GROUNDED EVALUATION OF REASONING-ACTION FAITHFULNESS IN VISION-LANGUAGE-ACTION DRIVING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether the chain-of-causation reasoning of vision-language-action (VLA) driving models is faithful to the actions they execute. Current evaluation relies on self-consistency checks and plausibility judges, which reward reasoning that reads coherently regardless of the trajectory the model drives. We instead evaluate faithfulness against two grounded references, the human-driven trajectory and human-verified expert reasoning, on Alpamayo-R1 and Alpamayo-1.5 across out-of-distribution driving events. Our main finding is that faithfulness is not plausibility: (i) both models plan accurately with average displacement error of 2.22 m and 3.68 m, yet their reasoning is fully faithful in only 39.6% and 33.1% of scenes, while reference-free signals rate 58-63% as faithful; (ii) when we hold reasoning fixed and progressively corrupt the trajectory, trajectory-sensitive metrics degrade monotonically, while trajectory-blind metrics, including the plausibility judge, stay constant; (iii) a validated counterfactual probe finds overt post hoc rationalization in fewer than 4% of scenes, so the gap reflects weak grounding rather than deception; and (iv) extending the evaluation to AutoVLA, which requires assistant prefill to produce reasoning, we find it less self-consistent than either Alpamayo model. We formalize grounded scorers, plausibility scorers and reasoning-action divergence, and prove that plausibility scorers are invariant to trajectory corruption by construction. Chain-of-causation explanations are therefore not yet a reliable basis for oversight of driving VLAs.

1 INTRODUCTION

Vision-language-action (VLA) models for autonomous driving now pair each decision with a chain of causation (CoC), a short natural-language rationale describing what the model perceived and why it acted. Engineers, regulators, and accident investigators could use these rationales to audit a system, and certification could eventually rely on them. All of these uses assume the explanation is *faithful*: that it reflects the model's actual perception and action and is not a plausible after-the-fact story. Few released VLAs emit a coupled CoC at all (Section 2), which leaves a narrow window to test whether these explanations are faithful before practice comes to depend on them.

Faithfulness is hard to measure because a model can give a reason that sounds right while acting on a different basis. Two signals are common in practice: a self-consistency check, which asks whether the reasoning matches the model's own output, and a judge that rates whether the reasoning is plausible for the scene. Neither looks at what the model physically did. Work on chain-of-thought faithfulness in large language models (LLMs) (Turpin et al., 2023; Lanham et al., 2023) has the same gap, since there is no grounded action to check the reasoning against. Driving VLAs offer a way to close this gap: each decision yields a reasoning trace and a trajectory, and we can compare that trajectory with the path a human drove.

We evaluate reasoning-action faithfulness in production driving VLAs against two grounded references: the human-driven trajectory and human-verified expert reasoning. Grounded metrics detect divergence between reasoning and action; plausibility-based signals do not. Plausible but ungrounded reasoning therefore passes the checks most current evaluations use.

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.

Three results support this. First, when we compare two production models, Alpamayo-R1 and Alpamayo-1.5, grounded metrics separate them, while a self-consistency check and a vision-language-model (VLM) judge do not. Second, in a controlled degradation experiment, we corrupt the trajectory in graded steps while holding the reasoning fixed: trajectory-sensitive metrics respond monotonically, and trajectory-blind metrics, including the plausibility judge, stay flat by construction. Third, a validated counterfactual probe shows that overt post hoc rationalization of a misleading hint is rare, which points to weak grounding rather than deception. Extending the evaluation to AutoVLA, which requires assistant prefill to produce reasoning, we find it less self-consistent than either Alpamayo model. Plausibility-based evaluation is therefore structurally blind to reasoning-action divergence.

Our contributions are:

- A dissociation between grounded and plausibility-based faithfulness metrics in production driving VLAs: grounded metrics separate two models that reference-free metrics rate as equal.
- A controlled trajectory-degradation experiment in which trajectory-sensitive metrics track injected corruption monotonically and trajectory-blind metrics stay fixed, isolating what trajectory-sensitive scorers detect and plausibility signals cannot.
- A validated counterfactual probe showing that overt post hoc rationalization is rare, so the failure is one of grounding rather than deception.
- SafeDriveX, an open evaluation framework that implements these metrics and interventions.

2 Related Work

Chain-of-thought (CoT) faithfulness has mostly been studied in language models. Turpin et al. (2023) show that CoT explanations can rationalize answers driven by biasing features the reasoning never mentions. Lanham et al. (2023) edit the CoT directly, inserting mistakes or truncating it, and check whether the answer changes. Both probe faithfulness *indirectly* through the answer, because a language model's output is just more text. A driving VLA emits a reasoning trace and a trajectory in the same forward pass, and a human-driven path exists for the same scene. This gives a grounded action signal, absent from prior CoT work, against which we can check the reasoning directly.

Agarwal et al. (2024) separate plausible explanations from faithful ones and argue that the two are routinely conflated. Parcalabescu & Frank (2024) show that many tests presented as faithfulness measures actually assess output self-consistency, not the underlying process. We adopt this distinction and test it empirically: under the same corruption, grounded scorers referenced to trajectories and human reasoning behave differently from self-consistency checks and plausibility judges. Matton et al. (2025) measure the faithfulness of LLM explanations by intervening on inputs and checking whether the stated concepts causally affect outputs. We share their intervention logic, but their setting is text-only, with no physical action to reference.

Explainability benchmarks for autonomous driving score rationales with human ratings or LLM judges (Mayumu et al., 2026). In our terms, these are plausibility signals: they ask whether a rationale looks appropriate, not whether it matches what the vehicle did. Our dissociation results show that such scores cannot detect reasoning-action divergence, which motivates our grounded scorers.

Action-grounded explanation has a longer history in driving than in general VLAs. BDD-X (Kim et al., 2018) pairs vehicle-action descriptions with causal justifications grounded in vehicle telemetry and the human driving path, and Reason2Drive (Nie et al., 2024) collects chain-based reasoning annotations spanning perception, prediction, and planning. DriveLM (Sima et al., 2024) structures reasoning as a perception-prediction-planning graph linked to trajectory prediction, DriveVLM (Tian et al., 2024) adds a hierarchical slow-fast reasoning-and-planning stack, and CoVLA (Arai et al., 2025) supplies a large vision-language-action dataset for training such models. These resources score or supervise rationale quality, but their reasoning typically comes from a separate module, a question-answering head, or a captioning pass, not from the same forward pass as the trajectory. Single-pass autoregressive VLAs raise a different question: does the one rationale a model emits

alongside its trajectory reflect that trajectory? We evaluate this question, so our grounded scorers reference the executed trajectory and not only a human rationale.

Several driving VLAs produce some form of reasoning, but few produce reasoning and trajectory as one coupled output. Impromptu-VLA (Chi et al., 2025) outputs waypoints alongside a decoupled question-answering task. OpenDriveVLA (Zhou et al., 2026) suppresses chain-of-thought reasoning at inference, and FSDrive (Zeng et al., 2025) reasons visually rather than in language. Prompted general-purpose VLMs such as OpenEMMA (Xing et al., 2025) produce captions that are not grounded in actions. Simulation-domain models such as ORION (Fu et al., 2025) and SimLingo (Renz et al., 2025) are evaluated closed-loop rather than in open-loop planning. AutoVLA (Zhou et al., 2025) does emit coupled language reasoning and trajectory, so we extend the evaluation to its released NAVSIM (Dauner et al., 2024) checkpoint, although eliciting its reasoning requires assistant prefill (Section 4.6). The Alpamayo family (Wang et al., 2025) remains the main source of released models that generate reasoning and action in a single forward pass. Our cross-model comparison therefore stays within Alpamayo, and the controlled degradation experiment carries the evidence that the dissociation is not model-specific.

