DRIVEATLAS: A CROSS-COUNTRY BENCHMARK FOR GROUNDED DRIVING-RULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) are becoming the perception-and-reasoning backbone of end-to-end driving systems, but the benchmarks used to assess them are built from one or two regions: they tell whether a model recognizes a scene, not whether it acts on the traffic rule that applies *where the scene was taken*. We introduce DRIVEATLAS, a cross-country, rule-grounded benchmark for *grounded driving-rule reasoning*, with 10,143 test and 26,539 training questions across eight countries and four tasks (perception, prediction, planning, and region). Each question is tied to a clause of the local traffic code, drawing on 5,962 unique rule texts, and is kept only if its correct answer changes under another country's rules; the country is never revealed to the model. We evaluate 16 open-weight VLMs (2B–9B parameters, three model generations) and report accuracy alongside a sampling-noise-corrected cross-country disparity D_c. All 16 models show significant differences in driving performance across countries: D_c ranges from 5.22 to 10.39, and its 95% interval excludes zero for every model. Scale does not remove these differences: a larger model of the same family is more accurate, but its country-dependent variation remains high, and in three of the four families where size buys accuracy, D_c moves by under half a point. An error analysis of two models finds that naming the wrong country is rarely the failure (under 3% of sampled errors), whereas about one item in five fails on a regional rule gap. We therefore propose DRIVEOPD, an on-policy self-distillation algorithm whose teacher is the same model reading a jurisdiction-redacted traffic-rule handbook: on both backbones we train, it raises accuracy on all four tasks and lowers D_c by about three points ($6.91 \rightarrow 3.87$, $7.07 \rightarrow 3.88$). Used as the backbone of a vision–language–action policy, DRIVEOPD lowers open-loop L2 error and collision rate in both Boston and Singapore, and narrows the gap between the two cities by 19–26%.

1 INTRODUCTION

Vision-language models (VLMs) have recently emerged as a promising foundation for autonomous driving end-to-end systems, demonstrating strong capabilities in scene understanding, behavior prediction, high-level decision making, and low-level planning (NVIDIA et al., 2025; Chi et al., 2025; Ma et al., 2024; Sima et al., 2024; Xu et al., 2024; Zhou et al., 2025b; Zeng et al., 2025; Luo et al., 2025; Zhou et al., 2025a). To support the rapid progress of VLMs in this domain, researchers have developed a diverse set of driving-oriented benchmarks (Sima et al., 2024; Caesar et al., 2020; Qian et al., 2024; Xie et al., 2025; Hao et al., 2025; Xu et al., 2025), covering a broad spectrum of tasks, from scene- and object-level perception to driver behavior reasoning and inference, from common driving scenarios to open-world corner cases, and from evaluating model effectiveness to assessing robustness.

However, existing benchmarks are typically built from data collected in one or two regions, implicitly assuming a *universal* set of driving rules and conventions. As a result, they may reward models that learn perception-centric shortcuts rather than rule-aware reasoning. In practice, however, driving is deeply shaped by local context (Li et al., 2024; Dong et al., 2024; Wang et al., 2025c). Traffic regulations, road-sign systems, right-of-way conventions, and pedestrian behaviors differ substantially across countries, and the *same* visual scene can correspond to *different* correct actions depending on the region. We argue that evaluating driving VLMs requires a new axis of analysis: **culturally and regionally grounded reasoning**. A competent driving foundation model must (i) recognize regional cues from visual context alone, such as road signs, license plates, lane markings, and vehicle

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.

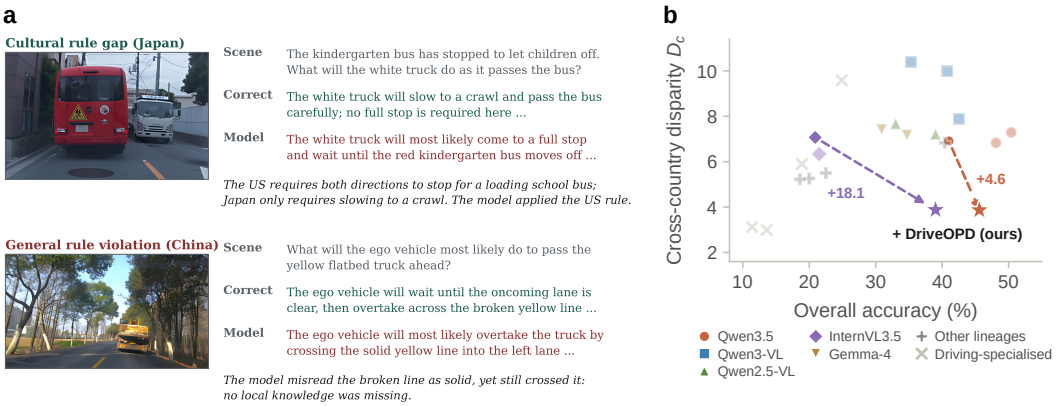

Figure 1: What DRIVEATLAS measures, and what moves it. a, Two illustrative errors: a *cultural rule gap*, the scene read correctly but another country's rule applied (top), and a *general rule violation* on top of a misread lane marking (bottom). **b**, Accuracy versus cross-country disparity D_c for all 20 models. D_c is the spread of a model's accuracy across countries, corrected for sampling noise: the higher it is, the more unevenly the model performs from country to country. DRIVEOPD raises accuracy *and* lowers D_c.

conventions; (ii) retrieve or apply the corresponding local traffic rules; and (iii) integrate both regional context and rule knowledge into perception, prediction, and planning decisions. Crucially, these abilities should be tested *without* explicit region labels, as simply looking up rules given a country name does not demonstrate visual–cultural reasoning. To our knowledge, no existing benchmark jointly addresses traffic-rule grounding, cultural reasoning, multi-country coverage, and audited evaluation at scale. As shown in Figure 1a, a model can read a Japanese scene correctly yet apply a US rule for driving (a *cultural rule gap*), or misread a broken line as solid and cross it anyway (a *general rule violation*). Benchmarks from only one or two regions cannot tell these apart.

To fill this gap, we introduce DRIVEATLAS, a cross-country benchmark for grounded driving-rule reasoning in VLMs, together with DRIVEOPD (**Drive On-Policy Distillation**), a rule-conditioned self-distillation algorithm that internalizes regional traffic knowledge directly into VLM parameters. DRIVEATLAS is built from nine public driving datasets spanning eight countries with distinct traffic systems—China (Mao et al., 2021), Japan (Arai et al., 2025), Singapore (Caesar et al., 2020), the United Kingdom (Marcu et al., 2023), India (Varma et al., 2019), the United States (Yu et al., 2020), Germany (Geiger et al., 2013; Cordts et al., 2016), and Switzerland (Sakaridis et al., 2021)—covering both left- and right-hand traffic and both Asian and Western regulatory traditions. These are all the countries for which we found a public dataset with forward-facing driving video, a research licence, and enough driving sequences to estimate a per-country score (Section 3.1). To find scenes in which these countries' rules diverge, we design a two-stage mining pipeline that combines handbook-derived keywords, Grounding DINO (Liu et al., 2024) filtering, and VLM-based (Bai et al., 2025a) semantic refinement (Section 3.2). For each mined scene, a VLM first extracts a structured scene state (road layout, traffic controls, dynamic agents, and interaction cues) and then, from that state and the governing handbook clause, writes an open-ended question with a reference answer (Section 3.2).

Overall, DRIVEATLAS contains 10,143 open-ended test questions and 26,539 training questions across four tasks—*Perception*, *Prediction*, *Planning*, and *Region*; the *Region* task diagnoses whether region-specific knowledge has been internalized rather than merely retrieved when visually cued. Each question is grounded in a specific clause of a country's traffic-rule handbook and is included only if its correct answer changes when another country's corresponding rule is substituted. The same driving scene can therefore have different correct answers under different regional regulations: for instance, turning right on red is generally prohibited in Japan, the UK, and Singapore, yet permitted under most conditions in China and the US. The resulting rule coverage is broad: the questions draw on 5,962 unique rule texts and 6,619 unique cross-country rule comparisons, organized into 28 regulatory driving scenarios (e.g., turning on red, yellow box junctions, bus priority) and eight decision-level families, from prohibitions painted on the road to signalized intersection control and speed regulation. Questions are open-ended rather than multiple-choice because a multiple-choice format can be solved from text alone: three text-only string heuristics that never see the image answer 49–67% of multiple-choice items correctly (chance is 25%), but only 0.7% of the open-ended ones. To make both the gold answers and their scoring reliable, we verify every item counterfactually, audit scene geometry (removing 20.4% of gold answers in the overtaking scenario), split the data by driving

Table 1: DRIVEATLAS against existing driving benchmarks. *Real*: recorded driving scenes; *Rule*: grounded in a traffic-code clause; *Culture*: requires a region-specific convention; *Region inferred*: jurisdiction withheld; *Rule-swap filter*: admitted only if the answer changes under another country's rules. *# QA* counts released items (ours: 10,143 test plus 26,539 training). (✓): mostly synthetic images; †a method, not a QA benchmark.

Benchmark	Items			Protocol		Countries	# QA
	Real	*Rule*	*Culture*	*Region inferred*	*Rule-swap filter*		
NuScenes-QA (Qian et al., 2024)	✓	✗	✗	✗	✗	SG, US	83K
SURDS (Guo et al., 2024)	✓	✗	✗	✗	✗	SG, US	9.3K
NuScenes-SpatialQA (Tian et al., 2025)	✓	✗	✗	✗	✗	SG, US	4K
DriveLM (Sima et al., 2024)	✓	✗	✗	✗	✗	SG, US, Sim	15K
DriveLMM-o1 (Ishaq et al., 2025)	✓	✗	✗	✗	✗	SG, US	4.6K
DriveBench (Xie et al., 2025)	✓	✗	✗	✗	✗	SG, US	20K
CODA-LM (Chen et al., 2025)	✓	✗	✗	✗	✗	CN, SG, US, DE	4.8K
DriveAction (Hao et al., 2025)	✓	✗	✗	✗	✗	CN	16K
OmniDrive (Wang et al., 2025a)	✓	✓	✗	✗	✗	SG, US	35K
LLaDA-AV† (Li et al., 2024)	✓	✓	✓	✗	✗	SG, US, CN, NL	–
DriveQA (Wei et al., 2025)	(✓)	✓	✗	✗	✗	US	474K
IDKB (Lu et al., 2025)	✗	✓	✓	✗	✗	multi-country	>1M
DRIVEATLAS (Ours)	✓	✓	✓	✓	✓	**CN, US, UK, JP, SG, IND, DE, CH**	**36,682**

sequence so that no test sequence appears in training, and audit the answer judge against two frontier judges of unrelated lineage and against human judges. Crucially, and unlike the prior benchmarks in Table 1, the country is never revealed to the model, so the applicable rule cannot be looked up at test time: the model must infer the region from the scene and already know its rules. DRIVEOPD targets exactly this setting: a single VLM serves as both teacher (conditioned on a *jurisdiction-redacted* handbook) and student (conditioned only on the scene and question), and the student is trained on its own on-policy rollouts to match the teacher's rule-grounded output distribution, so that regional traffic knowledge is absorbed into the parameters and no handbook is needed at inference.

