ACTING FOR THE RIGHT REASONS: DECISION-CAUSE ALIGNMENT IN TRAJECTORY PLANNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) are increasingly adopted as end-to-end trajectory planners trained by behavior cloning, which aligns actions but not necessarily the reasons behind them and can therefore mask safety-critical failures. Existing methods infer a planner's reasons indirectly from representations, interventions, or behavioral associations. However, such evidence is not the decision-driving event itself; if the inferred reason changes with the diagnostic, it reflects the diagnostic–planner pair rather than a property of the planner. We instead define the *decision-driving event*, the world event that actually drives the planner's response, and call a planner **decision-cause aligned** when this event is the human-relevant event rather than a co-occurring cue. When the human-relevant event and the cue share an onset and response form, static behavior cannot separate their contributions. When the cue appears first, ordinary driving logs can identify its contribution under explicit response-latency and disturbance-isolation assumptions. Building on this result, we introduce **TREAT** (loca**T**e–cu**R**e–c**E**rtify for decision-c**A**use alignmen**T**): LOCATE measures temporal locking to candidate cues, CURE removes diagnosed dependence with audit-targeted decorrelation records, and CERTIFY calibrates the audit by defect injection. Across nuScenes and Argoverse 2, four VLM families, the non-VLM planner UniAD, and 100+ model instances, replica fleets show cue-ward dependence that is graded and population-structured. Although the confirmed case occurs in the most cue-ward fleet observed, it remains rare at the instance level; certificates therefore bind individual trained weights. **TREAT** isolates this naturally occurring defect where four standard diagnostics do not, repairs it with less task degradation than compute-matched continued training (ADE 2.72 vs. 3.22 m; 1.62 m before repair), and calibrates audit resolution with per-planner false-positive rates and detection limits at 80% power. Code: `https://anonymous.4open.science/r/TREAT-5C14/README.md`

1 INTRODUCTION

Vision–language models (VLMs) are increasingly adopted for end-to-end trajectory planning (Hu et al., 2023; Wang et al., 2024; Steiner et al., 2024; Liu et al., 2024; Amini et al., 2025), typically trained by behavior cloning on human driving demonstrations. Behavior cloning, however, aligns actions, but not necessarily the reasons behind them, and this gap can mask safety-critical failures. In Figure 1, two planners brake identically at a crosswalk, yet one responds to the pedestrian and the other to the crosswalk markings; when the markings are absent, the latter fails to brake, exposing a failure mode reminiscent of the geometry of the 2018 Tempe fatality (National Transportation Safety Board, 2019). This concern, commonly framed as *acting for the right reasons*, becomes a concrete question in trajectory planning: which world event actually drives the planner's response?

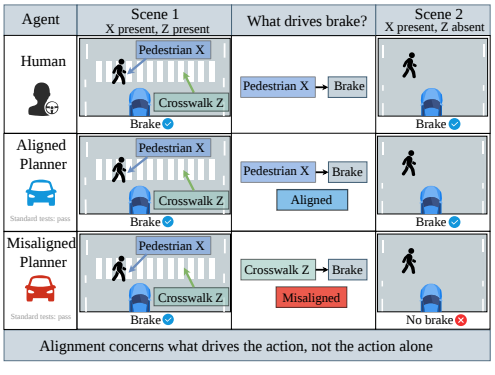


Figure 1: **Same action, different decision-driving event.** The same brake action can be driven by the pedestrian or the crosswalk cue.

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.

Existing methods fall into three paradigms: **representation-level** methods infer reasons from what the model encodes or highlights (Jacovi & Goldberg, 2020; Ross et al., 2017; Lapuschkin et al., 2019); **intervention-based** methods infer reasons from how behavior changes under edited or counterfactual inputs (McCoy et al., 2019; Geirhos et al., 2020; Dosovitskiy et al., 2017; de Haan et al., 2019); and **observational** methods infer them from behavioral associations across contexts, reaction timing, or training conditions (Swamy et al., 2022; Langosco di Langosco et al., 2022). All three provide evidence about a candidate reason, but such evidence does not, by itself, identify what actually drives the planner's response. The gap is twofold. First, a factor can be encoded, salient, or sensitive to edits without driving the action, as we confirm empirically (Table 1). Second, a reason that changes with the diagnostic is a property of the diagnostic–planner pair, not of the planner; a planner-level property must be defined on the planner and then estimated. This raises two questions. **(Q1) What counts as the reason that actually drives an action?** We answer with the *decision-driving event*, the world event that drives the planner's response: on the human side, a reason pairs a world event, the *human-relevant event*, with the response it demands, and a planner is **decision-cause aligned** when its decision-driving event is that event rather than a co-occurring cue. **(Q2) How can this alignment become a stable property of a trained planner?** We answer with **TREAT**, a LOCATE–CURE–CERTIFY framework for decision-cause alignment that fixes the target of assessment independently of the diagnostic.

The obstacle is identifiability. When the human-relevant event and a candidate cue share their onset and response form, aligned and misaligned planners can generate the same observations, so static trajectory scores (Hu et al., 2023) or behavioral correlations (Swamy et al., 2022; Langosco di Langosco et al., 2022) cannot distinguish them (Prop. 1). Ordinary driving logs contain the missing lever: naturalistic-driving studies have exploited event–response timing (Markkula et al., 2016; Wu & Lin, 2019; Engström et al., 2024; Varotto et al., 2026), and when a cue appears before the human-relevant event becomes active, a response in that cue-first interval cannot be driven by the later event (Granger, 1969), although other coincident events may still drive it; this possibility is addressed by the disturbance-isolation assumption and the cue scan. We formalize this with an onset-aligned response model under which, given a minimum response latency and disturbance isolation (Assumptions 1–2), temporal separation makes the cue contribution point-identifiable (Prop. 2), without constructed counterfactuals or access to model internals. **TREAT** builds on this identification in three stages: LOCATE measures temporal locking to candidate cues, CURE retrains on audit-targeted decorrelation records, and CERTIFY calibrates audit resolution by defect injection.