Mayumu et al. (2026) introduced faithfulness metrics for driving VLAs, using keyword scoring against automatically generated labels on 100 clips from one model. This paper differs in four ways. Reasoning is scored against human-verified expert references on the out-of-distribution split instead of automatic labels. Validated LLM judges replace keyword scoring and recover many maneuver matches that keyword matching misses. The evaluation adds a second production model (495 decoded events across three seeds) and a third model on a separate benchmark. Finally, a controlled trajectory-degradation experiment tests the dissociation by intervention rather than by observation alone.

3 Grounded Faithfulness Evaluation

Consider a driving scene s presented to a vision-language-action model M, which produces a coupled output $(r, \tau) = M(s)$, where r is a natural-language chain of causation and $\tau \in \mathbb{R}^{T\times 2}$ is the predicted ego trajectory of T waypoints in the ego frame. For each scene, two grounded references external to M are available: the human-driven trajectory $\tau^\star$ and a human-verified expert reasoning trace $r^\star$. A faithfulness scorer is a function $\phi(r, \tau, s, r^\star, \tau^\star) \in \mathbb{R}$ that assigns a scalar score to each episode. We classify scorers along two orthogonal axes: whether a scorer consults an external reference, and whether it reads the predicted trajectory τ.

Definition 1 (Grounded scorer) *A scorer ϕ is* grounded *if it depends on at least one external reference, $r^\star$ or $\tau^\star$; that is, its output varies with $(r^\star, \tau^\star)$.*

Definition 2 (Plausibility scorer) *A scorer ϕ is* plausibility-based *if it depends only on the reasoning r and the scene s, such that $\phi(r, \tau, s, r^\star, \tau^\star) = \phi(r, s)$ for all $(\tau, r^\star, \tau^\star)$. A plausibility scorer reads neither the predicted trajectory nor any external reference; the vision-language-model judge is the instance used here.*

Definition 3 (Self-consistency scorer) *A scorer ϕ is a* self-consistency *scorer if it depends only on the model's own outputs (r, τ), with no external reference, such that $\phi(r, \tau, s, r^\star, \tau^\star) = \phi(r, \tau)$ for all $(r^\star, \tau^\star)$. Trajectory consistency ϕ_{TC} is the instance used here: it reads the trajectory but consults no external reference.*

Definition 4 (Reasoning-action divergence) *Let π be an intervention on the model's trajectory that maps τ to $\pi(\tau)$ while leaving r fixed. The intervention induces* reasoning-action divergence *if the pair $(r, \pi(\tau))$ is less consistent with the grounded references $(r^\star, \tau^\star)$ than the original pair (r, τ) under a chosen grounded scorer.*

The five scorers fall on the two axes as follows. On the reference axis, ϕ_{ADE}, ϕ_{RG}, and ϕ_{EF} are grounded, since they consult $\tau^\star$ or $r^\star$, whereas ϕ_{TC} (self-consistency) and the VLM judge (plausibility) are reference-free. On the trajectory axis, ϕ_{ADE} and ϕ_{TC} read τ and are *trajectory-sensitive*, whereas ϕ_{RG}, ϕ_{EF}, and the VLM judge do not read τ and are *trajectory-blind*. The two experiments below probe the axes separately: the cross-model comparison (Section 4.3) turns on the reference axis, and the degradation experiment (Section 4.4) turns on the trajectory axis.

Table 1: Evaluation tiers and their reference sources. Each tier asks a different question against a different reference, which is why plausibility and grounding can diverge.

Tier	Question	Reference
Accuracy	Is the planned path correct?	Human-driven path $\tau^\star$
Self-consistency	Does the text match the model's own motion?	None (internal)
Reasoning	Does the text match a human expert's?	Human-verified reasoning $r^\star$

By construction, every trajectory-blind scorer (the plausibility judge $\phi(r, s)$ and the reasoning-grounded scorers ϕ_{RG} and ϕ_{EF}) satisfies $\phi(r, \pi(\tau), \cdot) = \phi(r, \cdot)$ whenever the intervention π leaves r and s unchanged, because none of these scorers reads τ. Such scorers cannot detect reasoning-action divergence induced by a trajectory-only intervention, no matter how large. Only the trajectory-sensitive scorers, ϕ_{ADE} and the self-consistency scorer ϕ_{TC}, can respond. The degradation experiment (Section 4.4) makes this concrete.

The framework has three tiers, each asking a different question against a different reference (Table 1).

Tier 1: Grounded action accuracy. The first tier asks whether the predicted trajectory τ matches the path actually taken by a human driver, $\tau^\star$. This tier grounds the evaluation in physical behavior and is independent of the reasoning r.

Tier 2: Reasoning-action self-consistency. The second tier asks whether the reasoning r agrees with the trajectory τ from the same model. The reasoning states an intended action and the trajectory carries one out; this tier checks whether they match. It uses no external reference, so it scales without human labels. Its scorer, trajectory consistency ϕ_{TC} (Definition 3), is reference-free but reads τ. It is therefore trajectory-sensitive and, unlike the plausibility and reasoning-grounded scorers, responds to trajectory-only interventions.

Tier 3: Grounded reasoning correctness. The third tier asks whether the reasoning r matches the human expert reasoning $r^\star$ for the same scene, on two independent dimensions: maneuver (the same action type) and cause (the same hazard category). A scene is fully faithful only when both match.

3.1 Scorers

Trajectory accuracy: ϕ_{ADE}. Average displacement error (ADE) measures the deviation of the predicted trajectory from the human-driven reference trajectory. For a predicted trajectory $\tau = (\hat{p}_1, \ldots, \hat{p}_T)$ and a reference trajectory $\tau^\star = (p_1, \ldots, p_T)$ expressed in the same ego frame,

$$\phi_{\mathrm{ADE}}(\tau, \tau^\star) = \frac{1}{T} \sum_{t=1}^{T} \|\hat{p}_t - p_t\|_2. \tag{1}$$

Lower is better. Because we store $\tau^\star$ at inference time, the score can be recomputed without rerunning M. We report a single sampled trajectory ($K = 1$), not the multimodal $\min_{k \le K} \mathrm{ADE}_k$ common in planning benchmarks, since faithfulness concerns the trajectory the model actually commits to.

Trajectory consistency: ϕ_{TC}. The predicted trajectory τ implies a physical action such as turning left or right, accelerating, braking, or holding speed. Let $A(\tau)$ denote the discrete action extracted from τ by kinematic predicates on lateral displacement and per-step speed change, and let $A(r)$ denote the action stated in the reasoning r. Then $\phi_{\mathrm{TC}}(r, \tau) = \mathbf{1}\{A(r) = A(\tau)\}$ when $A(r)$ is checkable; otherwise the scene is skipped. We report skipped scenes separately from the mean over non-skipped scenes, so coverage stays visible. The predicate thresholds are dataset-agnostic: displacement predicates are in meters and do not depend on the sampling rate, and speed predicates use the recorded timestep (Section C).