Our work makes the following key contributions: **(1) A Cross-Country, Rule-Grounded Benchmark.** Unlike prior benchmarks built from one or two regions, DRIVEATLAS spans eight countries with both driving sides and both Asian and Western regulatory traditions. To our knowledge, it is the first benchmark to combine real driving scenes, a withheld country label, and a rule-swap filter that keeps only questions whose answer depends on the local rule (Table 1). **(2) Comprehensive Evaluation and Insights.** Evaluating 16 open-weight VLMs (2B–9B parameters, three model generations), we find that all models show significant differences in driving performance across countries (sampling-noise-corrected disparity D_c of 5.22–10.39), and that scale does not remove them: a larger model is more accurate, but its country-dependent variation remains high. Models rarely misidentify the country (under 3% of errors); instead, about one item in five fails on a *cultural rule gap*, where the scene is read correctly but the local rule is not known. **(3) DRIVEOPD: Internalizing Regional Rules for Cross-Country Robustness.** Because models recognize the country yet often lack its rules, and scale does not close this gap, we propose DRIVEOPD, which trains a VLM on its own on-policy rollouts to match itself reading a jurisdiction-redacted handbook, so that regional rules are absorbed into its parameters and no handbook is needed at inference. On both backbones, DRIVEOPD raises accuracy on all four tasks (+4.6 on Qwen3.5-2B, +18.1 on InternVL3.5-2B) and, unlike scale, makes driving decisions more consistent across countries, lowering D_c by about three points. Used as the backbone of a vision–language–action policy, it lowers open-loop L2 error and collision rate in both Boston and Singapore, which drive on opposite sides of the road, and narrows the gap between the two cities by 19–26%.

2 RELATED WORK

VLMs and VLAs for driving. End-to-end systems now fold perception, reasoning and action into a single VLM (Hu et al., 2023; Jiang et al., 2023), whether by emitting control or explanations (Xu et al., 2024; Shao et al., 2024), putting chain-of-thought inside planning (Tian et al., 2024), casting the stack as graph-structured VQA (Wang et al., 2025a; Sima et al., 2024), or coupling a backbone directly to low-level actions (Chi et al., 2025; NVIDIA et al., 2025; Arai et al., 2025). All are trained and evaluated in a narrow set of regions, leaving open whether the backbone's reasoning survives a change of rulebook.

Driving benchmarks for VLMs. Most driving VQA benchmarks are built on one or two regions—nuScenes, and so Singapore and the United States (Qian et al., 2024; Inoue et al., 2024; Sima et al.,

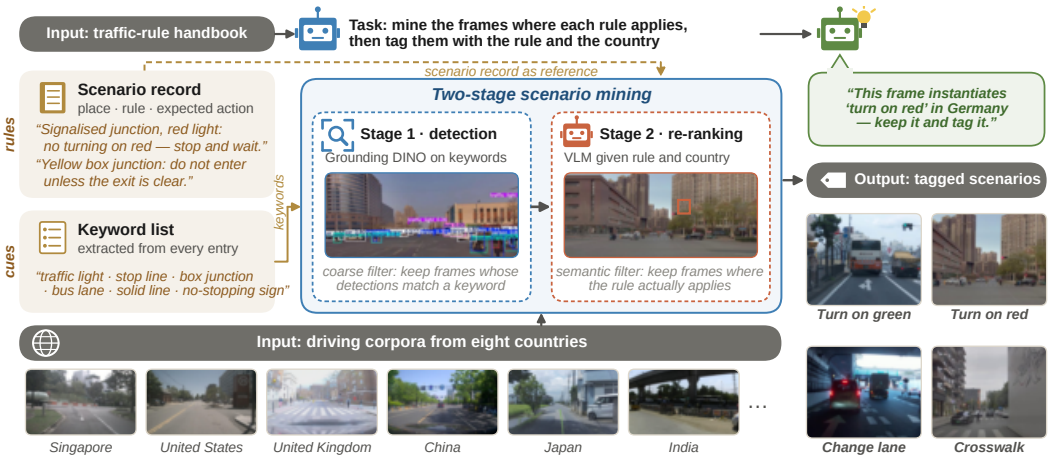

Figure 2: Two-stage scenario mining. Keywords taken from the scenario record drive open-vocabulary detection (Grounding DINO) as a coarse filter; a VLM given privileged country and rule context then re-ranks the survivors, yielding scenario-tagged frames.

2024; Ding et al., 2024; Guo et al., 2024; Tian et al., 2025; Xie et al., 2025; Meng et al., 2025), or Chinese corner cases (Chen et al., 2025; Hao et al., 2025; He & Shi, 2025)—and therefore assume a uniform rulebook. Two lines come closer. LLaDA-AV (Li et al., 2024) adapts a policy across regions, but evaluates open-loop planning on a small set and supplies the region label. On the knowledge axis, DriveQA (Wei et al., 2025) poses driving-theory questions in text and in largely synthetic scenes, and IDKB (Lu et al., 2025) assembles over a million handbook, theory-test and simulator items across many countries and shows that fine-tuning on them raises scores—so multi-country, handbook-grounded evaluation is not new, and neither is the observation that rule text helps. What is new here is where the question is asked from and what has to be inferred to answer it: DRIVEATLAS items are real forward-camera scenes, the jurisdiction is *withheld* so it must be read off the image, and each item is admitted only if its answer changes under a substituted rulebook (Table 1). We know of no benchmark that combines the three; like DriveLM and CODA-LM, DRIVEATLAS is open-ended rather than multiple-choice.

Supplying and distilling rules. A regional rule can be supplied at inference—LLaDA-AV prompts a planner with a region's traffic description, and retrieval agents fetch the regulation that fits the scene (Cai et al., 2024)—but both presuppose a known jurisdiction and pay for the rule text on every call. Context distillation (Snell et al., 2022) instead trains a model to behave as if a prompt block were present when it is not, and on-policy distillation under a reverse KL (Gu et al., 2024; Agarwal et al., 2024) has recently been run with the model's own context-conditioned copy as teacher (Shenfeld et al., 2026). DRIVEOPD is that recipe with a jurisdiction-redacted handbook as the context; we add the redaction, a measurement of the teacher–student gap before training, and an evaluation whose outcome includes cross-country evenness.

3 METHOD

DRIVEATLAS tests whether a VLM can recover a region's traffic norms from the scene and act on them. We describe its regulatory foundations (Section 3.1), its construction (Section 3.2) and its composition (Section 3.3), then DRIVEOPD (Section 3.4).

3.1 RULE-GROUNDED FOUNDATIONS OF DRIVEATLAS

Handbooks and culture-specific situations. Every item is grounded in a national traffic regulation through two artefacts that should be kept apart. The *scenario record* is what items are built from: for each regulatory scenario and country it stores a scene description, the governing rule and the expected ego action, and it is the shared supervision signal for scene mining and verification. Evaluated models never see it. The *handbook* is what a model can be given to read: a numbered transcription of each country's rules, compiled from the national code, the regional descriptions of LLaDA-AV (Li et al.,

2024), and prior studies of cross-country driving behaviour (Dong et al., 2024; Wang et al., 2025c). It runs to 170 sections across the eight countries, from which the items cite 158 distinct named provisions, and covers 95–100% of each country's items (per-country breakdown in Appendix B.1). Scenarios are organised at three levels: 13 *categories* of culture-specific traffic situation, retained only where national codes diverge in the legality of a manoeuvre, in enforcement norms or in regulatory infrastructure (turning on red, pedestrian crossings and bus lanes, box junctions and contraflow lanes); 28 *scenarios* refining them in the release; and eight decision-level *families* used for analysis in Section 4.2.

Source corpora. Corpora were admitted against four criteria fixed before collection: an own national code, forward-facing video with recoverable *sequence* identity, a research licence, and enough sequences for a country mean to be estimable. Nine qualify: CoVLA (Japan), CODA (China), nuScenes (Singapore), BDD100K (United States), LingoQA (United Kingdom), IDD (India), KITTI and Cityscapes (Germany), and ACDC (Switzerland) (Arai et al., 2025; Li et al., 2022; Caesar et al., 2020; Yu et al., 2020; Marcu et al., 2023; Varma et al., 2019; Geiger et al., 2013; Cordts et al., 2016; Sakaridis et al., 2021). China carries three conditions: CODA is assembled from ONCE (Mao et al., 2021), nuScenes and KITTI, so we draw China from its ONCE-derived frames alone, which also keeps the split image-disjoint from Singapore and Germany; CODA ships curated frames rather than video, so China's sequence identity is the parent ONCE recording rather than a window we cut; and because CODA-LM (Chen et al., 2025) uses the same images, China is not image-disjoint from that benchmark. Country and source corpus are *perfectly confounded* by construction, with one exception: Germany contributes both KITTI and Cityscapes. The noise-corrected spread is 1.65 $[0.00, 5.25]$ between those two corpora and 4.95 $[3.16, 11.39]$ between the twelve cities of Cityscapes, against 6.39 $[5.30, 7.74]$ across the eight countries (Appendix E.2); the intervals overlap, so these controls do not show that countries differ by more than recording campaigns or cities do. This is the principal threat to every country-level number here—the eight groups differ in weather, camera, resolution, era and road structure as well as in rulebook—so a country's score carries its corpus with it and country-level differences are descriptive, not causal. Coverage is also biased toward regions with mature autonomous-driving research.

3.2 Construction Pipeline of DriveAtlas

Scenario Mining. We target scenes that look alike but demand different decisions across countries, so that appearance matching alone cannot succeed. A keyword detector over-retrieves and a large VLM over every frame is prohibitive, so we use a two-stage cascade (Figure 2): Grounding DINO (Liu et al., 2024) keeps frames whose detections match a *visual entity keyword* from the scenario record (`traffic light`, `stop line`, `turn arrow` for turning on red), then a frontier VLM (Bai et al., 2025a) re-ranks the survivors, given privileged information the benchmarked models never see—the country label and the matching scenario-record entry—so that mining is an easier task than the benchmark itself. Frames are sampled at 2 Hz (1 Hz for LingoQA), one per short window.

