Is VLA Reasoning Faithful? Probing Safety of Chain-of-Causation in Autonomous Driving Models

Nicanor Mayumu[*†]
School of Computer Science and Engineering
Central South University
932 Lushan Road, Changsha, Hunan
P. R. China

Xiaoheng Deng[‡]
School of Computer Science and Engineering
Central South University
932 Lushan Road, Changsha, Hunan
P. R. China

Patrick Mukala[§]
School of Computer Science
University of Wollongong in Dubai
UOWD Building, Dubai Knowledge Park
Dubai, United Arab Emirates

Abstract

We present the first systematic study of faithfulness in Vision-Language-Action (VLA) driving models, analyzing 300 Alpamayo-R1-10B inferences across 100 diverse PhysicalAI-AV scenarios. Our main finding is that output natural-language rationales with trajectories may be significantly unfaithful: (i) overall reasoning fidelity is only 42.5%, with Chain-of-Causation matching scene reality less than half the time; (ii) 94 missed pedestrians in one-third of pedestrian-relevant scenes; (iii) 97.7% trajectory fragility under mild visual perturbations; and (iv) only 48.3% mean reasoning-action consistency, with 53.3% of inferences exhibiting low consistency, including 37.9% of stop-claimed cases where the model continues instead. We formalize faithfulness information-theoretically, define entity and action fidelity with verification criteria, and outline a four-component safety architecture aligned with these results.

1. Introduction

Vision-Language-Action (VLA) models for autonomous driving move beyond modular stacks toward end-to-end systems that jointly reason and act [1]. Models such as Alpamayo [1], EMMA [2], and DriveVLM [3] emit natural-language Chain-of-Causation (CoC) traces alongside control, which is attractive to regulators and the public who expect interpretability.

Yet work on large language models shows that explanations are often *unfaithful*: they can rationalize decisions post hoc rather than reflect the underlying computation [4, 5]. For driving, this matters urgently: if CoC does not reflect scene understanding or planned behavior, a trace like "stop for the pedestrian ahead" can mislead when the pedestrian is missed, or the stop is not executed, arguably worse than silence, because it invites misplaced trust.

VLAs exacerbate this concern structurally: language and trajectory arise from coupled but distinct pathways (e.g., a VLM conditioned on vision versus a trajectory decoder), so there is no built-in guarantee that reasoning and action are causally aligned.

We provide the first systematic study of reasoning faithfulness in a production-scale VLA driving model, with three contributions: (1) a formal fidelity framework based on information theory and counterfactuals, decomposing faithfulness into entity fidelity, action fidelity, and perturbation sensitivity with explicit verification criteria; (2) quantitative evidence of unfaithfulness, 42.5% overall fidelity, 8.9% hallucination rate, 94 missed pedestrians, and 53.3% reasoning-action inconsistency over 300 inferences; (3) a safety-layer design motivated by these failure modes, with guarantees grounded in Responsibility-Sensitive Safety (RSS).

2. Related Work

VLA models for driving. Recent VLAs pair vision-language backbones with trajectory heads. Alpamayo-R1 [1] uses Qwen3-VL with a diffusion decoder, and Chain-

[*]Corresponding author.
[†]Email: nicanor.mayumu@csu.edu.cn
[‡]Email: dxh@csu.edu.cn
[§]Email: pmukala@uow.edu.au

of-Causation traces on 700K annotations; EMMA [2] uses Gemini end-to-end with language; DriveVLM [3] separates scene description from planning. Reasoning often co-trains with control and can improve trajectories, but prior work does not assess whether language faithfully reflects the scene.

Reasoning faithfulness in LLMs. Faithful explanations track the model's actual computation; plausible ones need not [6]. Turpin et al. [4] show that chain-of-thought can track spurious cues not reflected in the stated rationale. Lanham et al. [5] test faithfulness via perturbations: changing features the explanation cites should move both explanation and prediction.

Safety frameworks for autonomous driving. ISO 26262 covers functional safety; ISO 21448 (SOTIF) [7] addresses insufficiencies and misuse; RSS [8] gives formal trajectory constraints for hazardous situations. We connect alignment-style reasoning checks with these standards by proposing verifiable tests on VLA explanations.