Across nuScenes (Caesar et al., 2020) and Argoverse 2 (Wilson et al., 2021), four VLM families (Amini et al., 2025; Wang et al., 2024; Steiner et al., 2024; Liu et al., 2024), the non-VLM planner UniAD (Hu et al., 2023), and $100+$ model instances, **TREAT** shows that cue-ward dependence is graded at the population level but thresholded at the instance: replica fleets of the four architectures differ in cue-ward tilt, and the single confirmed misalignment arises in the most cue-ward family yet only one of thirty runs of that pipeline crosses the confirmation threshold. On the confirmed instance, LOCATE detects braking locked to crosswalk onset rather than the agent entering the danger zone, while four standard diagnostics fail to isolate the defective planner; CURE removes the diagnosed dependence and triples the event response at a task cost (ADE $1.62 \rightarrow 2.72$ m) that is 0.5 m smaller than that of compute-matched continued training; and CERTIFY calibrates audit resolution with measured false-positive rates and detection limits.

Our contributions are summarized as follows:

- **Formulation.** Decision-cause alignment as a property of a trained planner defined by its decision-driving event, with conditions under which it is not identifiable from behavior (Prop. 1) and under which, given stated latency and disturbance-isolation assumptions, temporal precedence in ordinary logs identifies the cue contribution (Prop. 2).
- **Method. TREAT** locates a misaligned dependence by temporal locking, cures it with decorrelation records, and calibrates audit resolution with false-positive rates and detection limits.
- **Discovery.** Population-level susceptibility to cue-ward dependence is graded and population-structured, whereas confirmation is an instance-level, detectability-thresholded statement, so certificates must bind trained weights.
- **Evaluation.** A naturally misaligned planner that four standard diagnostics fail to isolate, repaired with less task degradation than compute-matched continued training, and with audit resolution calibrated by measured false-positive rate and detection limit.

2 PROBLEM SETUP

Consider a trained trajectory planner instance π (a *model instance*, one set of weights Θ from one training run) driving through a logged scene with frames $t = 1, \ldots, T$. At each frame it plans a future ego trajectory $\hat{\mathbf{y}}_t$, from which we derive a scalar response r_t, the mean planned speed over the 3 s horizon. Lower r_t corresponds to stronger braking; the audit converts this series to post-onset braking magnitudes in §4.1. Let $\mathcal{E}$ be the set of world events identified from the logged tracks and map layers, each with an onset τ_E, the first frame at which it becomes active. For a response r, the *human-relevant event* $X \in \mathcal{E}$ is the event that should drive it under the human driving specification; any other event $Z \in \mathcal{E} \setminus \{X\}$ is a *candidate cue*. In the running case, r is the braking response and X is an agent, a pedestrian or a vehicle, entering the ego vehicle's danger zone, per the duty to yield to road users in the path (United Nations, 1968) and its operationalization in AEB test protocols (Euro NCAP, 2024).

Definition 1 (Decision-driving events) *A planner's response may be driven jointly by the human-relevant event X and a candidate cue Z. Its* decision-driving events *are*

$$\mathcal{D}_\pi(r) = \{X : \beta_X \neq 0\} \cup \{Z : \beta_Z \neq 0\}, \tag{1}$$

where β_X and β_Z are their contributions in the onset-aligned model

$$r_t = \mu + \beta_X g_X(t - \tau_X) + \beta_Z g_Z(t - \tau_Z) + \varepsilon_t. \tag{2}$$

Here g_X and g_Z are response kernels, μ is the baseline, and ε_t absorbs noise and other events. Hence $\mathcal{D}_\pi(r)$ is a property of the planner, not of the audit.

Definition 2 (Decision-cause alignment) *Planner π is* decision-cause aligned *on response r when the human-relevant event X is its only decision-driving event,*

$$\mathcal{D}_\pi(r) = \{X\} \quad \Longleftrightarrow \quad \beta_X \neq 0 \ \textit{and} \ \beta_Z = 0 \ \ \forall Z \in \mathcal{E} \setminus \{X\}. \tag{3}$$

With finite data, exact zero cannot be established. We therefore certify operational alignment *relative to the audit: for every audited cue Z,* LOCATE *finds no detectable cue-ward dependence at or above the calibrated injected-defect strength $\delta_Z = \mathrm{LOD}_p(N_Z)$ (§4.3). This operational certificate is stated on the audit scale and is not a numerical upper bound on the structural coefficient β_Z. Thus, cue dependence β_Z is a graded property of the model instance, whereas an audit verdict is a thresholded statement at the audit's finite-sample resolution.*

Given ordinary driving logs and black-box access to a model instance π_Θ, we test each audited cue Z for detectable cue-ward dependence at the audit's calibrated resolution δ_Z. A cue with no detectable dependence is cleared at this resolution; a detected dependence is adjudicated and, if confirmed, repaired and re-audited.

3 IDENTIFIABILITY OF THE DECISION-DRIVING EVENT

3.1 IDENTIFICATION ASSUMPTIONS

Temporal separation identifies the cue's contribution under two conditions.