Question Construction. Following prior benchmarks (Sima et al., 2024; Qian et al., 2024; Xie et al., 2025) we adopt the **perception / prediction / planning** decomposition and add a **region** task, in which the model must infer the country from incidental cues and answer a question about that country's rule—a diagnostic of whether regional knowledge has been *internalised* rather than merely cued. Questions are written in two steps: *structured state extraction* records the facts any handbook rule could be conditioned on (road layout, traffic controls, dynamic agents, interaction cues) with a frontier VLM conditioned on the frame, ground-truth boxes, country and handbook; *open-ended question construction* then writes a question and a reference answer from the state and the cited clause, scored against that reference by the rubric judge of Section 4.1. A multiple-choice format was abandoned after three text-only string attacks (an answer-frequency prior, a question–option overlap heuristic, and nearest-question retrieval) solved 49–67% of it against a 25% chance rate; the same attacks score 0.7% against the open-ended rubric. That bounds what a string heuristic recovers, not what language priors do: without the image a VLM still scores 20.6, against 45.7 with it and 19.7 with a mismatched one, on the backbones for which we re-ran these controls (Appendix E.1). Accuracy should be read against that floor, not zero.

Verification and Audit. An item is admitted only if its answer depends on the governing rule (protocol in Appendix B.5). *Counterfactual verification* holds the scene fixed: a VLM verifier sees the image, the structured state and, in turn, the homologous clause of each other country, and answers under that clause. The item is kept only if the reference answer changes under at least

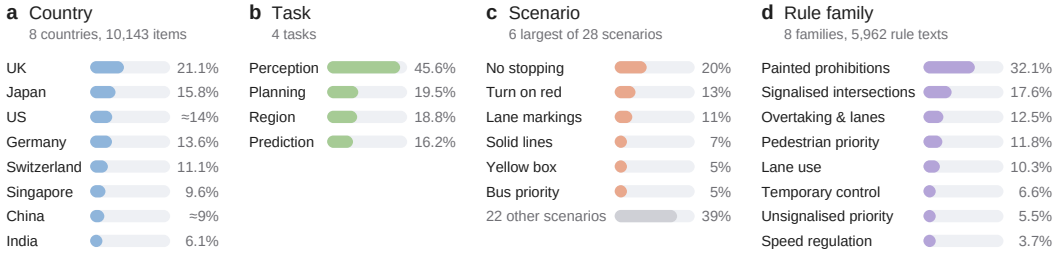

Figure 3: Composition of the DRIVEATLAS test split by country, task, and the six largest scenario categories.

one of them; items invariant under every substitution are dropped as universal driving sense. The criterion constrains the *question*, not every way of failing it: a model can still miss an admitted item by breaking a provision that holds everywhere (Figure 1a, bottom). Because generic review cannot catch errors that require reading scene *geometry*, a targeted audit of the overtaking scenario had a frontier VLM (Bai et al., 2025a) report the lane count, ego lane index, divider and target-vehicle direction *before* judging the gold answer; it found 20.4% of gold answers there wrong (596 of $2{,}928$) against 0.3% elsewhere.

3.3 BENCHMARK ANALYSIS

Figure 3 shows the composition of the test split. Coverage is spread across countries, from the United Kingdom (21.1%) to India (6.1%; panel a). Perception is the largest task ($4{,}627$ items), followed by planning ($1{,}973$), region ($1{,}903$) and prediction ($1{,}640$; panel b). No single scenario dominates: the largest, no stopping, holds 20% of items, and the 22 scenarios outside the top six still hold 39% (panel c). The questions draw on $5{,}962$ unique rule texts and $6{,}619$ unique cross-country rule comparisons. Because 28 scenarios split eight countries too finely for reliable estimates, we group them into eight decision-level families by the decision the driver must make (panel d), which raises the share of country cells with at least 30 questions from 39.2% (per scenario) to 81.2% (per family); Section 4.2 therefore reports results per family. Item counts still overstate the effective sample size, because questions from one recording are correlated (Singapore's 970 questions come from 35 sequences), so every interval on accuracy or D_c in this paper resamples *sequences*, not questions.

Diversity audit. Eight countries can amount to one question asked eight times, so we reduce diversity to four axes, each with a threshold fixed before the values were computed, so that a release can fail its own check: **visual** diversity against images that collapse onto a few scenes, **geo-cultural** against countries present in name but not in mass, **question** against a corpus of a few templates, and **rule** against many questions probing one regulation. Our first internal build failed the question axis, with 85.4% of perception items opening with the same six words. Appendix B.6 gives each axis its statistics, threshold, measured value and verdict.

3.4 DRIVEOPD: LEARNING FROM A HANDBOOK-CONDITIONED TEACHER

If a model reads the scene correctly but does not know which rule governs where it is, the missing ingredient should be expressible in weights rather than supplied at inference. DRIVEOPD does this by *context distillation* (Snell et al., 2022) under a per-token reverse KL (Gu et al., 2024) on the student's own rollouts (Agarwal et al., 2024)—in effect the on-policy self-distillation recipe of Shenfeld et al. (2026) with a frozen teacher; our additions are the context, a whole-country traffic-rule handbook, and the redaction of jurisdiction names from it. One checkpoint plays both roles. The teacher $\pi^T(\cdot \mid x, q) = \pi_{\theta_0}(\cdot \mid x, q, r_c)$ is the *frozen* initial model reading the image x, the question q and the entire redacted handbook r_c of the scene's country; the student $\pi^S(\cdot \mid x, q) = \pi_\theta(\cdot \mid x, q)$ is the same model with a trainable LoRA adapter and no handbook. We sample rollouts $y \sim \pi^S$ and minimise

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{(x,q,c)\sim\mathcal{D}}\, \mathbb{E}_{y\sim\pi^S}\left[\frac{1}{|y|}\sum_t \mathrm{KL}\big(\pi^S(\cdot \mid x, q, y_{<t}) \,\|\, \pi^T(\cdot \mid x, q, y_{<t})\big)\right], \tag{1}$$

restricted at each step to the teacher's top-k tokens ($k = 64$), over the DRIVEATLAS training split $\mathcal{D}$. No gold answer enters the loss, so the only training signal is what the handbook changes in the teacher's next-token distribution—a quantity that can be measured before any training. We train a rank-32 LoRA adapter with four rollouts per prompt.

Table 2: Evaluation of open-weight VLMs on DRIVEATLAS (Direct prompt). Equal-country macro accuracies (%) and the cross-country disparity D_c with its 95% sequence-bootstrap interval; best generalist accuracy in bold. A low D_c at low accuracy (driving-specialised block) is uniform failure, not evenness.

Model	Size	Task accuracy (%) ↑				Overall ↑	D_c [95% CI] ↓
		Perc.	*Pred.*	*Plan.*	*Region*		
Generalist VLMs, grouped by family							
Qwen3.5-2B (Qwen Team, 2026)	2B	46.8	32.5	39.7	44.8	41.0	6.91 [5.59, 8.55]
Qwen3.5-4B (Qwen Team, 2026)	4B	**55.9**	38.3	48.4	50.0	48.1	6.83 [5.73, 8.30]
Qwen3.5-9B (Qwen Team, 2026)	9B	54.8	**40.9**	**50.7**	**55.0**	**50.4**	7.29 [6.21, 8.60]
Qwen3-VL-2B (Bai et al., 2025a)	2B	39.3	26.5	34.1	41.2	35.3	10.39 [9.08, 11.94]
Qwen3-VL-4B (Bai et al., 2025a)	4B	46.2	32.8	38.5	45.9	40.8	9.98 [8.83, 11.32]
Qwen3-VL-8B (Bai et al., 2025a)	8B	52.8	30.2	39.9	47.3	42.5	7.88 [6.72, 9.28]
Qwen2.5-VL-3B (Bai et al., 2025b)	3B	37.1	22.3	33.4	39.4	33.0	7.67 [6.46, 9.04]
Qwen2.5-VL-7B (Bai et al., 2025b)	7B	46.3	27.3	36.2	46.4	39.0	7.21 [5.90, 8.73]
InternVL3.5-2B (Wang et al., 2025b)	2B	8.8	22.2	23.9	28.8	20.9	7.07 [5.85, 8.44]
InternVL3.5-8B (Wang et al., 2025b)	8B	10.8	18.1	27.2	29.9	21.5	6.34 [5.21, 7.54]
Gemma-4-E2B (Gemma Team, 2026)	2B	32.5	23.3	31.7	35.9	30.9	7.42 [6.20, 8.83]
Gemma-4-E4B (Gemma Team, 2026)	4B	38.1	25.7	35.2	39.8	34.7	7.17 [5.79, 8.87]
LLaVA-OV-1.5-8B (An et al., 2025)	8B	44.6	29.0	43.3	44.7	40.4	6.83 [5.68, 8.21]
MiniCPM-V-4.5 (Yu et al., 2025)	8B	13.1	16.1	30.7	30.2	22.5	5.50 [4.42, 6.75]
Molmo-7B-D (Deitke et al., 2025)	7B	13.4	16.1	24.1	26.4	20.0	5.26 [4.29, 6.55]
SmolVLM2-2.2B (Marafioti et al., 2025)	2.2B	5.0	17.2	22.6	29.5	18.6	5.22 [4.07, 6.66]
Mean (16 generalist models)	–	34.1	26.2	35.0	39.7	33.7	7.19
Driving-specialised models (reference only; excluded from the analysis)							
DriveMM (Huang et al., 2024)	8B	9.3	27.8	31.1	31.5	24.9	9.58 [8.34, 10.88]
Dolphins (Ma et al., 2024)	7B	7.6	19.4	25.1	23.4	18.9	5.90 [4.82, 7.13]
DriveLM-2B (Sima et al., 2024)	2B	9.8	8.0	3.7	24.1	11.4	3.12 [2.31, 4.30]
DriveLM-4B (Sima et al., 2024)	4B	10.6	7.7	7.8	28.4	13.6	2.99 [2.35, 4.08]

4 EXPERIMENTS

4.1 SETUP

Models and prompts. We evaluate 16 open-weight VLMs (2B–9B) under one inference stack with no extended-thinking mode; Table 2 names them and lists four older driving-specialised models for reference. The *Direct* prompt asks for an answer of at most two sentences, *Reasoning* requests a fixed four-stage trace (region, rule, visible evidence, answer), and *Rule-Given* additionally reveals the country and the governing clause—an oracle used only in Section 3.4.

Metrics. An image-aware rubric judge (Gemma-4-31B-IT, one deployment throughout) sees the image, the question and the reference answer and returns a binary *correct* verdict. Its rate is *accuracy*, always equal-weighted across countries (Overall is the mean of all 32 country–task cells). Reference answers are verified by human reviewers, and the judge is audited against a human expert and two judges of unrelated lineage (Appendix D.2). Evenness is a sampling-noise-corrected *cross-country disparity*,

$$D_c = 100\sqrt{\max(V_{\text{obs}} - V_{\text{noise}},\ 0)}, \tag{2}$$

where V_{obs} is the variance of the eight country means and V_{noise} the mean of their sequence-bootstrap variances (2,000 draws), so $D_c = 0$ means the spread is explained by sampling. Since its intervals in Table 2 have half-widths of 1.1–1.5 points, we call a difference in D_c *resolvable* only above two points.