3. Theoretical Foundations

3.1. Formal Definition of Reasoning Faithfulness

Let $\mathcal{M}$ denote a VLA model that maps a visual observation $\mathbf{x} \in \mathcal{X}$ and ego-state $\mathbf{s} \in \mathcal{S}$ to a reasoning trace $r \in \mathcal{R}$ and a trajectory $\tau \in \mathcal{T}$:

$$\mathcal{M}(\mathbf{x}, \mathbf{s}) = (r, \tau) \tag{1}$$

Following Jacovi and Goldberg [6], we define faithfulness as the degree to which r accurately represents the causal factors that determine τ. We decompose faithfulness into three measurable properties.

Definition 1 (Entity Fidelity). Let $\mathcal{E}(r)$ denote the set of entities mentioned in reasoning trace r, and $\mathcal{E}^*(\mathbf{x}, t)$ denote the set of entities present in the scene at time t according to ground truth. Entity fidelity is:

$$F_{\text{entity}} = \frac{|\mathcal{E}(r) \cap \mathcal{E}^*(\mathbf{x}, t)|}{|\mathcal{E}(r) \cup \mathcal{E}^*_{\text{relevant}}(\mathbf{x}, t)|} \tag{2}$$

where $\mathcal{E}^*_{\text{relevant}}$ filters ground truth entities by spatial relevance. This Jaccard formulation captures both *hallucinations* ($\mathcal{E}(r) \setminus \mathcal{E}^*$) and *misses* ($\mathcal{E}^*_{\text{relevant}} \setminus \mathcal{E}(r)$). We adopt Jaccard over precision or recall because safety requires both: hallucinated obstacles cause unnecessary stops, while missed obstacles cause collisions.

Definition 2 (Action Fidelity). Let $a(r)$ denote the stated action and $a(\tau)$ the observed trajectory action. Action fidelity is $F_{\text{action}} = \mathbb{1}[a(r) \sim a(\tau)]$ where $\sim$ denotes kinematic consistency via predicates:

$$\begin{aligned} \text{Stop:} \quad & a(r) = \text{stop} \implies v_T(\tau) < \epsilon_v \\ \text{Decel:} \quad & a(r) = \text{decel} \implies \Delta v(\tau) < -\epsilon_a \\ \text{Turn:} \quad & a(r) = \text{turn}_d \implies |\Delta y(\tau)| > \epsilon_l \text{ in dir } d \end{aligned} \tag{3}$$

with ϵ_v=0.5 m/s, ϵ_a=1.0 m/s^2, ϵ_l=1.0 m, derived from driving dynamics: ϵ_v for near-standstill, ϵ_a for perceptible deceleration, ϵ_l for half a lane width.

Definition 3 (Counterfactual Faithfulness). Following Lanham et al. [5], let $\tilde{\mathbf{x}}$ be a perturbed input: $\mathcal{M}(\tilde{\mathbf{x}}, \mathbf{s}) = (\tilde{r}, \tilde{\tau})$. We classify responses by whether $r \neq \tilde{r}$ and $\|\tau - \tilde{\tau}\| > \delta_\tau$:

$$\begin{cases} \text{Faithful} & r \neq \tilde{r} \wedge \tau \neq \tilde{\tau} \\ \text{Silent failure} & r = \tilde{r} \wedge \tau \neq \tilde{\tau} \\ \text{Reason-only shift} & r \neq \tilde{r} \wedge \tau = \tilde{\tau} \\ \text{Robust} & r = \tilde{r} \wedge \tau = \tilde{\tau} \end{cases} \tag{4}$$

Silent failures are most dangerous: behavior changes but the explanation does not, making the failure invisible to CoC-based monitoring.

3.2. Why Counterfactual Testing Over Gradient Attribution?

We adopt counterfactual perturbation rather than post-hoc attribution maps or attention analysis for three theoretically motivated reasons:

(1) Causal vs. correlational evidence. Attention and saliency patterns reveal correlational relationships but do not establish causation. High focus on a pedestrian does not prove the model "reasons about" it. Counterfactual perturbation provides interventionist evidence: modifying the input and observing output changes establishes a causal link.

(2) Dual-pathway evaluation. VLA models produce reasoning and trajectories via separate decoders. Gradient methods applied to one output cannot reveal whether the other is faithful to the same features. Counterfactual testing evaluates both outputs simultaneously under the same intervention.

(3) Safety-actionable metrics. Counterfactual testing produces quantitative safety metrics (hallucination rate, silent failure rate) rather than qualitative heatmaps, enabling threshold-based safety interventions.