Assumption 1 (Event-response latency) *A response cannot precede the event that drives it (Granger, 1969), and drivers and planners react with a nonzero delay (Markkula et al., 2016; Wu & Lin, 2019). We assume a minimum latency $\ell_{\min} > 0$: $g_E(t - \tau_E) = 0$ for $t < \tau_E + \ell_{\min}$ and every event E.*

The second condition is stated on the cue-first window $W = [\,\tau_Z + \ell_{\min},\ \tau_X + \ell_{\min})$, the frames in which the cue can already drive the response while the human-relevant event cannot yet.

Assumption 2 (Disturbance isolation) *On W, the disturbance ε_t of Eq. (2), which carries every event other than X and Z, is conditionally mean-zero given the cue's kernel (Spirtes et al., 2000; Pearl, 2009):*

$$\mathbb{E}[\varepsilon_t \mid g_Z(t - \tau_Z),\ t \in W] = 0. \tag{4}$$

A co-onset event that would violate Eq. (4) is not left in ε_t: it is added to the cue vocabulary and audited as its own cue (§4.1; App. C.1).

3.2 IDENTIFICATION PROPOSITIONS

We establish three results: static non-identifiability, temporal identifiability, and finite-sample detectability.

Proposition 1 (Static non-identifiability) *If X and Z share the onset, $\tau_X = \tau_Z =: \tau$, and the kernel, $g_X = g_Z = g$, then Eq. (2) reduces to*

$$r_t = \mu + (\beta_X + \beta_Z)\, g(t - \tau) + \varepsilon_t, \tag{5}$$

so only the sum $\beta_X + \beta_Z$ is identifiable from the response (Rothenberg, 1971).

Proof sketch. The two contributions collapse onto the same regressor $g(t-\tau)$, so infinitely many pairs (β_X, β_Z) generate the same response; in particular, a misaligned planner (β_X, β_Z) with $\beta_Z \neq 0$ is observationally equivalent to the aligned planner $(\beta_X + \beta_Z,\, 0)$. □

Proposition 2 (Temporal identifiability) *Let the cue appear first, $\tau_Z < \tau_X$, and let Assumptions 1 and 2 hold on W. If the cue has response energy on the window, $\mathbb{E}[g_Z(t-\tau_Z)^2 \mid t \in W] > 0$, then β_Z is point-identified from the response on W (MacKinlay, 1997; Jordà, 2005); in particular, the window mean identifies it up to the kernel's average height,*

$$\mathbb{E}[r_t - \mu \mid t \in W] = \beta_Z\, \bar{g}_Z, \qquad \bar{g}_Z = \tfrac{1}{|W|} \textstyle\int_W g_Z(t - \tau_Z)\, dt, \tag{6}$$

which requires no knowledge of the kernel's shape and is the quantity on which the locking contrast of §4.1 is built.

The energy condition fails only when W is empty or the cue does not shape the response there, in which case the scene carries no identifying information for this cue. LOCATE uses the conservative subset of W ending at τ_X (§4.1).

Proof sketch. For every $t \in W$, $t < \tau_X + \ell_{\min}$, so Assumption 1 gives $g_X(t - \tau_X) = 0$ and Eq. (2) reduces on W to $r_t - \mu = \beta_Z g_Z(t - \tau_Z) + \varepsilon_t$. By Assumption 2, the residual is orthogonal to $g_Z(t - \tau_Z)$ and mean-zero on W, so projecting $r_t - \mu$ onto the cue kernel isolates the weight,

$$\beta_Z = \frac{\mathbb{E}[(r_t - \mu)\, g_Z(t - \tau_Z) \mid t \in W]}{\mathbb{E}[g_Z(t - \tau_Z)^2 \mid t \in W]}, \tag{7}$$

and taking the plain expectation on W instead of the projection gives Eq. (6). □

Proposition 3 (Finite-sample detectability) *Consider testing $H_0 : \beta_Z = 0$ against $\beta_Z > 0$ with the window-mean statistic Eq. (6) pooled over N cue-first scenes, at significance level α and detection power p. The smallest driving weight the test can detect, its* limit of detection *(Currie, 1968; Bloom, 1995), scales as*

$$\text{LOD} \;\approx\; \frac{(z_{1-\alpha} + z_p)\, \sigma_\varepsilon}{\sqrt{N}\, \bar{g}_Z}, \tag{8}$$

with $z_{1-\alpha}$ and z_p the standard-normal quantiles at levels $1-\alpha$ and p, and σ_ε the standard deviation of ε_t. Detectability therefore improves as $1/\sqrt{N}$; a larger onset gap $\tau_X - \tau_Z$ provides a longer informative window and can improve detectability when the added portion contains cue-response signal; if the cue appears in only a fraction $\rho < 1$ of the scenes in which X occurs (partial redundancy), detectable effects scale by ρ.

Proof sketch. The window mean of $r_t - \mu$ in scene i has expectation $\beta_Z \bar{g}_Z$ and standard deviation σ_ε, so the mean over N scenes has standard error $\sigma_\varepsilon/\sqrt{N}$; a one-sided level-$\alpha$ test rejects above $z_{1-\alpha}\sigma_\varepsilon/\sqrt{N}$ and reaches power p iff $\beta_Z \bar{g}_Z \geq (z_{1-\alpha} + z_p)\sigma_\varepsilon/\sqrt{N}$, which is Eq. (8). □

4 TREAT: MODEL-INSTANCE DECISION-CAUSE ALIGNMENT

TREAT turns the identification result of §3 into a three-stage procedure on one trained model instance π_Θ (Figure 2), illustrated throughout with the running case of an agent entering the ego vehicle's danger zone (X) and a crosswalk coming into view (Z): LOCATE measures temporal locking, CURE supplies the missing distinction, and CERTIFY calibrates the claim.