4.2 BENCHMARK RESULTS: WHAT CURRENT VLMS DO

Every model is uneven across countries, and part of the unevenness is shared. Overall accuracy spans 18.6 (SmolVLM2-2.2B) to 50.4 (Qwen3.5-9B) with a mean of 33.7, so no model is close to solving the benchmark (Table 2). D_c spans 5.22 to 10.39, and its 95% interval excludes zero for all 16 models; Singapore is the lowest-scoring country for 10 of them. Figure 4a shows what the statistic compresses: Qwen3.5-9B is above Qwen3.5-2B in every country, but by anything from 2.6 points

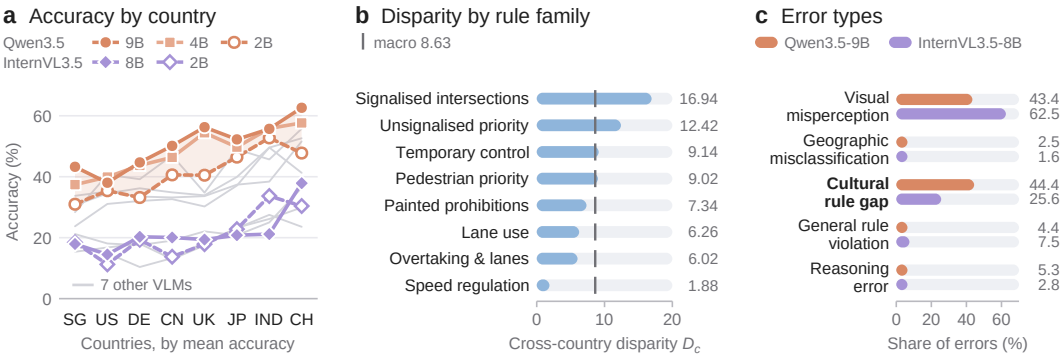

Figure 4: What current VLMs do on DRIVEATLAS. a, Per-country accuracy under the Direct prompt for 12 VLMs, countries ordered by their mean over all 12; Qwen3.5 (2B, 4B, 9B) and InternVL3.5 (2B, 8B) are highlighted and the other seven are grey. The shaded band is the Qwen3.5 gain from 2B to 9B (2.6–15.8 points). **b**, Cross-country disparity D_c per rule family (Direct prompt, macro over the 16 generalists); the dark tick on each bar marks the macro D_c of 8.63. **c**, Error types under the Reasoning prompt, as the mean share of 40 sampled errors per country.

Table 3: DRIVEOPD against its own frozen backbone. Same test split, Direct prompt and judge; the student (highlighted) is never told where it is. Task and Overall columns are equal-country macro accuracies.

Backbone	Variant	Task accuracy (%) ↑				Overall ↑	D_c ↓
		Perc.	*Pred.*	*Plan.*	*Region*		
Qwen3.5-2B	Base	46.8	32.5	39.7	44.8	41.0	6.91
	+ DRIVEOPD	49.2	39.2	44.5	49.6	**45.6**	3.87
InternVL3.5-2B	Base	8.8	22.2	23.9	28.8	20.9	7.07
	+ DRIVEOPD	40.9	33.7	37.8	43.4	**39.0**	3.88

(United States) to 15.8 (United Kingdom), so a high aggregate does not imply region-robustness. The profiles also move together: ordered by their mean, nearly every model rises from Singapore to Switzerland, and the profiles of Qwen3.5-9B, Qwen3.5-2B and InternVL3.5-2B correlate at $r = 0.73$–0.79. Part of D_c is therefore a difficulty profile of the eight country–corpus groups that all models share, and a text-only control keeps a comparable disparity (4.76 at 20.6% accuracy, against 7.10 at 45.7% with the image). We interpret only *differences* in D_c between models scored on the same items; Appendix C.1 gives the full 16×8 per-country matrix.

A larger model of the same family is usually more accurate but rarely more even. Sizes of one family differ in data and vision encoder as well as in parameters, so this is a released-size comparison, not a controlled scaling study. Overall accuracy rises with size in four of the five multi-size families (by 3.8–9.4 points) and is flat in InternVL3.5, whose 8B and 2B profiles nearly coincide (Figure 4a). D_c does not follow accuracy: it changes by $+0.38$, -0.46 and -0.25 in Qwen3.5, Qwen2.5-VL and Gemma-4, and by -0.73 in InternVL3.5, all far inside the two-point threshold. The exception, Qwen3-VL (-2.51), starts from the highest disparity of any generalist, and its 2B model sits about three points above both its predecessor and its successor ($7.67 \rightarrow 10.39 \rightarrow 6.91$); it is also the model we used to mine scenes, so we conclude only that evenness is neither fixed by a lineage nor bought with parameters.

The disparity concentrates in a few rule families, led by signalised intersection control. Among eight decision-level families fixed before inspecting model outputs (Figure 4b; Table 6), signalised intersection control is the widest ($D_c = 16.94$, about twice the macro of 8.63), followed by right of way without a signal (12.42). Speed regulation, a single scenario (Appendix C.2), is both the narrowest (1.88) and the best answered (62.7), plausibly because a speed limit is posted as a number while signal phasing and priority are not. Stating the region and rule first does not close the gap: under the Reasoning prompt, family-averaged accuracy rises only from 40.8 to 42.1 (families weighted equally, unlike the Overall of Table 2; Appendix D), and macro D_c does not fall ($8.63 \rightarrow 9.16$).

Naming the wrong country is rarely the failure; about one item in five fails on a regional rule gap. We sampled 40 errors per country for Qwen3.5-9B and InternVL3.5-8B under the Reasoning prompt, whose trace names the assumed country, and labelled each by its earliest failing stage (Figure 4c). Geographic misclassification accounts for only 2.5% and 1.6% of errors. Qwen3.5-9B

Table 4: DRIVEOPD as a VLA backbone. Open-loop planning on nuScenes in Boston (right-hand traffic) and Singapore (left-hand traffic): L2 error (m) at 1–3 s and its average, average collision rate CR (%), and the Singapore-minus-Boston Gap in average L2; lower is better, best in bold.

Backbone	Boston					Singapore					Gap
	1s	*2s*	*3s*	*Avg.*	*CR*	*1s*	*2s*	*3s*	*Avg.*	*CR*	
Qwen3.5-2B	0.135	0.262	0.468	0.288	2.91	0.195	0.379	0.692	0.422	0.67	0.134
+ DRIVEOPD	**0.125**	**0.253**	**0.463**	**0.280**	2.80	**0.151**	**0.335**	**0.649**	**0.379**	**0.60**	**0.099**
InternVL3.5-2B	0.140	0.273	0.492	0.302	2.93	0.198	0.393	0.720	0.437	0.72	0.135
+ DRIVEOPD	0.128	0.262	0.483	0.291	**2.74**	0.159	0.356	0.686	0.401	0.67	0.110

splits almost evenly between visual misperception (43.4%) and a *cultural rule gap* (44.4%), a scene read correctly but governed by a provision the model does not hold; InternVL3.5-8B is dominated by perception (62.5% vs. 25.6%). Per item, the rule gap is about one in five (19.5% and 21.3%), a lower bound given earliest-stage labelling, and the pattern holds for two further models (Table 11). This is the part a handbook could supply, and the target of Section 3.4.

4.3 DRIVEOPD

DRIVEOPD improves driving VLMs across countries. Table 3 compares each student with its own frozen backbone on the same test split, prompt and judge; neither is told the country at test time. We train two 2B backbones of different lineages, a strong one (Qwen3.5-2B, 41.0) and a weak one (InternVL3.5-2B, 20.9). Both improve on all four tasks, by 4.6 points overall for Qwen3.5-2B and by 18.1 for InternVL3.5-2B, and both become more even: D_c falls from 6.91 to 3.87 and from 7.07 to 3.88, reductions of 3.0 and 3.2 points that exceed the two-point threshold of Section 4.1. Because accuracy rises at the same time, the lower disparity is not the uniform failure of the driving-specialised block in Table 2, and both students are more even than every generalist in that table. The gain is not confined to reading the scene: Region, the task that tests whether regional knowledge is held rather than read off the frame, rises by 4.8 and 14.6 points. The weaker backbone gains most; after training, InternVL3.5-2B scores 17.5 points above its own 8B sibling (21.5).

A DRIVEOPD backbone yields a VLA that plans better in both cities and narrows the gap between them. We train the same VLA recipe on each frozen backbone and on its DRIVEOPD student, and evaluate open-loop planning on nuScenes separately for Boston (right-hand traffic) and Singapore (left-hand traffic; Table 4). On both backbones, DRIVEOPD lowers the L2 error at every horizon and the average collision rate in both cities (collision rate $2.91 \rightarrow 2.80$ and $0.67 \rightarrow 0.60$ on Qwen3.5-2B; $2.93 \rightarrow 2.74$ and $0.72 \rightarrow 0.67$ on InternVL3.5-2B). The gain is larger in Singapore, where the frozen backbones plan worst: average L2 falls by 10.2% and 8.2%, against 2.8% and 3.6% in Boston. The Singapore–Boston gap in average L2 therefore narrows from 0.134 to 0.099 m (-26%) and from 0.135 to 0.110 m (-19%). This is open-loop evaluation on two cities of one corpus, from one run per backbone; whether the gain holds in closed loop and across the other six countries is left to future work.

5 CONCLUSION

We introduced DRIVEATLAS, a cross-country benchmark for rule-grounded driving reasoning with 10,143 test and 26,539 training questions across eight countries and four tasks. All 16 open-weight VLMs we evaluate are uneven across countries, larger models are more accurate but rarely more even, and about one item in five fails on a regional rule gap rather than on naming the country. DRIVEOPD, whose teacher is the same model reading a jurisdiction-redacted handbook, raises accuracy on all four tasks, lowers D_c by about three points, and as a VLA backbone narrows the planning gap between Boston and Singapore. Scale did not buy cross-country evenness, but a handbook-conditioned teacher did. Limitations remain: country and corpus are confounded, items are generated with one model family, rules are modelled nationally without Latin America or Africa, and DRIVEOPD is trained once per backbone and evaluated open-loop in two cities.

AI USE STATEMENT

Large models are both the object of study and components of this work. Frozen open-weight VLMs are the evaluated systems; a detector and a frontier VLM mine scenarios, write questions and reference answers, and verify them (Section 3); a model of a different lineage serves as the rubric judge (Section 4.1); and DRIVEOPD trains a VLM against a handbook-conditioned copy of itself (Section 3.4). Reference answers are verified by human reviewers, and the judge is evaluated against a human expert and two other judges (Appendix D.2). Beyond these roles, we used generative AI tools only for code editing and for refining the writing; they were not used to form hypotheses, design experiments, or produce reported numbers. The authors take responsibility for all claims, code, and text.