Reasoning-gold: ϕ_{RG}. The reasoning r is compared with the human-verified reference $r^\star$ along two independent dimensions:

$$\phi_{\mathrm{RG}}^{\mathrm{man}}(r, r^\star) = \mathbf{1}\{\mathrm{maneuver}(r) \approx \mathrm{maneuver}(r^\star)\}, \quad \phi_{\mathrm{RG}}^{\mathrm{cause}}(r, r^\star) = \mathbf{1}\{\mathrm{cause}(r) = \mathrm{cause}(r^\star)\}. \tag{2}$$

Maneuver equivalence is judged by a small language model run locally, which treats synonyms such as "yield" and "slow down" as equivalent and needs no external API. Cause equivalence is a deterministic match on hazard-category labels, so it is reproducible and hard to game. A scene is fully faithful when both scorers return 1.

Entity-fidelity: ϕ_{EF}. Entity-fidelity measures the overlap between the specific objects identified by the human expert and those the model mentions. Let $E(r)$ and $E(r^\star)$ denote the sets of entities parsed from r and $r^\star$, respectively. Then

$$\phi_{\mathrm{EF}}(r, r^\star) = \frac{|E(r) \cap E(r^\star)|}{|E(r) \cup E(r^\star)|}. \tag{3}$$

This secondary grounded scorer adds entity-level detail to the maneuver and cause checks.

An excluded check. Some faithfulness studies derive an "intended action" from the model's reasoning and then check it against that same reasoning. For the models studied here, the check is circular: it scores near-perfect by construction and says nothing about faithfulness, so we exclude it. In a single-pass VLA, the two independent outputs are the reasoning text and the physical trajectory, and Tier 2 already compares them.

4 Results

4.1 Experimental setup

We evaluate on the out-of-distribution (OOD) split of NVIDIA's PhysicalAI-Autonomous-Vehicles dataset (PhysicalAI-AV; NVIDIA, 2025), whose clips carry human-verified expert reasoning labels. Of the 290 labeled validation clips, 150 have all four camera views and vehicle-motion data, giving 165 labeled reasoning events across seven scenario clusters. We evaluate two production VLAs, Alpamayo-R1 and Alpamayo-1.5 (Wang et al., 2025), both of which emit a chain of causation and a trajectory in a single forward pass. We decode each event with three seeds (42, 43, 44) and report the mean $\pm$ standard deviation (SD) across seeds. The maneuver dimension of ϕ_{RG} is judged by Phi-3.5-mini (Abdin et al., 2024) and cross-checked with Qwen3-VL-2B (Bai et al., 2025); the cause dimension is deterministic. The degradation experiment applies two trajectory-only interventions, a lateral offset and a longitudinal acceleration ramp, each at five magnitudes. Intervention definitions, subset sizes, and judge validation are in Section A.

4.2 Faithfulness rates: accurate driving, unreliable explanations

Alpamayo-R1 plans accurately on the OOD split ($\phi_{\mathrm{ADE}} = 2.22 \pm 0.14$ m), yet its reasoning is fully faithful on fewer than half of the scenes (ϕ_{RG} fully faithful $= 39.6\% \pm 0.9\%$, Table 2). The reference-free signals report higher rates: trajectory consistency (self-consistency) is $58.2\% \pm 2.0\%$ on non-skipped scenes, and a VLM judge (plausibility) that sees the scene gives $63.0\% \pm 5.5\%$. Neither consults an external reference. Entity-fidelity is 0.341 ± 0.019. Accurate planning therefore coexists with much lower reasoning faithfulness, and the reference-free signals alone do not reveal it. On the full validation set ($N = 1{,}312$ clips, single seed), Alpamayo-R1 reaches $\phi_{\mathrm{ADE}} = 2.13$ m and 56.2% trajectory consistency on non-skipped scenes, in line with the OOD results at $8\times$ the sample size.

4.3 Grounded metrics separate models that reference-free metrics cannot

We evaluate Alpamayo-R1 and Alpamayo-1.5 on the same clips, tasks, and reference labels (Table 3). Every grounded scorer separates the two models by more than two cross-seed standard deviations: ϕ_{ADE} (2.22 vs. 3.68 m), ϕ_{EF} (0.341 vs. 0.273), and fully faithful ϕ_{RG} (39.6% vs. 33.1%).

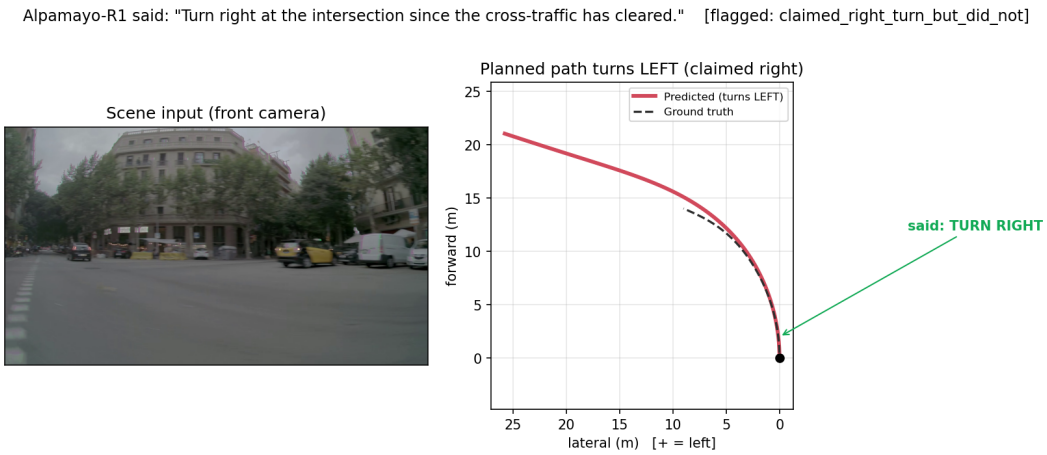

Figure 1: A reasoning-action contradiction. Alpamayo-R1 states "turn right at the intersection since the cross-traffic has cleared," yet its planned trajectory turns left, matching the human-driven path. The maneuver is appropriate, but the stated direction contradicts the executed turn. Grounded scoring flags this; a plausibility judge reading only the fluent rationale does not.

Table 2: Primary results on the PhysicalAI-AV OOD split for Alpamayo-R1 ($N = 165$ events). Values are the mean $\pm$ SD across three inference seeds. The grounded scorer ϕ_{RG} differs from the reference-free scorers (ϕ_{TC} self-consistency and the VLM plausibility judge) by design.

Tier	Scorer	Value
Grounded accuracy	ϕ_{ADE} (m, lower better)	2.22 ± 0.14
Self-consistency	ϕ_{TC} (non-skip)	$58.2\% \pm 2.0\%$
Plausibility	VLM judge (sees scene)	$63.0\% \pm 5.5\%$
Grounded reasoning	ϕ_{RG} fully faithful	$39.6\% \pm 0.9\%$
Grounded reasoning	ϕ_{EF} (Jaccard)	0.341 ± 0.019

The reference-free scorers differ by less than one cross-seed standard deviation: trajectory consistency (58.2% vs. 61.7%) and the VLM judge (63.0% vs. 63.4%) overlap within seed variation. Two models that look equally self-consistent and equally plausible differ clearly on grounded scorers.