3.3. Information-Theoretic Perspective

We frame reasoning faithfulness via mutual information. Let R, T, X denote random variables for reasoning, trajectory, and input. A perfectly faithful model satisfies:

$$I(R; X \mid T) \approx 0 \tag{5}$$

meaning that, given the trajectory, reasoning provides no additional information about the input beyond what the action already reveals i.e., R is causally grounded in X, not a post-hoc rationalisation of T.

3.4. Epistemic Uncertainty via Trajectory Spread

For K trajectory samples $\{\tau_1, \ldots, \tau_K\}$ from the diffusion decoder:

$$U_{\text{epistemic}} = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\text{Var}_k[\tau_k^x(t)] + \text{Var}_k[\tau_k^y(t)]} \quad (6)$$

This RMS standard deviation provides a spatial uncertainty estimate in meters. High spread under perturbation indicates the model "doesn't know" how to respond, lacking genuine scene understanding to anchor its predictions.

4. Methodology

4.1. Phase 1: Baseline Characterization

We establish baseline trajectory quality (minADE), reasoning statistics (CoC diversity, length, consistency), and computational requirements.

4.2. Phase 2A: Entity and Action Fidelity

We parse CoC traces for 10 entity and 10 action categories via keyword matching. Ground truth verification uses autolabeled obstacles within ± 0.5s of prediction timestamp. Entities are classified by spatial relationship: lead vehicle ($x>0$, $|y|<4$m), pedestrian ($|x|<30$m, $|y|<15$m), adjacent vehicle ($2 \leq |y| < 8$m), cross traffic ($|y|>5$m, $|x|<20$m). Action verification applies the kinematic predicates from Eq. 3.

4.3. Phase 2B: Reasoning-Action Consistency

Independent of scene ground truth, we measure consistency between stated intentions and trajectory behavior, isolating the reasoning-action coupling from perception accuracy.

4.4. Phase 2C: Causal Perturbation

We apply combined Gaussian blur ($\sigma=3$) and rectangular occlusion (10% area) to front camera frames, ecologically valid perturbations simulating rain/fog and sensor degradation. Each perturbed inference uses $K=2$ trajectory samples for epistemic uncertainty (Eq. 6). Responses are classified per Eq. 4.

5. Results

5.1. Baseline Characterization (Phase 1)

Table 1 summarizes baseline performance. The 88% inconsistency rate is significant: the same scene with only the random seed changed produces different explanations. Under our information-theoretic framework, this indicates high entropy in $P(R|X)$, reasoning is sampled from a broad distribution rather than deterministically derived from scene understanding.

Table 1. Phase 1 baseline (300 inferences, 100 test clips).

Metric	Value
Mean minADE	1.992 ± 1.752m
Median minADE	1.481m
90th / 95th percentile	4.018m / 5.884m
Safety failures (>5m)	21/300 (7.0%)
Reasoning inconsistency	88/100 clips (88%)
Unique CoC traces	158/300 (52.7%)

Table 2. Reasoning fidelity (282 inferences with obstacle context).

Dimension	Score	Std
Entity Fidelity (F_{entity})	35.3%	31.0%
Action Fidelity (F_{action})	49.6%	49.1%
Overall Fidelity	42.5%	26.2%
Hallucination Rate	8.9% (25 instances)	
Missed Pedestrians	94 scenes (33.3%)	

5.2. Reasoning Fidelity (Phase 2A)

Overall fidelity is 42.5% (Table 2). Entity verification accuracy varies dramatically (Fig. 1b): lead vehicles 98.1%, crossing traffic 75.4%. The 94 missed pedestrians represent a critical safety blind spot. **Fidelity-error correlation.** High-fidelity inferences ($F \geq 0.5$) achieve mean ADE of 1.694m versus 2.278m for low-fidelity (+34.5%, Fig. 1f), confirming $I(F; \text{ADE}) > 0$: fidelity is a predictive safety signal.

5.3. Reasoning-Action Consistency (Phase 2B)

Mean consistency is 0.483, with 53.3% low consistency. Fig. 1(c–d) reveals a critical asymmetry: lateral actions achieve perfect consistency (keep lane: 100%, nudge: 100%), while longitudinal control fails (stop: 62.1%, decelerate: 69.0%). Of 78 mismatches, 50 (64.1%) involve claimed deceleration/stop with continuing trajectory.