ETHICS STATEMENT

This work is diagnostic: it measures how unevenly driving VLMs apply traffic rules across countries, so that the gap can be found and reduced before such models are deployed. DRIVEATLAS is built from publicly released driving datasets under their original licences and adds no identifying information beyond what the source data already contain; the traffic-rule handbooks are compiled from publicly available regulations. A score on DRIVEATLAS is not a certification: it is not evidence that a system is safe or legally compliant in any jurisdiction, and DRIVEOPD is evaluated only in open-loop planning. Traffic codes change, so the handbooks reflect the rules at construction time and should be checked against current official sources before any deployment. The eight countries do not include Latin America or Africa, and evenness across them does not imply evenness elsewhere.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, et al. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1933–1943. IEEE, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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*, pp. 11621–11631, 2020.

Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z. Zhao, Zhiwen Wu, Xu Han, and Jiaqi Ma. Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via LLM. *arXiv preprint arXiv:2410.04759*, 2024.

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

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Yongqi Dong, Chang Liu, Yiyun Wang, and Zhe Fu. Towards understanding worldwide cross-cultural differences in implicit driving cues: Review, comparative analysis, and research roadmap. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1569–1575. IEEE, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. *arXiv preprint arXiv:2411.13112*, 2024.

Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, and Xianpeng Lang. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025.

Yuankai He and Weisong Shi. Carscenes: Semantic vlm dataset for safe autonomous driving. *arXiv preprint arXiv:2511.10701*, 2025.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023.

Zhijian Huang, Chengjian Feng, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. DriveMM: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024.

Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 930–938, 2024.

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20501–20508. IEEE, 2025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14948–14957, 2024.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, and Hang Xu. CODA: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. Can LVLMs obtain a driver's license? a benchmark towards reliable AGI for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, et al. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*, 2025.

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

Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, et al. One million scenes for autonomous driving: Once dataset. 2021.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. Lingoqa: Visual question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023.

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 preprint arXiv:2511.14592*, 2025.

NVIDIA, Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, Liang Feng, Greg Heinrich, Jack Huang, Peter Karkus, Boyi Li, Pinyi Li, Tsung-Yi Lin, Dongran Liu, Ming-Yu Liu, Langechuan Liu, Zhijian Liu, Jason Lu, Yunxiang Mao, Pavlo Molchanov, Lindsey Pavao, Zhenghao Peng, Mike Ranzinger, Ed Schmerling, Shida Shen, Yunfei Shi, Sarah Tariq, Ran Tian, Tilman Wekel, Xinshuo Weng, Tianjun Xiao, Eric Yang, Xiaodong Yang, Yurong You, Xiaohui Zeng, Wenyuan Zhang, Boris Ivanovic, and Marco Pavone. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail, 2025. URL https://arxiv.org/abs/2511.00088.

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.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, pp. 10765–10775, 2021.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15120–15130, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

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.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4567–4576, 2025.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1743–1751. IEEE, 2019.

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

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.

Yuting Wang, Zhaocheng He, Wangyong Xing, and Chengchuang Lin. Impact of regional driving behavior differences on traffic flow. *Scientific Reports*, 15(1):9027, 2025c.

Maolin Wei, Wanzhou Liu, and Eshed Ohn-Bar. DriveQA: Passing the driving knowledge test. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 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*, pp. 6585–6597, 2025.

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

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

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

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

A Overview

This appendix gives the details needed to reproduce DriveAtlas and DriveOPD, the controls behind our validity claims, and the full results summarized in the main text. It is organized as follows.

1. **Appendix B: Benchmark Construction Details.** The per-country traffic-rule handbooks (170 sections across eight countries) and how many of each country's items they cover; the 13 culture-specific categories, 28 scenarios and eight regulatory families; the nine source corpora and the sequence split; the question-construction prompts; the counterfactual verification and geometry-audit protocol; the diversity audit; and the identities and decoding settings of all evaluated models.
2. **Appendix C: Additional Results.** The full 16×8 per-country accuracy matrix, D_c per regulatory family, the error analysis broken down by country, and a per-task diagnosis of the prompting settings.
3. **Appendix D: Metric Consistency.** How a free-text answer becomes a reported number (the judge template, equal-country aggregation, and the sequence-level bootstrap); the audit of the answer judge against a human expert and two judges of unrelated lineage; and accuracy and the cross-country disparity D_c recomputed on the judge's graded rubric scale.
4. **Appendix E: Validity Controls.** Text-only and mismatched-image controls, which bound how much of the accuracy comes from language priors, and a within-country comparison of corpora and cities, which bounds the country–corpus confound.
5. **Appendix F: Demonstration.** Qualitative comparisons across diverse VLMs on DriveAtlas.

B Benchmark Construction Details

B.1 Traffic-Rule Handbooks and Coverage per Country

Table 5 reports what each country's handbook contains and how much of that country's items it reaches. Across the eight countries the handbooks run to 170 numbered sections and 10,125 words, and the items cite 158 distinct named provisions from them: a handbook is a few pages of rules, not a national code in full. Each is transcribed only as deeply as its own items require, which is why the counts differ between countries (Japan needs 25 sections and 52 named provisions, Switzerland 16 and 8); they describe how far the benchmark reaches into each code, not the code itself. Every section is cited by at least one item, and the handbooks cover 95–100% of each country's items (Section 3.1).

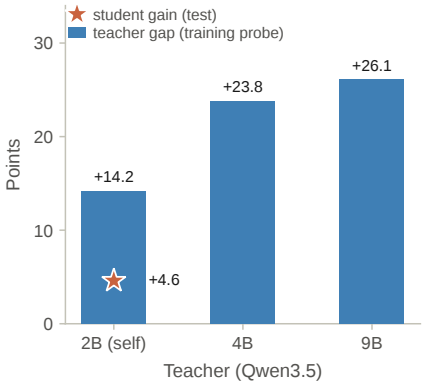

Figure 5: What the handbook is worth in context. Accuracy gain of a Qwen3.5 model on 6,154 held-out training items when the handbook is placed in its context (bars), and the test gain of the DriveOPD student distilled from the 2B teacher (star).

Coverage bounds what the teacher of Section 3.4 can transmit: a provision that an item depends on but the handbook omits is one the teacher cannot supply and the student cannot receive. Figure 5 measures what the handbook is worth before any training, by scoring the same model with and without it in context on 6,154 held-out training items. It is worth +14.2 points to Qwen3.5-2B, the teacher of Section 3.4, and +23.8 and +26.1 to the 4B and 9B models, so a more capable reader extracts more from the same text; we nevertheless keep teacher and student the same model, so that they differ only by the handbook (Section 3.4). The student distilled from the 2B teacher gains +4.6 on the test split, about a third of its teacher's gap.

B.2 Culture-Specific Scenarios and Regulatory Families

The eight decision-level families of Section 3.3 were fixed before any model output was inspected. They are numbered here as F1–F8, the numbering used in Table 6:

Table 5: How much regulation the benchmark covers. Each country contributes a traffic-rule handbook of numbered sections transcribed from its national code, deep enough to cover the provisions its own items cite. *Sections* is the length of that handbook and *words* its size; neither is a property of the country's code, only of how much of it this benchmark reaches. *Provisions* counts the distinct named clauses the items cite (turning on red, bus-lane entitlement, crossing obligations, and so on) and *scenarios* the distinct regulatory situations they are drawn from. Every one of the 36,682 items is admitted only if its correct answer changes under another country's rules, so each rests on a provision that is genuinely local rather than universal.

Country	Sections	Words	Provisions	Scenarios	Items
China	24	1,327	19	27	6,803
United States	23	1,403	33	28	6,957
United Kingdom	22	1,367	21	21	8,408
Japan	25	1,459	52	28	6,165
Singapore	18	985	16	22	1,668
India	18	1,339	15	22	1,700
Germany	24	1,360	12	24	2,928
Switzerland	16	885	8	21	2,053
All eight	**170**	**10,125**	**158**	**28**	**36,682**

Table 6: Country-dependent performance differences concentrate in intersection rules. Noise-corrected cross-country disparity D_c for every evaluated checkpoint under Direct prompting. Columns are ordered by the generalist macro disparity: signalised intersection control (F1, 16.94) and unsignalised priority (F5, 12.42) show the largest disparities, whereas speed regulation (F8, 1.88) is substantially more even across countries. Values of at least 12 are bold. A country enters a family only when it contains at least 30 items. A value of 0.00 means that the observed spread is entirely explained by sampling noise, not that the cell is missing. F1: signalised intersection control; F5: right of way without a signal; F7: hazards and temporary control; F2: pedestrian and vulnerable-user priority; F4: prohibitions painted on the road; F3: lane use; F6: overtaking and lane discipline; F8: speed regulation.

		Highest disparity		Intermediate disparity					Lowest
Model	Size	Signalised F1	Unsignalised F5	Temporary F7	Pedestrian F2	Painted F4	Lane use F3	Overtaking F6	Speed F8
Macro (16 generalists)	–	**16.94**	**12.42**	9.14	9.02	7.34	6.26	6.02	1.88
Qwen3.5-2B	2B	**12.49**	6.08	8.87	10.23	8.51	6.67	8.46	4.63
Qwen3.5-4B	4B	**15.23**	**12.21**	**13.15**	7.43	6.24	7.34	6.02	0.88
Qwen3.5-9B	9B	**18.86**	**15.01**	10.26	**12.83**	6.94	5.71	0.00	3.33
Qwen3-VL-2B	2B	**17.10**	11.73	10.61	11.70	7.42	10.87	**13.60**	0.00
Qwen3-VL-4B	4B	**21.10**	6.47	10.15	9.73	6.93	11.53	6.46	4.19
Qwen3-VL-8B	8B	**16.75**	11.23	**14.38**	8.07	9.49	7.71	9.00	0.00
Qwen2.5-VL-3B	3B	**16.72**	8.93	9.53	11.96	5.56	**13.14**	6.24	4.03
Qwen2.5-VL-7B	7B	**15.55**	10.90	9.21	9.25	6.54	10.61	8.29	3.32
InternVL3.5-2B	2B	11.60	3.53	0.00	**12.00**	0.75	0.00	**12.44**	1.36
InternVL3.5-8B	8B	**16.60**	**12.56**	2.72	5.93	6.38	1.02	6.80	0.86
Gemma-4-E2B-it	2B	**18.30**	3.41	7.08	10.72	6.40	9.72	2.68	5.63
Gemma-4-E4B-it	4B	11.46	5.28	8.93	9.88	3.93	10.53	8.34	0.00
LLaVA-OneVision-1.5-8B	8B	**15.42**	8.85	5.52	7.90	6.37	8.71	5.51	5.18
MiniCPM-V 4.5	8B	10.93	9.02	2.92	6.54	3.86	0.00	0.00	0.00
Molmo-7B-D	7B	11.78	2.10	0.00	7.85	7.05	4.70	3.43	0.00
SmolVLM2-2.2B	2.2B	8.36	4.54	0.00	3.59	2.02	0.00	9.98	0.00
DriveMM	8B	**16.94**	9.03	0.00	8.76	5.94	4.92	8.31	**13.52**
Dolphins	7B	**13.53**	10.17	0.00	3.76	0.66	3.37	0.00	8.23
DriveLM-2B	2B	5.65	0.75	0.00	0.00	5.39	0.00	5.91	0.00
DriveLM-4B	4B	4.61	2.36	0.00	3.20	5.54	0.00	4.82	4.43

- **F1 Signalised intersection control** (5 scenarios) — what a signal phase permits: turn-on-red exceptions, protected versus permissive arrows, the amber dilemma zone, reservoir boxes, reversible-lane signals.
- **F2 Pedestrian and vulnerable-user priority** (5) — when a driver must yield to people outside a vehicle: uncontrolled crossings, school zones and buses, cycle crossings, scramble phases.
- **F3 Who may use this lane** (5) — lanes reserved by vehicle class, occupancy, direction, or payment.
- **F4 Prohibitions painted on the road** (3) — what a marking forbids: stopping, parking, crossing a centreline, entering a box.
- **F5 Right of way without a signal** (5) — priority resolved by rule rather than by a light: roundabouts, multi-way stops, uncontrolled junctions, bus re-entry, U-turns.