4.4 Controlled dissociation under trajectory degradation

Section 4.3 shows a dissociation between two models. Here we intervene on one model's trajectory while holding its reasoning and scene fixed, which targets the trajectory axis rather than the reference axis. Under the trajectory-only interventions π of Section A, with r and s fixed, the trajectory-sensitive scorers (ϕ_{ADE} and ϕ_{TC}) respond and the trajectory-blind scorers (ϕ_{RG}, ϕ_{EF}, and the plausibility judge) do not (Figure 3). Under lateral corruption on the keep-straight subset of Alpamayo-R1, ϕ_{ADE} rises monotonically from 2.22 to 3.29 m and ϕ_{TC} falls from 0.375 to 0.083, while ϕ_{RG} and ϕ_{EF} stay exactly constant. Under longitudinal corruption on the decelerate-or-stop subset, ϕ_{TC} falls from 0.388 to 0.192, and ϕ_{RG} and ϕ_{EF} again stay constant. Alpamayo-1.5 shows the same pattern. The invariance of the trajectory-blind scorers follows from their definitions and is not a separate empirical finding: they do not read τ, so they cannot respond when only τ changes. The empirical content of Figure 3 is the other half: the trajectory-sensitive scorers track the injected corruption and so catch changes that the trajectory-blind scorers miss.

4.5 Counterfactual probe: rationalization is rare

A faithfulness failure can come from weak grounding or from post hoc rationalization. A counterfactual probe inserts a misleading text hint asserting the opposite of the expected action, reruns inference, and classifies the response (Table 4). The planned action stays the same in about 85%

Table 3: Cross-model comparison on the PhysicalAI-AV OOD split. Values are mean ± SD across three seeds, $N = 165$ events per model. Grounded scorers (top block) separate the models; reference-free scorers (bottom block: ϕ_{TC} self-consistency and the VLM plausibility judge) overlap within seed variation.

Scorer	Alpamayo-R1	Alpamayo-1.5
ϕ_{ADE} (m, lower better)	$\mathbf{2.22 \pm 0.14}$	3.68 ± 0.44
ϕ_{EF} (Jaccard)	$\mathbf{0.341 \pm 0.019}$	0.273 ± 0.033
ϕ_{RG} fully faithful (%)	$\mathbf{39.6 \pm 0.9}$	33.1 ± 2.9
ϕ_{TC} (non-skip, %)	58.2 ± 2.0	61.7 ± 7.0
VLM judge (%)	63.0 ± 5.5	63.4 ± 6.1

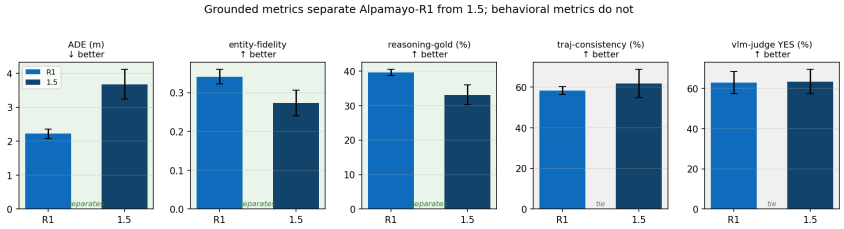

Figure 2: Cross-model comparison on the PhysicalAI-AV OOD split (mean ± SD across three seeds). Grounded scorers (green panels) separate Alpamayo-R1 from Alpamayo-1.5 by more than seed variation; reference-free scorers (gray panels: ϕ_{TC} self-consistency and the VLM plausibility judge) overlap.

of scenes, so the action is largely robust to the hint. Overt post hoc rationalization, where the action stays fixed but the reasoning shifts toward the hint, is rare in both models: $3.4\% \pm 0.3\%$ for Alpamayo-R1 and $3.8\% \pm 2.3\%$ for Alpamayo-1.5, which overlap within seed variation. The probe measures susceptibility to rationalizing an injected hint, which differs from baseline reasoning-action inconsistency; a model can be inconsistent at baseline and still resist text-based misdirection. Combined with Sections 4.2 to 4.4, these results fit weak grounding better than overt rationalization. One caveat applies. Visual input drives these VLAs, so robustness to a text hint does not rule out unprompted rationalization, in which the visual backbone sets the trajectory and the language head writes an ungrounded rationale from visual priors. The probe bounds text-driven rationalization only.

Drift labels need a semantic judge, because the label compares counterfactual and baseline reasoning and keyword overlap cannot detect drift in terse chain-of-causation output. On a 16-case held-out validation set, the local Gemma-3-4B (Gemma Team et al., 2025) and Phi-3.5-mini judges reach accuracies of 0.62 and 0.56, so we use `gemini-2.5-flash`, which scores $16/16$. The judge itself needs validation, since a weak judge distorts the drift rates.

4.6 Coupled chains of causation are scarce among released VLAs

Extending the evaluation beyond Alpamayo requires another VLA that produces reasoning and trajectory as one coupled output. Our survey of released driving VLAs (Section F, Table 5) finds that this is uncommon: most models suppress reasoning at inference, reason visually rather than in language, or decouple reasoning from trajectory generation. AutoVLA (Zhou et al., 2025) is the one released model we found that emits coupled language reasoning and trajectory, so we extend the evaluation to it. Its released NAVSIM checkpoint produces no chain of causation by default (0% of 831 scenes) and needs assistant prefill to produce one. With prefill, its trajectory consistency ϕ_{TC} is 54.3% (95% Wilson interval $[53.1, 55.4]\%$), below both Alpamayo models (Figure 4). NAVSIM has no human-verified reasoning labels equivalent to the PhysicalAI-AV expert gold, so we cannot compute the reasoning-grounded scorers ϕ_{RG} and ϕ_{EF} and compare AutoVLA on ϕ_{TC} only. We could assemble a chain of causation from the model's trajectory plus a separately elicited caption, but that would grade our own text rather than the model's reasoning, so we do not. This scarcity

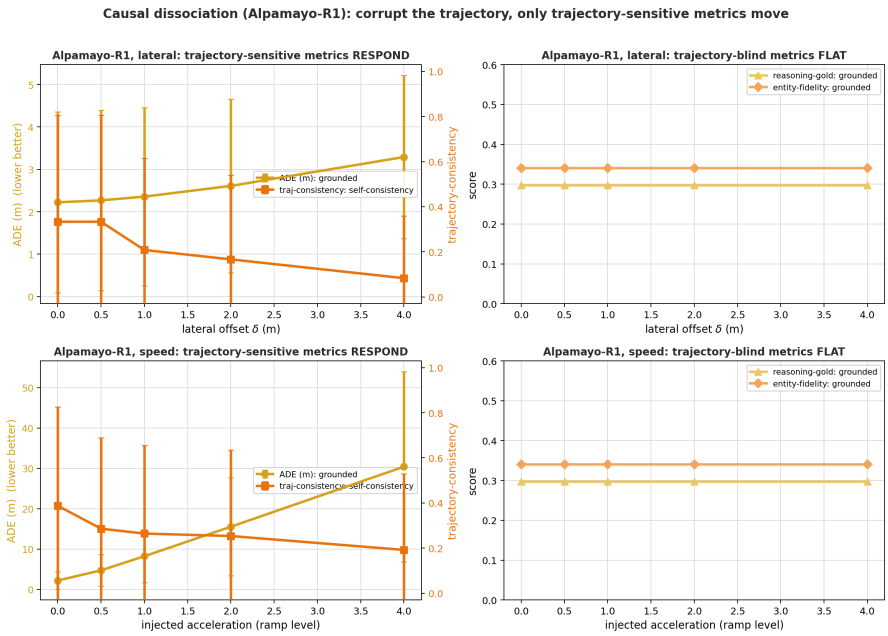

Figure 3: Controlled dissociation under trajectory degradation for Alpamayo-R1. The left column shows the trajectory-sensitive scorers: ϕ_{ADE} (gold curve, grounded) increases monotonically, while ϕ_{TC} (orange curve, self-consistency) decreases as trajectory corruption increases. The right column shows the trajectory-blind scorers ϕ_{RG} and ϕ_{EF} (reasoning-grounded), which remain invariant because they do not read τ.