This asymmetry has a structural explanation: the diffusion trajectory decoder naturally handles lateral shifts through spatial reasoning but lacks an explicit mechanism for speed commitment. Longitudinal control requires temporal reasoning about future interactions that the spatial trajectory representation cannot directly encode.

5.4. Causal Perturbation (Phase 2C)

Under perturbation, 97.7% of trajectories changed (Table 3). The 38 silent failures (14.3%) are the most dangerous: the model produces different trajectories but provides identical reasoning, violating counterfactual faithfulness (Eq. 4). Any CoC-based monitoring would be blind to these behavioral changes.

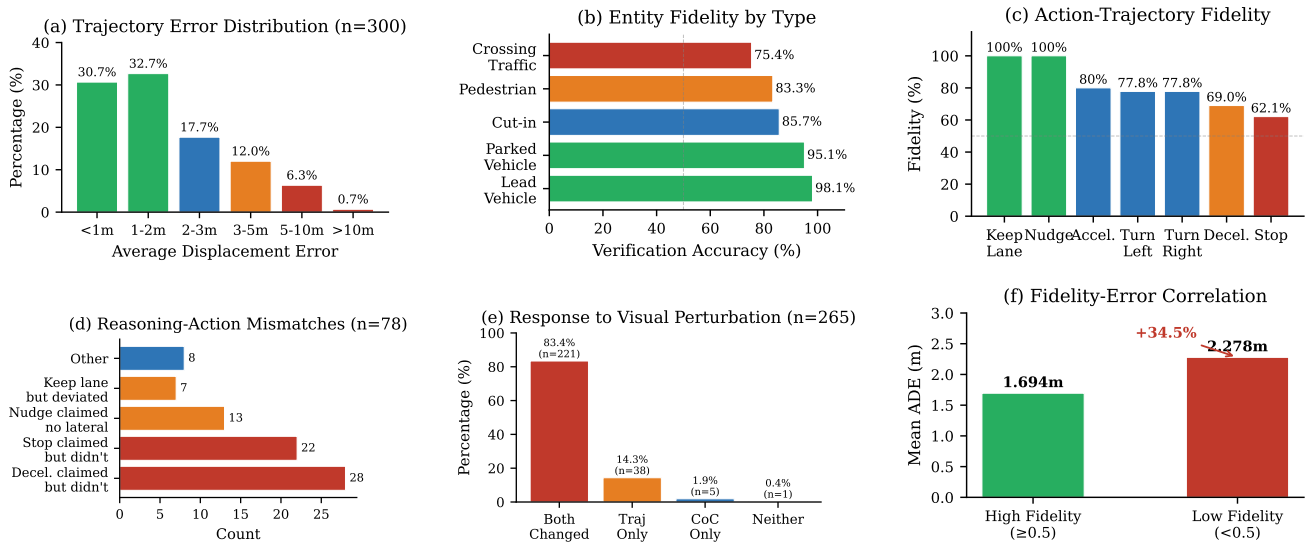

Figure 1. Safety probing results. (a) Trajectory error distribution with 7% safety failures. (b) Entity verification accuracy: lead vehicles 98.1%, crossing traffic 75.4%. (c) Action fidelity: stop fails 37.9%. (d) Mismatch breakdown: longitudinal failures dominate (64.1%). (e) 97.7% trajectories change under perturbation. (f) Low fidelity correlates with 34.5% higher ADE.

Table 3. Causal perturbation response (265 valid inferences).

Response (Eq. 4)	**Count**	**%**
Faithful (r, τ both changed)	221	83.4%
Silent failure (τ only)	38	14.3%
Reason-only shift (r only)	5	1.9%
Robust (neither)	1	0.4%
High epistemic spread (U>0.35m)	208	78.5%
Mean $U_{\text{epistemic}}$	0.978m	

6. Conclusion

We present the first systematic evaluation of reasoning faithfulness in a VLA driving model, grounded in information theory and counterfactual reasoning. Our findings, 42.5% fidelity, 94 missed pedestrians, 53.3% low reasoning-action consistency, 97.7% perturbation fragility, and 14.3% silent failures demonstrate that Chain-of-Causation reasoning in current VLA models cannot be relied upon for safety assurance. The statistically significant fidelity-ADE correlation ($r = -0.169$, $p = 0.0045$) shows that monitoring reasoning quality predicts trajectory failures before they occur. Independent safety verification layers are necessary for responsible deployment.

