reas. | × | × | × | × |
| AgentThink (Qian et al., 2025) | Reas. | × | × | ✓ | ✓ |
| DriveAgent-R1 (Zheng et al., 2026a) | Reas. | ✓ | × | ✓ | ✓ |
| InfoCoordiBridge (Liu et al., 2026) | Sup. | ✓ | ✓ | × | × |
| DriveJudge (Sun et al., 2026) | Reas. | × | × | ✓ | ✓ |
| **MosaicDrive (ours)** | **Task** | ✓ | ✓ | ✓ | ✓ |

View acquisition: the system obtains views beyond its initial input (not that it receives multi-view images). Trigger: what determines which evidence is obtained (Sup.: fixed view set from the benchmark/pipeline; Reas.: retrieval conditioned on current evidence and confidence; Task: requirements derived from the question before any observation).

**Question-conditioned view and frame selection.** Outside driving, question-conditioned selection of visual inputs is an established problem: Frame-Voyager learns which frames to query for a given question, and FOCUS selects keyframes under an explicit budget (Yu et al., 2025b; zhu).

**Active perception and tool-use policies.** Outside driving, DeepEyes, Active-o3, and Pixel Reasoner train zoom and crop actions with reinforcement learning (Zheng et al., 2026b; Zhu et al., 2026; Wang et al., 2025a), and ActiView evaluates active perception directly (Wang et al., 2025b); DeepEyesV2 and agentic policy optimization train tool invocation with reinforcement learning (Hong et al., 2026; Dong et al., 2026). As with the driving-specific systems discussed in Section 2, the decision to acquire is conditioned on the agent's current reasoning state rather than on task identity established prior to acquisition.

**Knowledge conflict and verification.** Work on knowledge conflicts distinguishes conflict types and the behavior each warrants (Xu et al., 2024; Cattan et al., 2025), and multimodal models detect such conflicts yet tend to favor parametric knowledge over external evidence (Jia et al., 2025). Search-

augmented models remain unreliable when retrieval returns conflicting or noisy results (Pham et al., 2026), faithfulness degrades under inconsistent or counterfactual context (Ming et al., 2025), and conflicts are markedly harder to detect when the sources are heterogeneous (Wu et al., 2025), as they are across camera views and perception tools.