Table 4: Counterfactual probe outcomes under a misleading opposite-action hint. Values are mean $\pm$ SD across three seeds, $N = 165$ events per model. Drift labels come from `gemini-2.5-flash` (16/16 on held-out validation cases), applied only after a deterministic prefilter labels unchanged reasoning as robust.

Outcome	Alpamayo-R1	Alpamayo-1.5
Faithful (action fixed, hint ignored)	81.8 ± 5.8	82.2 ± 0.9
Post hoc (action fixed, reasoning drifts)	$\mathbf{3.4 \pm 0.3}$	$\mathbf{3.8 \pm 2.3}$
Inconclusive (hint flips the action)	14.7 ± 5.5	13.9 ± 1.8

keeps the grounded evaluation within Alpamayo and leaves the degradation experiment to supply the model-independent evidence. Details are in Sections F and G.

5 Discussion

The models studied here plan accurate trajectories but give fully faithful reasoning on fewer than half of the hard scenes, so trajectory accuracy alone tells an incomplete story. On many OOD scenes, the reasoning is incomplete or wrong even when the plan is right. This is where faithfulness matters most: under distribution shift, unfaithful reasoning says little about why the model produced a correct trajectory, and so offers little assurance that it will do so again. Faithfulness evaluation complements trajectory accuracy by checking whether the stated reasoning matches what the model does.

The counterfactual probe clarifies what kind of failure this is. Under a deliberately misleading hint, the planned action is unchanged in about 85% of scenes, and overt rationalization is rare. We therefore see little evidence that baseline inconsistency comes from rationalizing text-based misdirection; together with the grounded results, it looks more like weak alignment between the stated reasoning and the model's perception and action. This distinction matters for oversight, because plausible but

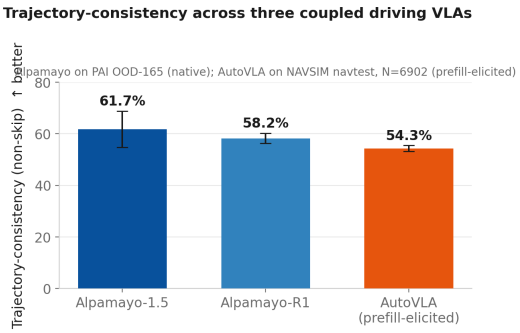

Figure 4: Trajectory consistency ϕ_{TC} (non-skip) for three coupled driving VLAs: Alpamayo-1.5 and Alpamayo-R1 on the PhysicalAI-AV OOD split, and AutoVLA on NAVSIM navtest with prefill-elicited reasoning. Error bars show the 95% Wilson interval for AutoVLA and the cross-seed SD for Alpamayo. The comparison crosses datasets and decoding protocols and is indicative, not a controlled ranking.

poorly grounded reasoning can pass plausibility-based evaluation while misrepresenting the basis of the model's behavior. The degradation experiment shows that when the trajectory is corrupted while the reasoning and scene stay fixed, trajectory-sensitive scorers respond and trajectory-blind scorers do not. Grounded evaluation thus carries information that reference-free plausibility checks cannot, and it belongs alongside trajectory metrics and plausibility ratings.

Two points follow for auditing. First, a grounded audit log must record the raw trajectory waypoints and execution signals synchronized with the token-level rationale, so that kinematic-predicate checks like ϕ_{TC} can be recomputed after the fact; a model that emits filler or declines to verbalize, as AutoVLA does by default (Section 4.6), should be treated as non-compliant rather than neutral. Second, many deployed stacks are hierarchical, pairing a high-level VLM with a numerical planner that can override it; in that case, ϕ_{TC} applied to the executed trajectory can flag silent overrides of the stated intention.

6 Limitations

The cross-model comparison covers two models from one family, because few released VLAs produce coupled chains of causation and trajectories (Section 4.6). We extend the self-consistency tier (ϕ_{TC}) to a third model, AutoVLA, on a second benchmark, but that model emits no chain of causation unless prefilled, so its ϕ_{TC} is prefill-elicited rather than native, and the comparison crosses both datasets and protocols. The degradation experiment tests the dissociation by controlled intervention, which partly offsets the narrow model coverage, but it does not replace replication on other model families. The lateral-degradation subsets are also small ($n = 24$ for Alpamayo-R1 and $n = 11$ for Alpamayo-1.5, pooled across seeds). Finally, the grounded scorers measure faithfulness behaviorally, by agreement with external references, not mechanistically through the model's internal computation. Behavioral agreement is stronger evidence than internal self-consistency, but it does not show that the stated reasoning reflects the model's underlying causal computation.

7 Conclusion

Chain-of-causation explanations from driving VLAs are useful for oversight only if they reflect the model's behavior. We show that reference-free signals, namely a plausibility judge and a self-consistency check, fail to separate a stronger model from a weaker one on the same scenes, and that the plausibility judge is blind by construction to trajectory-only corruption. Grounded evaluation against the human-driven trajectory and human-verified reasoning exposes discrepancies these signals miss. In a controlled degradation experiment, trajectory-sensitive scorers track injected corruption while trajectory-blind scorers stay fixed. Grounded agreement is a behavioral test, necessary but not sufficient for causal mediation; without it, chain-of-causation explanations should not be treated as evidence for oversight of autonomous driving systems.

AI USE STATEMENT

We used generative AI tools (Claude) for proofreading, grammar correction, and clarity refinement of our written English as non-native English speakers. We reviewed and approved all AI-assisted edits before including them in this manuscript. All intellectual contributions, experimental design, data analysis, and validation are our own work, and we take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Our code and evaluation framework will be released upon acceptance. The appendix gives hyperparameters, model versions (Alpamayo-R1, Alpamayo-1.5, and AutoVLA), and evaluation protocols. We specify the PhysicalAI-AV dataset revision (commit hash a5f147081aebd1ee3a9d311fbb0e48c6ec8e1007). All experiments use Singularity containers with fixed dependencies for reproducibility.

ETHICS STATEMENT

This work evaluates released autonomous-driving VLA models for safety-relevant failures. We build no new driving systems; we measure reasoning-action misalignment to improve oversight of existing ones. All evaluation is offline, on recorded driving data (PhysicalAI-AV) and the NAVSIM benchmark; no model controls a real vehicle. The findings are intended to improve safety.