- **F6 Overtaking and lane discipline** (1) — which side to pass on and which lane to hold; this is the scenario whose gold answers the geometry audit of Appendix B.5 corrected.
- **F7 Hazards and temporary control** (3) — control that overrides the permanent layout: work zones, emergency vehicles, level crossings.
- **F8 Speed regulation** (1) — posted and enforced limits, including camera zones.

F6 and F8 hold one scenario each, so their D_c is a within-scenario statistic and is not comparable to the others' as a measure of family breadth. They are reported because omitting the two extremes of the range would flatter the grouping.

B.3 Source Corpora and the Sequence Split

Nine corpora contribute: CoVLA (Japan), CODA (China), nuScenes (Singapore), BDD100K (United States), LingoQA (United Kingdom), IDD (India), KITTI and Cityscapes (Germany), and ACDC (Switzerland). Admission was decided against four criteria fixed before collection — an own national code, forward-facing video with recoverable sequence identity, a research licence, and enough sequences for a country mean to be estimable — and the criteria, not the resulting country list, are what we would defend: the list is what those criteria admitted, and coverage is consequently biased toward regions with mature autonomous-driving research.

The release is split by *driving sequence*, not by frame, so no sequence and no image appears on both sides: 5,509 distinct sequences on the test side and 14,328 on the training side. Every interval on accuracy or D_c in this paper resamples sequences rather than questions, which typically doubles an interval's width relative to an item-level bootstrap.

B.4 Question Construction and Prompt Templates

The templates below are reproduced from `data_generation.prompts.PROMPTS`, the object the inference runner renders, so what is printed is what the models were asked. `evaluation/export_prompts.py --check` fails if the two drift apart. That guard exists because an earlier version of this appendix printed multiple-choice-era templates ending in `Output only the answer letter (A/B/C/D)`, long after every reported number came from open-ended scoring. `{q}` is the question; the rule block is `{rule_block}`, holding either the cited clause or a whole country handbook, with jurisdiction names replaced by "this country" in every setting, including during DriveOPD training.

Direct, the setting all headline numbers use:

```
Look at the driving scene and answer the question.
Question: {q}

Answer in at most two sentences. State the answer itself first, then at most
one clause of justification. Do not list alternatives.
Write only the answer, with no preamble.
```

Reasoning adds a scaffold and no new information:

```
Step 1 - Geographic inference: Based on road infrastructure, traffic signs,
lane markings, and environmental cues in the image, identify the most likely
country or region where this scene was captured.
Step 2 - Cultural traffic rule: State the relevant traffic rule(s) for that
country that apply to this question.
Step 3 - Visual reasoning: Describe what you observe in the image that is
relevant to answering the question.
Step 4 - Answer.
Question: {q}

Output your step-by-step reasoning, then a final line that begins with
'Final answer:'. Answer in at most two sentences. [...]
```

B.5 Counterfactual Verification and Geometry Audit

Section 3.2 describes the counterfactual filter that admits an item only if its answer changes under another country's homologous clause. Table 7 reports the audit that filter cannot perform: it constrains

Table 7: From verified candidates to the released split. The geometry audit had a frontier VLM state the lane count, ego lane, divider and target direction before judging each gold answer. It found 20.4% of gold answers wrong in the overtaking scenario against 0.3% in all others, so the problem is specific to one scenario rather than systemic. An item is removed only when the audit finds its gold answer wrong.

Stage	Removed	Items
Candidates after counterfactual verification		11,039
Remove gold answers found wrong by the geometry audit	−596	10,443
Cap each scenario at 25% of a country's items	−224	10,219
Remove duplicates (question, country, scenario, image set)	−76	10,143
Test split (5,509 driving sequences)		**10,143**
Training split (disjoint sequences, images and questions)		**26,539**
Geometry audit of the overtaking scenario ($n = 2{,}928$)		
Gold confirmed		2,140 (73.1%)
Gold wrong, removed		596 (20.4%)
Undetermined, kept		192 (6.6%)
All other scenarios, gold wrong		0.3%

Table 8: Diversity audit of the DRIVEATLAS test split. Rows with a threshold are the checks: each threshold was fixed before its value was computed, so the release could fail, as our first internal build did on the question axis (Section 3.3). Rows without a threshold are reported for context.

Axis / statistic	Threshold	Value	Pass
Visual			
Distinct driving sequences	—	5,509	
Largest share of items from one sequence	—	1.0%	
Items sharing a question and an image set	$\leq 5.0\%$	0.0%	✓
Geo-cultural			
Countries (both driving sides)	—	8	
Smallest country share of items	—	6.1%	
Country × family cells with ≥ 30 items	$\geq 55.0\%$	52 of 64 (81.2%)	✓
Question			
Distinct 6-word question openings	—	3,316	
Largest share of any 3-word opening	$\leq 25.0\%$	22.7%	✓
Test questions appearing verbatim in training	$\leq 2.0\%$	0.0%	✓
Rephrased questions still asking what the gold answers	$\geq 99.0\%$	100.0%	✓
Rule and answer			
Unique rule texts / scenarios / families	—	5,962 / 28 / 8	
Largest share of items citing one clause	—	1.7%	
Largest share of any gold short answer	$\leq 5.0\%$	2.0%	✓

what a question asks, but it cannot see whether the gold answer is right, and neither can a generic quality review when being right requires reading scene geometry. For the overtaking scenario a frontier VLM had to state the lane count, the ego lane index, the divider and the target vehicle's direction before being shown the gold answer to judge. It found 20.4% of gold answers there wrong against 0.3% elsewhere. We report the whole funnel rather than only the final count, because a $68\times$ concentration in one scenario is the kind of defect a single aggregate quality figure hides.

B.6 DIVERSITY AUDIT

Table 8 gives each axis its statistic, a threshold declared *before* the value was computed, and a verdict. An earlier version of the audit passed every check while restricted to a single country, which means it was measuring nothing; each threshold is now paired with the population it is computed over, and the audit runs as a gate on the release build rather than as a report written afterwards.

B.7 MODEL IDENTITIES AND DECODING SETTINGS

Table 9 records, for every evaluated checkpoint, the snapshot it was loaded from, the number of items judged, and the judge deployment. All 20 checkpoints share one inference stack, one judge deployment and one input split hash; no reported cell mixes them. Extended-thinking modes are disabled throughout.

Three loads were excluded by name rather than by a score threshold, because they produce fluent but degenerate text and a degenerate model is indistinguishable from a weak one in a score column:

Table 9: Evaluated checkpoints. All 20 models are scored on the DRIVEATLAS test split by the same judge, which returned a parseable verdict for every response it received. *Scored* counts test items with a verdict; the three models below 10,143 lack a response on at most 24 items (0.24%), which are excluded from their scores. Gemma-4 sizes are effective parameter counts.

Model	Size	Scored	Model	Size	Scored
Qwen3.5			*Gemma-4*		
Qwen3.5-2B	2B	10,143	Gemma-4-E2B-it	2B	10,143
Qwen3.5-4B	4B	10,142	Gemma-4-E4B-it	4B	10,143
Qwen3.5-9B	9B	10,143	*Other lineages*		
Qwen3-VL			LLaVA-OneVision-1.5-8B	8B	10,143
Qwen3-VL-2B	2B	10,143	MiniCPM-V 4.5	8B	10,143
Qwen3-VL-4B	4B	10,142	Molmo-7B-D	7B	10,143
Qwen3-VL-8B	8B	10,143	SmolVLM2-2.2B	2.2B	10,143
Qwen2.5-VL			*Driving-specialised*		
Qwen2.5-VL-3B	3B	10,143	DriveMM	8B	10,143
Qwen2.5-VL-7B	7B	10,143	Dolphins	7B	10,143
InternVL3.5			Mini-InternVL2-2B-DA-DriveLM	2B	10,143
InternVL3.5-2B	2B	10,143	Mini-InternVL2-4B-DA-DriveLM	4B	10,143
InternVL3.5-8B	8B	10,119			

two `-HF` conversions of InternVL3.5 that repeat tokens, and MiniCPM-V-4.5 with hybrid thinking enabled, which spends its budget inside an unclosed thinking block.

C ADDITIONAL RESULTS

C.1 PER-COUNTRY RESULTS FOR ALL BACKBONES

Table 10 reports the accuracy of every evaluated model in each country, together with its D_c. Accuracies are computed as in Table 2: a country's score is the equal-weighted mean of its four task accuracies, and Overall is the mean of all 32 country–task cells (Appendix D.1). The ordering of countries is largely shared across models: Singapore is the weakest country for 9 of the 16 generalists and among the two weakest for 12, whereas Switzerland is among the two strongest for all 16 and India for 14.

Table 10: Accuracy by country for every evaluated model (Direct prompt). Each country column averages that country's four task accuracies equally, and Overall averages the eight columns, as in Table 2, which also gives the 95% intervals on D_c. Countries are ordered by mean accuracy across models, as in Figure 4a. Bold marks each row's strongest country and underline its weakest. Country and source corpus are confounded, so gaps are descriptive.