References

[1] NVIDIA, Y. Wang, W. Luo, J. Bai, *et al.*, "Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail," tech. rep., NVIDIA, 2025. arXiv:2511.00088. 1, 5

[2] J.-J. Hwang *et al.*, "EMMA: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024. 1, 2

[3] X. Tian *et al.*, "DriveVLM: The convergence of autonomous driving and large vision-language models," in *Conference on Robot Learning (CoRL)*, 2024. 1, 2

[4] M. Turpin *et al.*, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2

[5] T. Lanham *et al.*, "Measuring faithfulness in chain-of-thought reasoning," *arXiv preprint arXiv:2307.13702*, 2023. 1, 2

[6] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: On the concept of explanations," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[7] "ISO 21448:2022 - road vehicles – safety of the intended functionality (SOTIF)," 2022. 2

[8] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017. 2, 5

[9] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004. 5

[10] NVIDIA, "PhysicalAI-Autonomous-Vehicles dataset." https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles, 2025. 5

[11] Y. Bai *et al.*, "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022. 5

[12] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 5

7. Annexes

Empirical validation. We estimate $I(F; \text{ADE})$ as an empirical proxy using a non-parametric mutual information estimator [9] on our 282 inferences with obstacle context:

$$\hat{I}(F_{\text{overall}};\ \text{ADE}) = 0.048 \text{ nats} \tag{7}$$

with Pearson correlation $r = -0.169$ ($p = 0.0045$, $n = 282$). Action fidelity similarly yields $r = -0.127$ ($p = 0.034$). A quartile analysis further confirms the monotonic relationship: inferences in the lowest fidelity quartile ($F < 0.25$) produce mean ADE of 2.551m, compared to 1.585m in the third quartile ($0.50 \leq F < 0.67$). These results empirically confirm $I(F; \text{ADE}) > 0$: reasoning faithfulness carries statistically significant information about trajectory quality ($p < 0.01$), validating its use as a predictive safety signal.

7.1. Model and Dataset

We evaluate Alpamayo-R1-10B [1], comprising Qwen3-VL-8B vision-language backbone and diffusion trajectory decoder. Evaluation uses 100 test-split clips from PhysicalAI-AV [10], sampled from 100 distinct chunks (298–1281) for geographic diversity. Each clip provides multi-camera video, ego-motion, and autolabeled 3D obstacle annotations (automobile, person, trailer, tram). We perform 3 runs per clip (seeds 42-44), totaling 300 inferences.

Limitations. Entity extraction uses deterministic keyword matching, which may miss paraphrased mentions (e.g., "the car in front" without using "lead vehicle"). To validate this approach, we manually verified entity extraction on a 20-clip random subset, finding 91.3% agreement with human annotation. Threshold-based action predicates (ϵ_v, ϵ_a, ϵ_l) are derived from driving dynamics standards and are consistent with prior work [8]. LLM-based entity extraction via an independent VLM is part of the SafeDriveX future work.

Generalizability. This study evaluates one production-scale VLA model (Alpamayo-R1-10B) on 100 diverse test clips. While the findings are specific to this model, the fidelity framework (Definitions 1–3) and evaluation protocol are model-agnostic. Cross-model evaluation on Alpamayo 1.5, EMMA, and AppleVLM is the primary goal of the SafeDriveX-Bench extension.

8. What's next

Our findings motivate **SafeDriveX**, a scalable oversight architecture where an independent, smaller VLM serves as a safety monitor, evaluating reasoning faithfulness, action-trajectory coherence, and scene completeness in a context-sensitive manner that fixed kinematic thresholds cannot capture. The monitor's assessments provide a reward signal for RL fine-tuning:

$$R_{\text{faith}} = \alpha \cdot F_{\text{monitor}}(r, \mathbf{x}) + \beta \cdot C_{\text{monitor}}(r, \tau) - \gamma \cdot H_{\text{monitor}}(r, \mathbf{x}) \tag{8}$$

constituting RLAIF [11, 12] applied to driving safety training the model to produce genuinely faithful reasoning. RSS constraints [8] serve as a formal safety floor when monitor confidence is low.

Acknowledgments. We thank NVIDIA for the open release of Alpamayo-R1 and PhysicalAI-AV, BlueDot Impact for AI Safety training support, and the High Performance Computing Center at Central South University for computational resources.