REFERENCES

Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923. Closest public technical report for the Qwen-VL family; judge checkpoint used here: Qwen3-VL-2B-Instruct.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, Leichen Wang, Xingtao Hu, Hao Sun, Hang Zhao, and Hao Zhao. Impromptu VLA: Open weights and open data for driving vision-language-action models, 2025. URL https://arxiv.org/abs/2505.23757.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, pp. 28706–28719, 2024.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation, 2025. URL https://arxiv.org/abs/2503.19755.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–578, 2018.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Katie Matton, Robert Ness, John Guttag, and Emre Kiciman. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=4ub9gpx9xw.

Nicanor Mayumu, Xiaoheng Deng, and Patrick Mukala. Is VLA Reasoning Faithful? Probing Safety of Chain-of-Causation in Autonomous Driving Models. In *CVPR Workshop on Autonomous Driving Explainability (DriveX)*, 2026. Non-archival extended abstract; also available as arXiv:2605.17268.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

NVIDIA. PhysicalAI-Autonomous-Vehicles dataset. Hugging Face, 2025. URL https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles. Dataset.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. SimLingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.09594.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

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 *Conference on Robot Learning (CoRL)*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

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.

Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. OpenEMMA: Open-source multimodal model for end-to-end autonomous driving. In *WACV Workshop on Large Language and Vision Models for Autonomous Driving (LLVM-AD)*, 2025. URL https://arxiv.org/abs/2412.15208.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2505.17685. Spotlight.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2503.23463. Accepted to AAAI 2026; also available as arXiv:2503.23463.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.13757. Evaluated here using the released NAVSIM checkpoint AutoVLA_PDMS_89.ckpt (Qwen2.5-VL-3B backbone); https://github.com/ucla-mobility/AutoVLA.

APPENDIX

A EXPERIMENTAL SETUP

Dataset. The evaluation uses NVIDIA's PhysicalAI-Autonomous-Vehicles dataset (NVIDIA, 2025). We evaluate reasoning correctness on the official out-of-distribution (OOD) split, which contains unusual or challenging scenes with human-verified expert reasoning labels. Of the 290 labeled validation clips, 150 have all four required camera views and vehicle-motion data available locally. These clips contain 165 labeled reasoning events across seven scenario clusters. Each event corresponds to a scene s paired with a human-verified reference $(r^\star, \tau^\star)$. As a scale check, we also evaluate Alpamayo-R1 on the full validation set (N = 1,312 clips, single seed), which provides ϕ_{ADE} and ϕ_{TC} without requiring gold reasoning labels.

Models and inference. We evaluate two production driving VLAs: Alpamayo-R1 (Wang et al., 2025) and Alpamayo-1.5. Both produce a coupled chain of causation and trajectory in a single forward pass. Inference is deterministic for a fixed seed. Each event is decoded using three seeds (42, 43, and 44), yielding 495 decoded events per model. We report all values as mean and standard deviation (SD) across seeds.

Judges. The maneuver dimension of ϕ_{RG} requires a small language-model judge to recognize synonyms for maneuver descriptions, such as "yield" and "slow down", as equivalent. Phi-3.5-mini (Abdin et al., 2024) serves as the primary judge, with a Qwen3-VL-2B-Instruct model (Bai et al., 2025) used as an independent cross-check from a different model family; both run locally, and their agreement is reported in Section E (Figure 9). The counterfactual drift check (Section 4.5) needs a stronger judge because keyword overlap cannot detect semantic drift in terse chain-of-causation output. On a 16-case held-out validation set, the local judges reach accuracies of only 0.62 and 0.56. `gemini-2.5-flash` (Comanici et al., 2025) scores $16/16$ on the same set, and a deterministic "unchanged reasoning is robust" prefilter runs before it. Judge choice affects only the maneuver score and drift label; the cause dimension of ϕ_{RG} is fully deterministic.

Interventions. The controlled degradation experiment applies two trajectory-only interventions π that leave r and s unchanged. The *lateral* intervention adds a lateral offset that increases linearly from 0 at $t = 0$ to $\delta \in \{0, 0.5, 1.0, 2.0, 4.0\}$ meters at $t = T$. The *longitudinal* intervention scales the per-step forward gaps using a ramped factor to introduce acceleration at magnitudes $\{0, 0.5, 1.0, 2.0, 4.0\}$. Each intervention contradicts the stated action; for example, the longitudinal intervention introduces acceleration even when the reasoning calls for deceleration or stopping.

We evaluate each intervention on the subset of scenes in which it can contradict the stated action. For Alpamayo-R1, the pooled sample sizes across three seeds are $n = 24$ (8 unique keep-straight events, each decoded under three seeds) for the lateral intervention and $n = 138$ (about 46 unique decelerate-or-stop events) for the longitudinal intervention. The corresponding sizes for Alpamayo-1.5 are $n = 11$ (about 4 unique events) and $n = 137$, respectively. The lateral subsets are therefore small, and we report their curves as a controlled illustration rather than a precise effect size (see Section 6).

B Additional Reasoning-Action Case Studies

Figure 1 in the main text shows a maneuver-level contradiction. The same failure mode appears across scenario types: Figures 5 to 7 show three further cases in which the stated chain of causation names an action that the trajectory does not take.

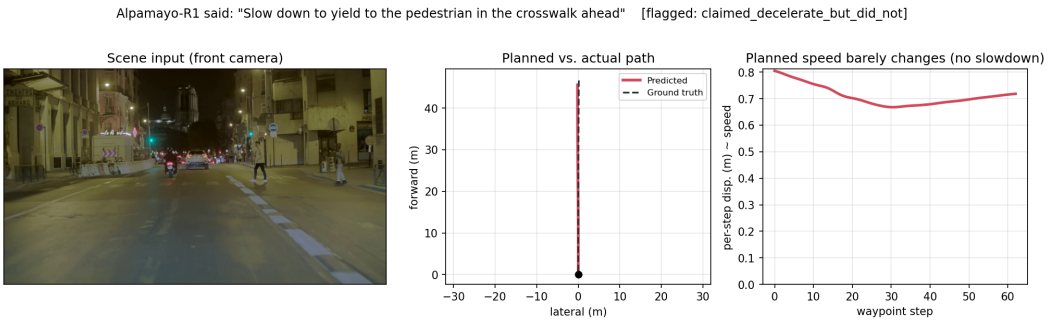

Figure 5: Alpamayo-R1 states "slow down to yield to the pedestrian in the crosswalk ahead," yet the planned speed does not decrease. Flagged `claimed_decelerate_but_did_not`.



Figure 6: Alpamayo-R1 states "slow down to yield to the cut-in vehicle from the left," yet the planned speed stays roughly constant. Flagged `claimed_decelerate_but_did_not`.

C Scorer Implementation Details

Kinematic action predicates $A(\tau)$. The discrete action of a trajectory is computed from the ego-frame waypoints (forward x, lateral y, with left positive). Let $\Delta y = y_T - y_0$ be the net lateral displacement and Δv the net speed change over the horizon. The predicates are: *stopped* if the final speed is below 0.5 m/s; *decelerates* if $\Delta v < -1.0$ m/s and *accelerates* if $\Delta v > +1.0$ m/s; *turns left* if $\Delta y > +1.0$ m and *turns right* if $\Delta y < -1.0$ m; *straight* if $|\Delta y| < 1.0$ m; and *nudge* if $|\Delta y| > 0.5$ m. The displacement predicates are independent of the sampling rate; the speed predicates assume the recorded timestep (0.5 s on NAVSIM, 0.1 s otherwise) so that the m/s thresholds are meaningful. We fixed thresholds on PhysicalAI-AV clips and reused them unchanged on NAVSIM.