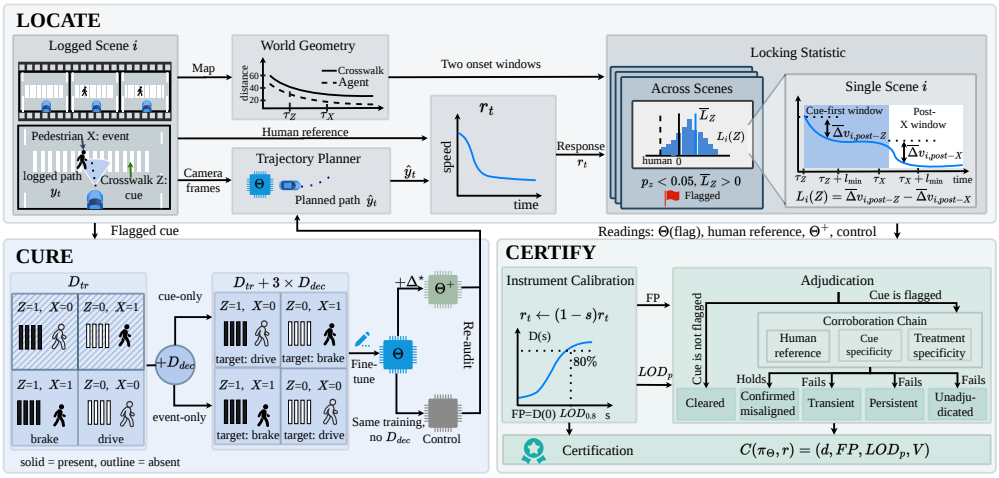


Figure 2: **The TREAT workflow on one model instance.** LOCATE flags cues by temporal locking on cue-first scenes; CURE fine-tunes on a four-cell decorrelation mixture whose off-diagonal configurations break the cue–event ambiguity, against a compute-matched control, both re-audited by LOCATE; CERTIFY calibrates the audit with injected defects and issues the verdict and certificate.

4.1 LOCATE: MEASURE TEMPORAL LOCKING

For each cue Z in the audited cue set V, we retain the cue-first scenes, $\tau_{Z,i} < \tau_{X,i}$, replay the planner, and read its response r_t on two windows: the cue-first window of Prop. 2, in which the crosswalk is visible but the agent has not yet entered the danger zone, and the window following the agent's onset. Each window is referenced to the planner's speed just before its own onset, giving the post-onset braking magnitudes $\overline{\Delta v}_{i,\text{post-}Z}$ and $\overline{\Delta v}_{i,\text{post-}X}$ and the scene's temporal-locking contrast

$$L_i(Z) = \overline{\Delta v}_{i,\text{post-}Z} - \overline{\Delta v}_{i,\text{post-}X}, \tag{9}$$

Because each event window is referenced to the speed immediately preceding its own onset, anticipatory slowing before the agent onset does not by itself count as cue locking; only a stronger cue-aligned response relative to the X-aligned response yields a positive locking contrast. Across the N_Z scenes, LOCATE reports the lock count n_Z, the number of scenes with $L_i(Z) > 0$, together with n_Z/N_Z, the mean contrast $\bar{L}_Z$, and the two-sided paired-t p-value p_Z of the contrast, treating scenes as the unit of replication; the same statistic on the logged human trajectories gives the *human reference*. A cue is *flagged* when $p_Z < \alpha$ with $\bar{L}_Z > 0$, with α=0.05 in all experiments, a screening rule whose realized error rates are calibrated by CERTIFY. The audit is black-box and intervention-free, consuming only replayed trajectories; a cue is audited only when $N_Z \geq 5$ cue-first scenes remain, and an instance is recorded as *inadmissible* only if no audited cue meets this requirement. Windows, baselines, the aggregate statistic, and the cue scan are specified in App. C.1.

4.2 CURE: SUPPLY THE MISSING DISTINCTION

Proposition 4 (Training-support insufficiency) *If Z and X always co-occur on the training support, planners with equal behavior-cloning likelihood can differ on cue-only inputs $(Z{=}1, X{=}0)$; a repair objective that sees the data only through this likelihood therefore cannot guarantee removal of a β_Z-type dependence (proof sketch in App. C.4).*

CURE supplies the missing distinction by adding the two off-diagonal configurations absent from the confounded training support, a crosswalk without an agent in the danger zone, which should not trigger braking, and an agent in the danger zone without a crosswalk, which should, together with matched on-diagonal records that preserve the cue–event contexts (App. C.2), and fine-tunes the weights Θ on these decorrelation records $\mathcal{D}_{\text{dec}}$ mixed with the task data $\mathcal{D}_{\text{tr}}$:

$$\Delta^\star = \arg\min_{\Delta} \; \mathcal{L}_{\text{main}}\big(\Theta + \Delta; \mathcal{D}_{\text{tr}}\big) + \lambda\, \mathcal{L}_{\text{main}}\big(\Theta + \Delta; \mathcal{D}_{\text{dec}}\big), \tag{10}$$

where $\mathcal{L}_{\text{main}}$ is the original training objective and λ is realized as the oversampling factor of the decorrelation records (3× in our runs); task preservation is not constrained during training but reported afterwards as held-out ADE. The repaired instance is π_{Θ^+} with $\Theta^+ = \Theta + \Delta^\star$. A *compute-matched control* receives the same additional training without the decorrelation records, so that the effect of the evidence is separated from that of continued training. Both the repaired instance π_{Θ^+}

and the control are returned to LOCATE for a full re-audit and evaluated on the main task: repair is measured by the same quantity that diagnosed the defect, not assumed.