Model	**Size**	SG	US	DE	CN	UK	JP	IND	CH	**Overall**	D_c
Qwen3.5											
Qwen3.5-2B	2B	31.0	35.4	33.2	40.6	40.5	46.4	**52.7**	47.8	41.0	6.91
Qwen3.5-4B	4B	37.4	39.9	43.7	46.4	54.4	49.7	55.9	**57.7**	48.1	6.83
Qwen3.5-9B	9B	43.3	38.0	44.7	50.1	56.3	52.3	55.7	**62.6**	50.4	7.29
Qwen3-VL											
Qwen3-VL-2B	2B	19.0	28.6	31.1	33.9	27.1	44.3	**52.9**	45.4	35.3	10.39
Qwen3-VL-4B	4B	26.6	34.6	38.2	40.4	29.6	47.7	52.9	**56.8**	40.8	9.98
Qwen3-VL-8B	8B	28.4	40.7	39.2	46.9	34.8	49.0	45.7	**55.8**	42.5	7.88
Qwen2.5-VL											
Qwen2.5-VL-3B	3B	21.4	30.2	30.6	35.0	28.7	29.4	39.7	**49.3**	33.0	7.67
Qwen2.5-VL-7B	7B	30.6	34.6	36.2	34.9	33.9	39.9	49.6	**52.6**	39.0	7.21
InternVL3.5											
InternVL3.5-2B	2B	18.4	11.2	19.3	13.7	17.9	22.8	**33.7**	30.4	20.9	7.07
InternVL3.5-8B	8B	17.9	14.5	20.3	20.1	19.4	20.8	21.2	**37.9**	21.5	6.34
Gemma-4											
Gemma-4-E2B	2B	21.4	25.1	29.2	28.1	27.9	29.6	38.5	**47.1**	30.9	7.42
Gemma-4-E4B	4B	23.7	31.1	32.0	32.6	30.2	37.1	**49.6**	41.3	34.7	7.17
Other lineages											
LLaVA-OV-1.5-8B	8B	37.0	38.4	36.8	36.5	36.9	40.7	41.7	**54.9**	40.4	6.83
MiniCPM-V-4.5	8B	21.2	18.1	17.6	19.0	22.1	20.8	25.0	**36.4**	22.5	5.50
Molmo-7B-D	7B	15.4	16.8	18.0	13.0	17.5	23.3	26.1	**29.9**	20.0	5.26
SmolVLM2-2.2B	2.2B	17.3	14.9	10.4	13.3	18.5	23.1	**27.5**	23.6	18.6	5.22
Driving-specialised (reference only)											
DriveMM	8B	21.2	16.0	18.3	19.2	24.0	23.7	32.7	**44.4**	24.9	9.58
Dolphins	7B	15.5	12.8	16.6	14.3	16.6	20.1	21.9	**33.2**	18.9	5.90
DriveLM-2B	2B	15.4	7.5	7.6	8.0	13.3	10.3	**16.4**	12.8	11.4	3.12
DriveLM-4B	4B	14.9	10.3	8.2	11.5	**18.4**	15.0	17.2	13.5	13.6	2.99

Table 11: Error types per country. Share (%) of 40 sampled errors per country under the Reasoning prompt; each column sums to 100 within a model, and Mean averages the eight countries. Each error is labelled with the earliest stage that failed. A *cultural rule gap* involves a provision that differs between jurisdictions, so the correct action would change in another country; a *general rule violation* involves a provision shared by essentially every country.

Model	Error type	Share of sampled errors per country (%)								Mean
		CN	US	UK	JP	SG	IND	DE	CH	
Qwen3.5-9B	Visual misperception	37.5	47.5	37.5	57.5	47.5	30.0	42.5	45.0	43.1
	Geographic misclassification	0.0	0.0	7.5	2.5	0.0	5.0	0.0	2.5	2.2
	Cultural rule gap	45.0	42.5	42.5	35.0	35.0	52.5	55.0	50.0	44.7
	General rule violation	15.0	7.5	0.0	0.0	12.5	5.0	0.0	0.0	5.0
	Reasoning error	2.5	2.5	12.5	5.0	5.0	7.5	2.5	2.5	5.0
Qwen3-VL-8B	Visual misperception	55.0	27.5	47.5	55.0	40.0	32.5	65.0	50.0	46.6
	Geographic misclassification	0.0	0.0	2.5	5.0	7.5	12.5	0.0	0.0	3.4
	Cultural rule gap	37.5	42.5	40.0	10.0	42.5	45.0	32.5	47.5	37.2
	General rule violation	7.5	30.0	7.5	22.5	7.5	5.0	2.5	0.0	10.3
	Reasoning error	0.0	0.0	2.5	7.5	2.5	5.0	0.0	2.5	2.5
Qwen2.5-VL-7B	Visual misperception	45.0	32.5	42.5	67.5	42.5	42.5	52.5	40.0	45.6
	Geographic misclassification	0.0	0.0	7.5	0.0	2.5	5.0	2.5	2.5	2.5
	Cultural rule gap	20.0	35.0	40.0	22.5	32.5	47.5	35.0	57.5	36.2
	General rule violation	27.5	25.0	7.5	7.5	20.0	5.0	10.0	0.0	12.8
	Reasoning error	7.5	7.5	2.5	2.5	2.5	0.0	0.0	0.0	2.8
InternVL3.5-8B	Visual misperception	65.0	70.0	82.5	55.0	50.0	50.0	72.5	50.0	61.9
	Geographic misclassification	0.0	2.5	0.0	2.5	0.0	10.0	0.0	0.0	1.9
	Cultural rule gap	17.5	15.0	7.5	32.5	35.0	32.5	22.5	42.5	25.6
	General rule violation	10.0	12.5	10.0	5.0	12.5	2.5	2.5	2.5	7.2
	Reasoning error	7.5	0.0	0.0	5.0	2.5	5.0	2.5	5.0	3.4

C.2 Disparity per Regulatory Family

Table 6 reports D_c per regulatory family for every model, the per-model detail behind Figure 4b, with families numbered as in Appendix B.2. By Eq. 2, a value of 0.00 means that the spread across countries within that family is within sampling noise. Because each family covers only part of the test split, its per-model values carry wider intervals than the model-level D_c of Table 2, and we read the table for the pattern that holds across models rather than for any single cell.

C.3 Error Analysis per Country

Table 11 breaks down sampled errors by country for the strongest model of four families: the two of Figure 4c (Qwen3.5-9B and InternVL3.5-8B) together with Qwen3-VL-8B and Qwen2.5-VL-7B. For each model we sample 40 errors per country under the Reasoning prompt, whose trace names the country the model assumed, and assign each error to the earliest stage that failed. An answer that misreads a sign and then cites the wrong rule is therefore counted as visual misperception, so the share of cultural rule gaps is a lower bound.

The pattern of Figure 4c holds in all four models. Geographic misclassification accounts for at most 3.4% of sampled errors in any model; the rest split mainly between visual misperception (43.1–61.9%) and cultural rule gaps (25.6–44.7%), with general rule violations at 5.0–12.8%. With 40 errors per country, an individual country's share has a standard error of up to 8 points, so we read the per-model totals rather than single countries. Appendix F shows one rejected answer per country from these four models, illustrating visual misperception and the cultural rule gap.

D Metric Consistency

D.1 From an Answer to a Number

Answers are free text, so the metric has two parts: a judge that turns one answer into one verdict, and an aggregation that turns the 10,143 verdicts into the number a table reports.

What the judge sees, and what it returns. The judge is Gemma-4-31B-IT under a single deployment for every reported cell. It receives the scene frames in chronological order, the question, the

reference answer with its key points, and a list of acceptable alternative answers. It is not told which model produced the candidate. The template is reproduced verbatim from the released evaluation code:

```
You grade one answer to a driving-scene question. You see the frames in
chronological order; the last frame is the current observation. You are given a
reference answer, the key points a correct answer must contain, and answers
that are also acceptable. You do not know which model produced the candidate,
and you must not speculate about it.

Score these four dimensions from 0 to 5:
- key_point_coverage: how many of the key points the candidate states, with no
  contradiction of any of them.
- correctness: whether the candidate's core claim matches the reference or an
  acceptable alternative. A fluent answer that asserts the wrong fact scores 0.
- local_rule_compliance: whether the candidate respects the traffic convention
  the item depends on. Score 0 when it applies a different region's rule.
- grounding: whether the candidate's stated evidence is actually visible in the
  frames. Penalise invented objects, signs, or signal states.

Then set correct to true only when the candidate would be accepted by a careful
human grader: it states the reference answer or an acceptable alternative and
contradicts none of the key points. Verbosity, hedging, and extra correct
detail neither help nor hurt.

Return one json object with keys: key_point_coverage, correctness,
local_rule_compliance, grounding, total_20, correct, reason.
```

Accuracy is computed from the binary `correct` verdict, which the judge gives as a separate judgement rather than as a threshold on `total_20`; Figure 6a,b compares the two scales. The last clause of the rubric keeps answer length and style from affecting the verdict.

From verdicts to the reported accuracy. Let $y_i \in \{0, 1\}$ be the judge's verdict on item i. Every item carries a country c and a task t, and the reported number is the unweighted mean of the 32 country–task cells rather than the item-level mean:

$$a_{c,t} = \frac{1}{|\mathcal{I}_{c,t}|} \sum_{i \in \mathcal{I}_{c,t}} y_i, \qquad a_c = \frac{1}{4} \sum_t a_{c,t}, \qquad \text{Overall} = \frac{1}{8} \sum_c a_c. \tag{3}$$

A country's task mix is not something we want to measure: Perception makes up between 33% and 61% of a country's items, so an item-level mean would weight each country by how much Perception it happens to contain. Equal weighting also makes Overall and D_c two statistics of the same eight country scores a_c, their mean and their noise-corrected spread. The difference matters in practice: for Qwen3.5-9B the item-level mean is 51.91, whereas Eq. 3 gives 50.37, the value reported in Table 2.

Intervals. Every interval resamples driving sequences within each country–task cell, never individual items, because consecutive questions from one recording are not independent. The same bootstrap supplies the sampling variance that D_c subtracts (Eq. 2).

D.2 Judge Audit

We check the primary judge, Gemma-4-31B-IT, against a human expert and against two judges of unrelated lineage, Llama-3.3-Nemotron-Super-49B and GPT-5.5. All three regraded the same stratified sample of 384 answers, 8 per country for each of six models (Qwen3.5-9B, Qwen3.5-2B, Qwen3-VL-8B, Gemma-4-E4B, Molmo-7B-D and InternVL3.5-8B), without seeing the primary judge's verdicts. The primary judge agrees with the human expert on 89.2% of answers and with the two external judges on 89.8% ($\kappa = 0.77$) and 88.0% ($\kappa = 0.72$), close to the agreement between the two external judges themselves (91.9%, $\kappa = 0.79$); Table 12 summarises these checks. The judge returned a parseable verdict for every answer in every reported run.

Figure 6c compares the model rankings that the three judges induce on the sample. All three rank Qwen3.5-9B first and place Molmo-7B-D and InternVL3.5-8B last; the three models in between change order, which is within the sampling error of the sample (64 answers per model, a standard error of up to about six accuracy points). The primary judge shares a family with two evaluated models, Gemma-4-E2B and Gemma-4-E4B; it ranks Gemma-4-E4B fourth, below the third and second places the two external judges give it, so we see no sign of self-preference.

Table 12: Metric consistency. Checks on the answer judge and on the provenance of every reported number. Judge agreement is measured on 384 answers (8 per country for each of six models) regraded by two judges of unrelated lineage (Appendix D.2).