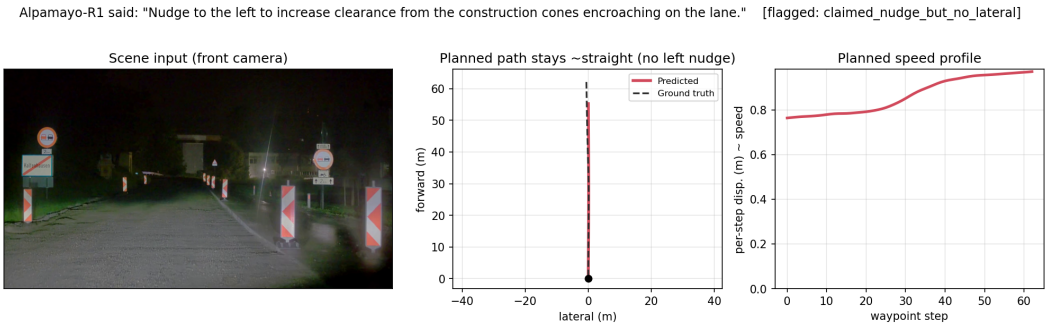

Figure 7: Alpamayo-R1 states "nudge to the left to increase clearance from the construction cones," yet the planned path stays straight with no lateral movement. Flagged `claimed_nudge_but_no_lateral`.

Reasoning action $A(r)$ and skips. We parse the action claimed by the reasoning by matching a fixed phrase vocabulary: stop ("stop", "halt", "standstill", "deceleration to zero"), decelerate ("decelerate", "slow", "brake"), accelerate ("accelerate", "speed up"), turn or change lane left and right ("turn left", "bear left", "change lane to the left", and the right-hand equivalents), keep straight ("continue straight", "maintain lane"), maintain speed ("maintain constant speed"), and nudge ("nudge"). A scene is *not checkable* and is skipped when the reasoning states no action that maps to a kinematic predicate; models that emit no trajectory are also skipped. Skipped scenes are reported separately from the non-skip mean rather than scored as neutral.

Entity vocabulary $E(\cdot)$. Entity-fidelity is computed over a closed category set: vehicle, pedestrian, cyclist, lead vehicle, cross traffic, parked vehicle, traffic light, stop sign, work zone, and oncoming vehicle. Each category is matched by a synonym list (for example, pedestrian from "pedestrian", "person", "walker"; work zone from "work zone", "construction", "cone"). Specific vehicle subtypes (lead, cross-traffic, parked) suppress the generic "vehicle" category for the same mention to avoid double counting.

D Model Input



Figure 8: Model input for one scene: multi-camera video across surround cameras and time steps. Alpamayo consumes all views and time steps jointly to produce a single trajectory and chain of causation.

E JUDGE VALIDATION AND PER-SCENARIO BREAKDOWN

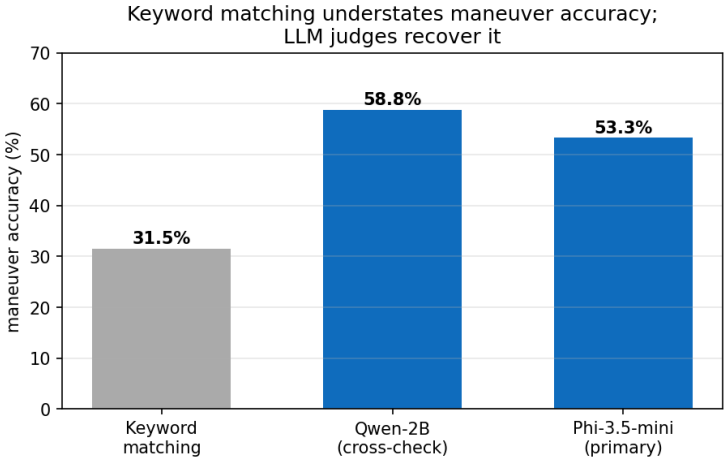

Figure 9: Maneuver accuracy under three judges on the same data. Keyword matching understates accuracy; two independent language-model judges recover it, and their agreement indicates that the headline result is not an artifact of one judge.

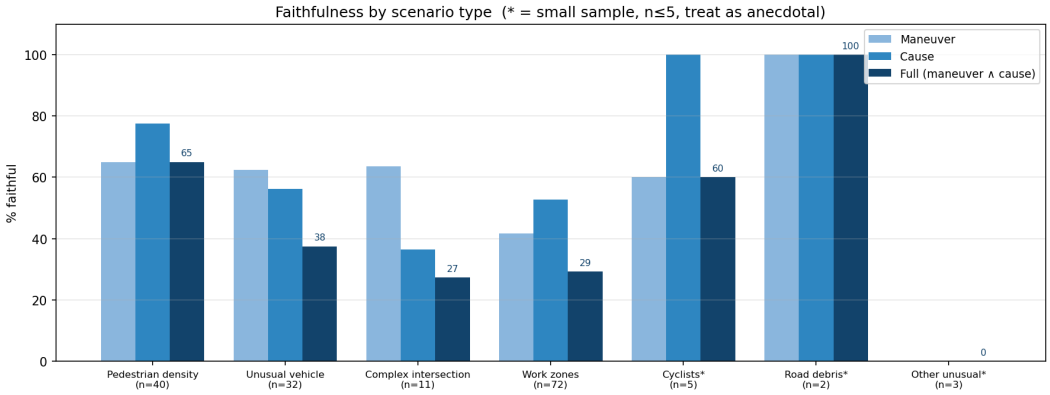

Figure 10: Faithfulness by scenario cluster (maneuver, cause, and full). Work zones, the largest cluster, are the weakest at 29% fully faithful. Clusters marked with an asterisk have five or fewer examples and are anecdotal.

F SCARCITY OF COUPLED CHAINS OF CAUSATION

Extending the cross-model comparison beyond the Alpamayo family requires another VLA that produces reasoning and trajectory as a single coupled output. Reviewing released driving VLAs, and running the accessible ones end-to-end, shows that this capability is uncommon (Table 5). Impromptu-VLA (Chi et al., 2025) outputs waypoints alongside a decoupled question-answering task; on its released nuScenes planning task, none of the 3,000 inspected training samples couple reasoning with trajectory generation. OpenDriveVLA (Zhou et al., 2026) suppresses chain-of-thought reasoning during inference by design. FSDrive (Zeng et al., 2025) produces visual rather than language-based reasoning and therefore does not satisfy the coupled language chain-of-causation requirement. AutoVLA (Zhou et al., 2025) is the one released model we found that emits coupled language reasoning and trajectory, so we extend the evaluation to its NAVSIM checkpoint. The result sharpens the scarcity finding rather than resolving it (Section G).