4.3 CERTIFY: CALIBRATE THE CLAIM

Identifiability does not imply finite-sample detectability (Prop. 3). CERTIFY therefore calibrates LOCATE as an instrument (Currie, 1968): it injects cue-locked braking of known strength $s \in [0, 1]$ on cue-present frames of each cleared planner's own response series (a flagged instance inherits the calibration of its closest cleared peer), reruns LOCATE, and reads the detection curve

$$D(s) = \Pr[\text{flagged} \mid s], \qquad \text{FP} = D(0), \qquad \text{LOD}_p = \min\{s : D(s) \geq p\}. \tag{11}$$

These quantities map each LOCATE reading to a certificate $\mathcal{C}(\pi_\Theta, r) = (d, \text{FP}, \text{LOD}_p, V)$, where V is the audited cue set and d records the cue-wise verdicts. A non-flagged cue is *cleared* at its detection limit; a flagged cue is *confirmed misaligned* only if it survives the corroboration chain (human reference, cue specificity, and repair specificity; App. C.3) (Hill, 1965). Failed flags are reported as *transient* if the flag disappears under both arms without a CURE-specific reduction, and *persistent* if no arm reduces the cue-ward contrast below the reading; any other outcome is reported as flagged but unadjudicated. Under $\beta_Z = 0$, a confirmed verdict arises with probability at most FP because confirmation requires a flag; the corroboration chain is descriptive rather than null-calibrated and provides no identification beyond Prop. 2; App. D.1 reports how many nominal flags the whole study should expect under a global null (Westfall & Young, 1993). The certificate applies only to the weights, response, cue set, and scene population on which the audit is calibrated; a repaired instance Θ^+ is re-certified after re-audit (App. C.3).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

Datasets. We evaluate on ***nuScenes*** (Caesar et al., 2020) and ***Argoverse 2 Sensor*** (Wilson et al., 2021), yielding N_Z=11 and 17 cue-first crosswalk scenes that satisfy the identification condition in §3; finite-sample resolution is calibrated per planner by FP and $\text{LOD}_{0.8}$ (§4.3). Event onsets are fixed geometrically before evaluation (d_X=20 m; d_Z=30 m for the crosswalk), and Argoverse 2 uses disjoint training and audit splits.

Models. We evaluate **TREAT** across four VLM families—LFM2-VL (450M/1.6B) and LFM2.5-VL (1.6B/3B) (Amini et al., 2025), Qwen2-VL (2B/7B) (Wang et al., 2024), PaliGemma 2 (3B) (Steiner et al., 2024), and LLaVA-1.6 (7B) (Liu et al., 2024)—together with SmolVLM2-500M and the non-VLM planner **UniAD** (Hu et al., 2023). SmolVLM2-500M is inadmissible (Table 8).

Evaluation Metrics. We evaluate **TREAT** along two axes. **(1) Decision-cause measurement: lock count** n_Z/N_Z, **mean locking contrast** $\bar{L}_Z$ (m/s), **locking-test** p**-value** p_Z, and the resulting **screening outcome or verdict** (baselines yield a thresholded outcome, flagged or not; **TREAT** refines a flag into confirmed misaligned, transient, or persistent and certifies a non-flag as cleared, §4.3), together with calibrated measurement reliability: **FP** (↓, %) and $\textbf{LOD}_{0.8}$ (↓, %). **(2) Decision-cause alignment after repair: post-repair locking** (↓), **human-relevant event response** (↑, m/s), and **braking reproduction** (↑, %; values above 100% mean the planner brakes harder than the human driver). **ADE** (↓, m) is reported separately as a task-preservation metric.

Baselines and controls. We compare against detection and repair baselines and stage-specific controls. **(1) Detection baselines: open-loop evaluation** (Hu et al., 2023), ranked by ADE; **linear probing** (Alain & Bengio, 2016; Belinkov, 2022), measuring cue decodability from frozen features; **pixel-deletion attribution** (Zeiler & Fergus, 2014), removing the agent or cue by inpainting; and **cue-conflict factorial scoring** (Geirhos et al., 2019), evaluating cue/event presence in a 2×2 grid. **(2) Repair baselines: activation steering** (Turner et al., 2023; Rimsky et al., 2024); **direction ablation** (Ravfogel et al., 2020; Arditi et al., 2024); **targeted behavior cloning** (Kaushik et al., 2020); and a **contrastive likelihood margin** (Welleck et al., 2020). **(3) Experimental controls:** a **constructed positive control** with known cue-driven braking, which LOCATE flags as intended (Table 8); a **compute-matched control** with the same training budget as CURE but no decorrelation records; and the **human reference** on the same scenes. Detection methods share the same planners and scenes; repair methods start from the same confirmed instance and are re-audited identically. The detection baselines were not designed to estimate decision-cause alignment; we therefore compare their native diagnostics and whether they isolate the same instance, rather than treating them as equivalent estimators of our target. See App. A.4.