Check	Criterion	Result
Judge validity		
Agreement with a human expert	$\geq 85\%$	89.2% ($\kappa = 0.76$)
Agreement with Llama-3.3-Nemotron-Super-49B	$\geq 85\%$, $\kappa \geq 0.60$	89.8% ($\kappa = 0.77$)
Agreement with GPT-5.5	$\geq 85\%$, $\kappa \geq 0.60$	88.0% ($\kappa = 0.72$)
Agreement between the two external judges	—	91.9% ($\kappa = 0.79$)
Two best and two worst models, all judges	identical	yes
Parse failures of the primary judge	$\leq 1\%$	0.0% on every run
Evaluation protocol		
Bootstrap resampling unit	—	driving sequence, 2,000 draws
One judge deployment for all reported numbers	required	yes (Gemma-4-31B-IT)
One test split for all reported numbers	required	yes

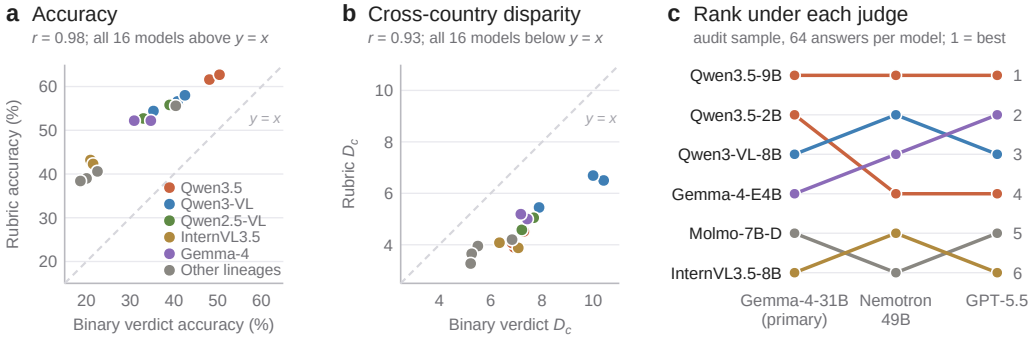

Figure 6: The judge and the scoring scale. a, **b**, Accuracy and D_c of the 16 generalists under the judge's binary verdict, used throughout the paper, and under its graded rubric score; dashed line, $y = x$. **c**, Rank of the six audited models on the 384-answer sample under the primary judge and two judges of unrelated lineage.

Rubric scale. Figure 6a,b rescores all 16 generalists with the judge's graded rubric score (`total_20`, scaled to 0–100) in place of its binary verdict. Every model scores higher on the rubric scale (by 12.3–22.3 points) and every D_c is lower (by 1.6–3.9 points), but the two scales agree on the models ($r = 0.98$ for accuracy and 0.93 for D_c), and D_c stays between 3.3 and 6.7 for every model, so no model becomes even under the rubric. The only pairs of models that change order differ on the binary scale by at most 1.7 accuracy points or 1.6 D_c points, below the two-point difference that Section 4.1 treats as resolvable.

E Validity Controls

E.1 Input Controls: Text-Only and Mismatched Image

Table 13a tests how much of the benchmark can be answered without the matching image, on Qwen3.5-2B and Qwen3.5-9B with the Direct prompt unchanged. Removing the images lowers mean accuracy from 45.7 to 20.6, and pairing each question with the frames of another sequence lowers it to 19.7: a mismatched image is worth no more than none, and more than half of the accuracy obtained with the correct image depends on that image.

Some disparity remains without the image: the text-only control has a D_c of 4.76, against 7.10 with the image, so questions differ across countries in how far they can be answered from the text alone. This component is shared by every model scored on the same items, consistent with the country ordering being largely common across models (Appendix C.1), and it is why the paper compares D_c between models scored on the same items rather than reading its absolute level.

E.2 Within-Country Corpus Comparison

Country and source corpus are confounded by construction, with one exception: Germany contributes both KITTI and Cityscapes. Table 13b uses this exception, and the twelve cities within Cityscapes,

Table 13: Validity controls. a, Input controls, mean of Qwen3.5-2B and Qwen3.5-9B under the unchanged Direct prompt: *text only* removes every image, and *mismatched image* pairs each question with the frames of another sequence under a fixed permutation. **b**, Noise-corrected spread over groups (D_c computed over the groups instead of the countries, with its 95% sequence-bootstrap interval) when the rulebook is held fixed, against the spread between countries.

a Input controls

Input	Acc.	D_c
Image and question	45.7	7.10
Text only	20.6	4.76
Mismatched image	19.7	6.04

b Spread with the rulebook held fixed

Groups compared	Groups	Held fixed	Spread [95% CI]
Cities within Cityscapes (DE)	12	rulebook, corpus	4.95 [3.16, 11.39]
Cityscapes vs. KITTI (DE)	2	rulebook	1.65 [0.00, 5.25]
Countries	8	—	6.39 [5.30, 7.74]

to measure the spread that appears when the rulebook is held fixed. Between the two German corpora the noise-corrected spread is 1.65 $[0.00, 5.25]$, and between the twelve Cityscapes cities 4.95 $[3.16, 11.39]$, against 6.39 $[5.30, 7.74]$ between the eight countries. Both point estimates lie below the spread between countries, but the intervals overlap, so these controls bound the corpus effect rather than separate it from the country effect, and we report country differences as descriptive throughout (Section 3.1). The error analysis gives more direct evidence that rules matter: for the two models of Figure 4c, about one item in five fails on a cultural rule gap, where the scene is read correctly and the answer follows another country's rule (Appendix C.3), which differences in camera, weather or scene composition alone would not produce.

F Demonstration

Figures 7 and 8 provide qualitative examples from all eight countries. The country name is shown in each panel only for exposition; it was withheld from the evaluated model, which received only the scene and question. For readability, the question and reference-answer wording is lightly normalized, while the model responses are reproduced from the corresponding logged runs.

The examples also clarify that a rejected answer is not necessarily a cultural rule gap. Several failures occur before regional knowledge is applied. In the United Kingdom and United States examples, the model fails to recognize the relevant sign. In Japan, it reaches the correct action but misreads both the center-line color and the truck's lane. In Germany, it attends to the nearby mandatory-direction sign instead of the queried no-entry sign. These cases are consistent with the visual-misperception component of the error analysis rather than evidence that the model inferred the wrong jurisdiction.

The Switzerland example is the clearest cultural rule gap: the model correctly refers to Swiss traffic rules but applies the US convention that generally allows a right turn on red after stopping. The remaining cases expose related grounding failures. The Singapore response confuses keeping left of an obstacle with making a left turn; the China response converts a stopping restriction into a no-entry restriction; and the India response predicts a right turn instead of joining the left-hand traffic flow. Together, the examples illustrate why evaluates both scene grounding and jurisdiction-conditioned rule application.

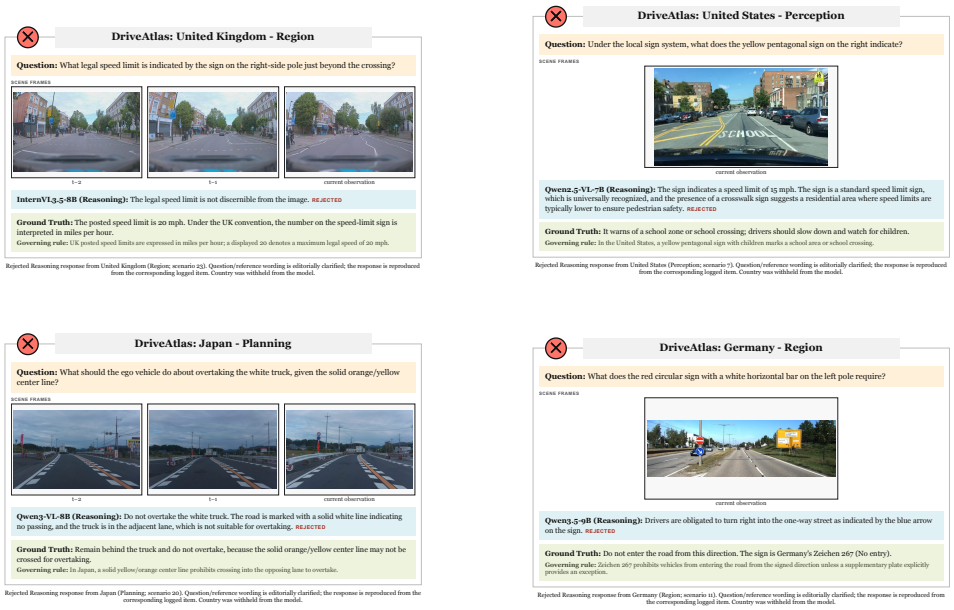

Figure 7: Examples dominated by visual or sign-grounding errors. From top left: United Kingdom, United States, Japan, and Germany. The responses fail to read the relevant sign or scene element correctly; they should therefore not be interpreted as evidence of geographic misclassification or a cultural rule gap.

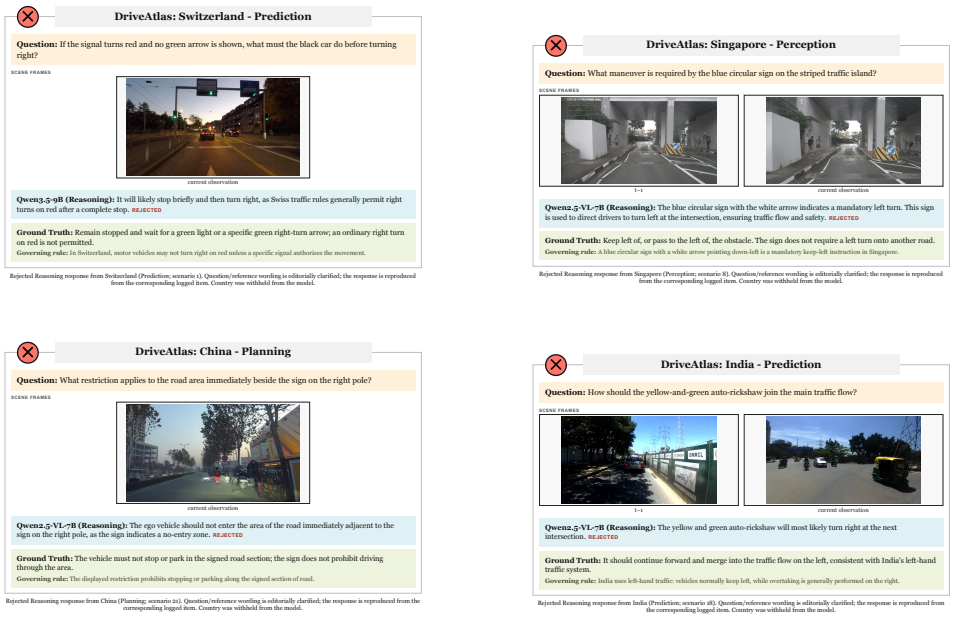

Figure 8: Examples involving regional rules or sign semantics. From top left: Switzerland, Singapore, China, and India. The Switzerland case is an explicit cultural rule gap: the model identifies the jurisdiction but applies another country's right-turn-on-red rule. The other cases show incorrect interpretation of a mandatory direction, stopping restriction, or left-hand traffic flow.