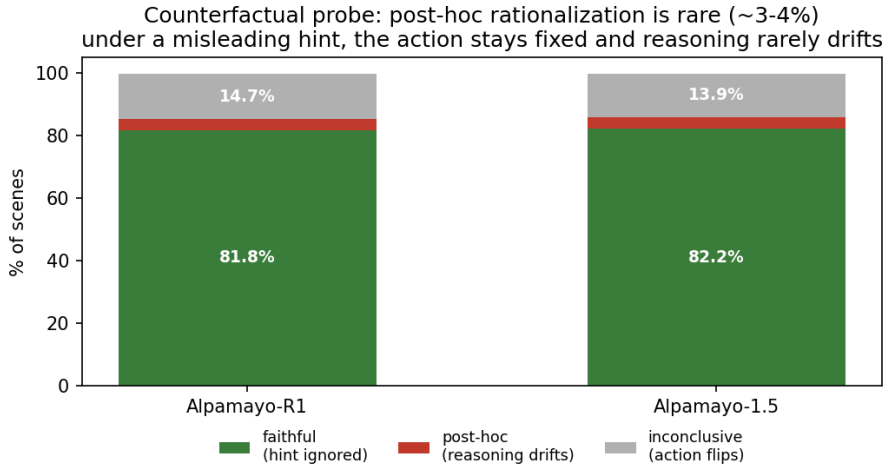

Figure 11: Counterfactual probe outcomes under a misleading opposite-action hint. On about 85% of scenes, the action stays fixed; overt post hoc rationalization is rare and indistinguishable between the two models.

Table 5: Released driving VLAs assessed against the coupled chain-of-causation requirement, defined as language-based reasoning and trajectory generation within a single forward pass. We assessed models by reviewing released resources and, where feasible, running direct inference.

Model	Dataset	Coupled language CoC?
Alpamayo-R1 / 1.5	PhysicalAI-AV	Yes (anchor)
AutoVLA	NAVSIM (released)	Yes; run end-to-end (0% CoC by default, prefill $\phi_{\mathrm{TC}} = 54.3\%$)
Impromptu-VLA	nuScenes	No; trajectory-only, reasoning is separate QA
OpenDriveVLA	nuScenes	No; suppresses CoT at inference
FSDrive	nuScenes / NAVSIM	No; visual CoT (generated frames)
OpenEMMA / LightEMMA	nuScenes	No; prompted VLM caption, not action-grounded
ORION / SimLingo	CARLA / Bench2Drive	Reasoning present; closed-loop simulation

G AutoVLA Diagnostics on NAVSIM

Running AutoVLA on the NAVSIM navtest split surfaces two findings. First, the released checkpoint emits *no* substantive chain of causation by default: across 831 decoded scenes, 0 produce a maneuver-and-cause rationale (0.0%, 95% Wilson interval $[0.0, 0.5]\%$); on straightforward scenes, the model uses an escape hatch written into its own prompt and returns a fixed filler string. Only assistant prefill elicits a chain of causation, so we label every AutoVLA consistency number as prefill-elicited. Second, when so elicited, its stated maneuver matches its own trajectory on 54.3% of checkable scenes (95% Wilson interval $[53.1, 55.4]\%$; 6,902 checkable scenes, coverage 57%), below both Alpamayo-R1 (58.2%) and Alpamayo-1.5 (61.7%) on the trajectory-consistency scorer ϕ_{TC}, whose predicate thresholds are dataset-agnostic and therefore comparable across substrates (Figure 4). NAVSIM has no human-verified reasoning labels equivalent to the PhysicalAI-AV expert gold, so ϕ_{RG} and ϕ_{EF} cannot be computed, and we compare AutoVLA on ϕ_{TC} only. AutoVLA thus extends the self-consistency tier to a third model on a second dataset. Its lower ϕ_{TC} fits the paper's central pattern, but with only ϕ_{TC} available, it cannot test the grounded-versus-plausibility dissociation itself. The comparison also spans different datasets and decoding regimes (native decoding for the Alpamayo models, prefill-elicited decoding for AutoVLA), and conditioning an autoregressive

decoder on a forced prefix may itself affect self-consistency. We therefore read it as indicative, not as a controlled ranking.

Figure 12 shows the scarcity result directly: the released checkpoint emits a substantive chain of causation on essentially none of the default (adaptive-CoT) scenes. It produces one only when the assistant prefill forces it. Figure 13 breaks down the reasoning-action contradictions among the forced-CoT scenes. Most are longitudinal: stated decelerations and stops that the trajectory does not execute, consistent with AutoVLA's lower trajectory consistency in Figure 4.

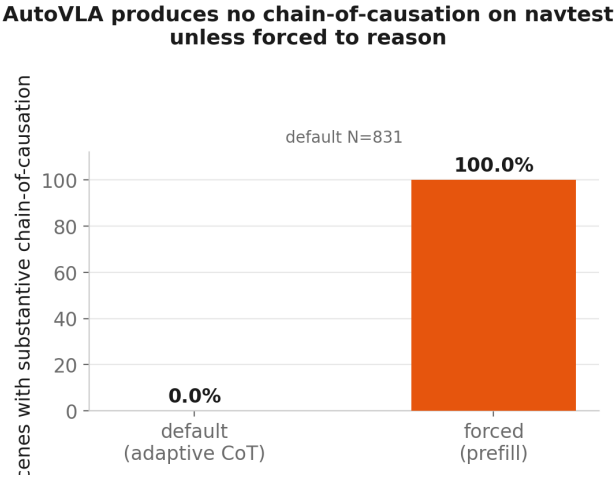

Figure 12: AutoVLA on NAVSIM navtest produces no substantive chain of causation by default (adaptive CoT) and only does so when forced by assistant prefill. On straightforward scenes, the released checkpoint uses an escape hatch written into its own prompt. This is the scarcity finding of Section 4.6.

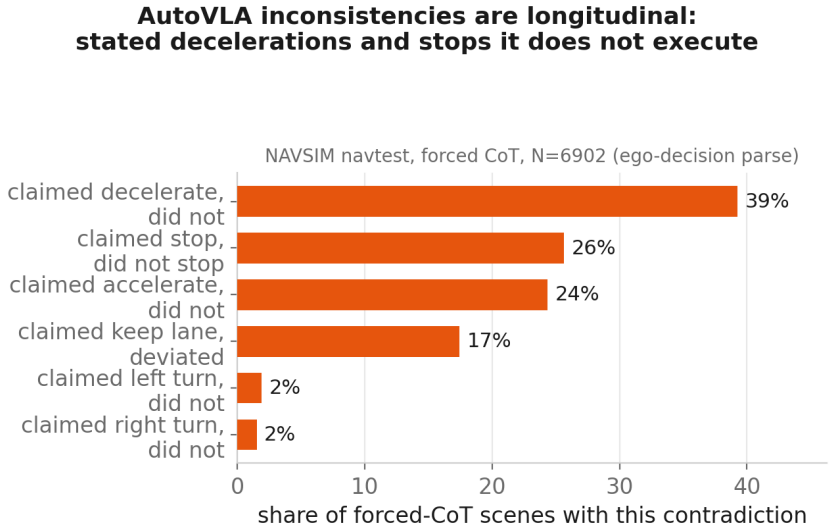

Figure 13: Reasoning-action contradictions among AutoVLA's forced-CoT scenes on NAVSIM navtest (ego-decision parse). Longitudinal cases dominate the share of scenes flagged with each "claimed X but did not" contradiction: stated decelerations and stops that the executed trajectory does not carry out.