5.2 MEASURING, REPAIRING, AND CERTIFYING DECISION-CAUSE ALIGNMENT

Table 1: **Decision-cause measurement across planners.** Baselines report their native diagnostic and a thresholded screening outcome; **TREAT** reports n_Z/N_Z, $\bar{L}_Z$, p_Z, its calibrated verdict, FP, and $\mathrm{LOD}_{0.8}$. § FP/LOD inherited from the closest cleared peer; ΔADE: agent-deletion effect. † text-recipe instance. $\mathrm{LOD}_{0.8} > 50\%$ indicates that 80% detection was not reached at the maximum tested strength $s = 0.5$.

Dataset	Method	Reading	UniAD	PaliGemma 3B	Qwen2-VL 2B	Qwen2-VL 7B†	LLaVA 7B	LFM2 450M	LFM2 1.6B	LFM2.5 3B
nuScenes	Open-loop eval.	ADE (m) outcome	1.07 not flagged	1.85 flagged	1.82 not flagged	1.75 not flagged	2.23 flagged	1.36 not flagged	**1.62** not flagged	1.19 not flagged
	Linear probing	decodability outcome	1.00 flagged	0.98 flagged	0.95 flagged	0.92 flagged	0.98 flagged	0.98 flagged	1.00 flagged	0.99 flagged
	Deletion attr.	ΔADE (agent) outcome	+0.54 flagged	+1.04 not flagged	+0.38 not flagged	+0.37 flagged	+0.53 flagged	+1.31 not flagged	+1.55 not flagged	+1.47 not flagged
	Cue-conflict	cue effect (m/s) outcome	-0.85 flagged	-2.03 flagged	-2.53 flagged	-3.87 flagged	-1.62 flagged	-1.53 flagged	-1.66 flagged	-1.82 flagged
	TREAT	n_Z/N_Z $\bar{L}_Z$ (m/s) p_Z verdict FP $\mathrm{LOD}_{0.8}$	6/11 -0.01 0.990 cleared 4.0% 50%	7/11 +0.32 0.736 cleared 11.8% 50%	5/11 +0.62 0.259 cleared 14.3% 20%	5/11 -0.19 0.800 cleared 1.2% 40%	5/11 -0.15 0.852 cleared 1.7% >50%	5/11 -0.74 0.553 cleared 0.2% >50%	9/11 +1.22 0.049 **confirmed** 0.5%§ 40%§	4/11 -0.73 0.134 cleared 0.1% 40%
	Human	Logs	0/11, -2.67 m/s, p<0.001: brakes for the agent, not the crosswalk							
Argoverse 2	Open-loop eval.	ADE (m) outcome	1.39 not flagged	2.24 flagged	2.19 not flagged	2.02 not flagged	2.52 flagged	1.56 not flagged	1.43 not flagged	1.37 not flagged
	Linear probing	decodability outcome	1.00 flagged	0.97 flagged	0.95 flagged	0.91 flagged	0.98 flagged	0.96 flagged	1.00 flagged	0.99 flagged
	Deletion attr.	ΔADE (agent) outcome	+0.53 flagged	+1.02 not flagged	+0.41 not flagged	+0.39 flagged	+0.57 flagged	+1.22 not flagged	+1.53 not flagged	+1.40 not flagged
	Cue-conflict	cue effect (m/s) outcome	-0.80 flagged	-2.00 flagged	-2.49 flagged	-3.82 flagged	-1.60 flagged	-1.51 flagged	-1.71 flagged	-1.86 flagged
	TREAT	n_Z/N_Z $\bar{L}_Z$ (m/s) p_Z verdict FP $\mathrm{LOD}_{0.8}$	10/17 +0.12 0.980 cleared 6.8% 50%	10/17 +0.67 0.474 cleared 12.4% 50%	8/17 +0.00 0.995 cleared 3.7% 5%	9/17 +0.01 0.901 cleared 0.4% 40%	10/17 +0.05 0.948 cleared 2.9% 30%	12/17 +1.85 0.014 **persistent** 1.0%§ 30%§	12/17 +1.61 0.049 transient 1.0%§ 30%§	10/17 -0.14 0.828 cleared 1.9% 30%
	Human	Logs	5/17, -0.05 m/s, p=0.902: no locking either way							

Three findings follow, one per stage of **TREAT**. **First, underlying dependence is graded, but alignment is adjudicated at the trained-instance level.** Table 1 shows that LOCATE flags the nuScenes LFM2-VL-1.6B instance ($9/11$, $\bar{L}_Z{=}{+}1.22$ m/s, $p{=}0.049$), whereas four standard diagnostics do not isolate the same planner (they measure trajectory accuracy, cue encoding, or intervention sensitivity rather than the decision-driving event; App. A.4). Figure 3(a) places this instance-level reading in population context: replica fleets occupy different positions on the locking-contrast axis, with LFM2-VL-1.6B showing the strongest observed cue-ward tendency on nuScenes, but the confirmed instance remaining a tail member and none of the 29 additional replicas yielding another confirmation (Table 8). The calibrated detection limits are approximately $+1.5$–$+3.0$ m/s in contrast units, above the fleet-level shifts in Figure 3(a), so fleet position characterizes population susceptibility rather than an individual verdict. The human reference strongly favors the human-relevant event on nuScenes and shows no cue locking on Argoverse 2, further separating planner-specific dependence from the audited scene geometry. **Second, alignment cannot be inferred from the screening boundary alone.** The boundary is necessarily discrete: activation steering leaves the cue-ward effect essentially unchanged ($+1.23$ vs. $+1.22$ m/s) while moving the nominal p-value from 0.049 to 0.051. Table 2 therefore uses post-repair behavior to adjudicate the initial flag. On the confirmed nuScenes instance, CURE reduces $\bar{L}_Z$ from $+1.22$ to $+0.04$ m/s (compute-matched control: $+0.60$ m/s) and raises the human-relevant event response from 0.25 to 0.90 m/s, at a held-out ADE of 2.72 m, up from 1.62 m before repair but 0.5 m below the 3.22 m of the matched continued training. Standard repair baselines do not remove the diagnosed dependence: activation steering leaves the contrast unchanged, while direction ablation, targeted behavior cloning, and the contrastive likelihood margin retain cue-ward contrasts of $+0.99$–$+1.35$ m/s. On Argoverse 2, the LFM2-VL-1.6B

reading falls below the screening threshold after both CURE and the matched control and is therefore transient, whereas LFM2-VL-450M retains a cue-ward reading across both CURE budgets and matched controls and is persistent. Thus, flags can correspond to targeted reversibility, nonspecific disappearance, or persistence under the current repair. **Finally, finite-sample calibration determines the resolution of an alignment claim.** A non-flag is therefore not interpreted as zero cue dependence. CERTIFY measures planner-specific false-positive rates and detection limits by injecting cue-locked defects of known strength and rerunning LOCATE (Figure 3(b); Argoverse 2 in Figure 4(b)). The resulting certificate states the resolution at which decision-cause alignment is supported for the audited weights, response, cue set, and scene population on which the audit is calibrated, rather than asserting exact structural independence.

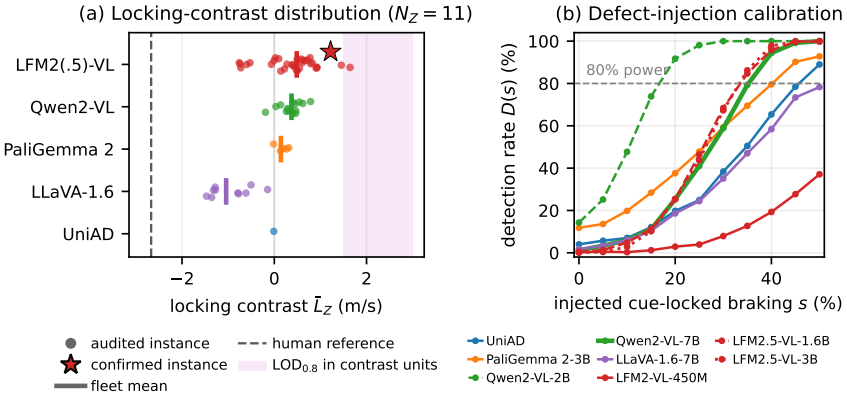


Figure 3: **Population structure and audit calibration on nuScenes.** (a) Locking contrast $\bar{L}_Z$ (cue-ward positive) across audited planner families. (b) Detection rate $D(s)$ under injected cue-locked braking; $D(0)$ gives FP and the 80% crossing defines $\text{LOD}_{0.8}$.

Table 2: **Post-repair outcomes for flagged instances.** "Screen" denotes LOCATE status, not certification. 1×/2× denote multiples of the base decorrelation-record budget.

Dataset	Model	Method	Post-repair locking n_Z/N_Z	$\bar{L}_Z$ (m/s)	p_Z	screen	Event response (↑, m/s)	Braking repro. (↑, %)	ADE (↓, m)
nuScenes	*LFM2 1.6B*	Original instance (confirmed)	9/11	+1.22	0.049	flagged	0.25	42	1.62
		+ CURE	6/11	+0.04	0.960	not flagged	0.90	54	2.72
		+ compute-matched control	6/11	+0.60	0.466	not flagged	0.73	36	3.22
		+ activation steering	9/11	+1.23	0.051	not flagged	0.27	39	1.61
		+ direction ablation	9/11	+0.99	0.074	not flagged	0.23	37	1.83
		+ targeted behavior cloning	8/11	+1.35	0.121	not flagged	0.41	65	2.04
		+ contrastive likelihood margin	10/11	+1.02	0.103	not flagged	0.19	41	2.11
Argoverse 2	*LFM2 1.6B*	Original instance (seed 1, transient)	12/17	+1.61	0.049	flagged	0.23	116	1.43
		+ CURE, 1× budget	10/17	+0.96	0.347	not flagged	0.37	133	2.92
		+ compute-matched control, 1× budget	10/17	+1.17	0.237	not flagged	0.22	128	3.47
	LFM2 450M	Original instance (persistent)	12/17	+1.85	0.014	flagged	-0.26	20	1.56
		+ CURE, 1× budget	14/17	+2.12	0.050	not flagged	0.03	107	3.35
		+ CURE, 2× budget	13/17	+1.96	0.042	flagged	-0.28	52	3.63
		+ compute-matched control, 1× budget	15/17	+2.25	0.030	flagged	-0.24	85	3.84
		+ compute-matched control, 2× budget	14/17	+2.30	0.020	flagged	-0.36	29	4.19

5.3 ROBUSTNESS AND SENSITIVITY OF DECISION-CAUSE MEASUREMENT

Reproducibility. The confirmed dependence is reproducible for fixed weights: two independent replays of the audit return 9/11 with p=0.048/0.049 (Table 8, App. D.4). This establishes reproducibility of the reading, not an independent held-out confirmation: discovery and corroboration use the same pre-specified audit set, and the same weights are not flagged when replayed zero-shot on Argoverse 2 (App. D.4).

Table 3: **Cue-wide robustness scan.** Cells give n_Z/N_Z and p_Z. ^{c} denotes significant locking toward the human-relevant event.

Cue Z	Co-occurrence	Confirmed instance Original	Confirmed instance + CURE	Other planners Qwen2-VL-7B	Other planners PaliGemma 2-3B
Crosswalk	0.659	**9/11, 0.049**	6/11, 0.96	5/11, 0.80	7/11, 0.74
Intersection	**0.841**	3/14, 0.35	7/14, 0.63	2/14, 0.033^c	3/14, 0.23
Stop line	0.687	6/15, 0.93	6/15, 0.72	5/15, 0.13	8/15, 0.92
Traffic light	0.331	3/8, 0.58	2/8, 0.17	0/8, 0.001^c	1/8, 0.010^c

Cue robustness. Across four candidate cues (Table 3), only crosswalk produces a cue-ward flag, despite intersection having the strongest cue–braking co-occurrence in training ($P(\text{cue} \mid \text{braking})$,

0.841 vs. 0.659); the other cues are cleared or lean toward the human-relevant event. Thus, **TREAT** does not simply recover the most correlated cue.

Measurement sensitivity. Detection increases monotonically with injected defect strength (Figure 3(b)): on the reference Qwen2-VL-7B series it reaches 100% at s=0.50 with $\mathrm{LOD}_{0.8} \approx 40\%$, while the other planners' limits spread from 5% (the unusually sensitive Argoverse 2 Qwen2-VL-2B series) to above 50%. Planner-specific FP and detection limits therefore set the resolution of each negative verdict (Table 1).

Identification and coverage. Adding onset-diluted scenes preserves the cue-ward direction but removes significance, consistent with Prop. 3: increasing scene count need not improve power when the added scenes weaken temporal identification (App. D.2). Such windows grow with deployment diversity ($0 \to 25$ across nuScenes regions; Figure 5, App. D.2).

6 Related Work

Related work falls into three approaches to inferring a planner's reason, plus assurance for deployed instances. **(1) Representation-level evidence** asks what a model encodes, highlights, or explains through probing, attribution, deletion, and explanation faithfulness (Alain & Bengio, 2016; Belinkov, 2022; Hewitt & Liang, 2019; Zeiler & Fergus, 2014; Adebayo et al., 2018; Hooker et al., 2019; Jacovi & Goldberg, 2020; Ross et al., 2017; Lapuschkin et al., 2019). These methods characterize representation or salience rather than which world event drives the action. **(2) Intervention-based evidence** infers reliance from behavior under counterfactual inputs, cue conflicts, simulation, or expert interventions (McCoy et al., 2019; Geirhos et al., 2020; 2019; Dosovitskiy et al., 2017; de Haan et al., 2019); related repair methods require the suspected confound, group, or direction to be specified (Arjovsky et al., 2019; Sagawa et al., 2020; Kaushik et al., 2020; Welleck et al., 2020; Turner et al., 2023; Rimsky et al., 2024; Ravfogel et al., 2020; Arditi et al., 2024). The resulting evidence depends on the chosen intervention. **(3) Observational and temporal evidence** uses behavioral associations, dataset confounding, temporal precedence, or reaction timing (Swamy et al., 2022; Langosco di Langosco et al., 2022; Granger, 1969; MacKinlay, 1997; Jordà, 2005; Markkula et al., 2016; Wu & Lin, 2019; Engström et al., 2024; Varotto et al., 2026). **TREAT** instead uses onset separation as an identification condition for a specific trained planner, distinguishing when a cue contribution is unidentifiable, identifiable, and detectable, and linking it to repair and certification. **(4) Safety assurance and instance-level certification** motivates evidence about latent hazards and deployed instances (International Organization for Standardization, 2022; Underwriters Laboratories, 2023; D'Amour et al., 2022), but does not provide calibrated measurement of the decision-driving event. **TREAT** adds a certificate bound to the audited weights, with measured false-positive rates and detection limits (Currie, 1968; Hill, 1965; Westfall & Young, 1993).

7 Conclusion

Behavior cloning aligns actions, but not necessarily their decision-driving events. We define *decision-cause alignment* and show that static behavior with a shared onset and response form cannot separate a human-relevant event from a coincident cue, whereas, under explicit latency and disturbance-isolation assumptions, temporal separation identifies the cue contribution. **TREAT** operationalizes this idea: Locate measures temporal locking, Cure repairs diagnosed dependence with targeted decorrelation records, and Certify calibrates audit resolution. Across two datasets, four VLM families, UniAD, and $100+$ model instances, **TREAT** detects and corroborates one naturally occurring misalignment that four standard diagnostics fail to isolate and repairs it at a smaller task cost than compute-matched continued training. Cue-ward dependence is graded and population-structured across replica fleets, while confirmation remains a rare, instance-level event; certificates therefore bind individual trained weights.

Limitations. Our strongest evidence concerns braking for an agent entering the danger zone and four candidate cues; events without onset-separated scenes remain untestable. Detection limits are finite and instance-specific, and flagged instances currently inherit calibration from cleared peers. The audit is open-loop and makes no closed-loop safety claim, while Cure changes behavior without requiring removal of the cue representation. Both repair arms raise held-out ADE, so Cure improves decision-cause alignment at a measurable task-accuracy cost. The confirmed case is supported by one pre-specified audit set of 11 cue-first scenes, eight of which overlap the training manifest, without an independent held-out replication of the defect. Future work should broaden the audited event and cue sets, automate cue discovery, and re-audit as deployment provides new identifying scenes.

AI USE STATEMENT

In this work we used generative AI tools (large language model assistants) for the following tasks with required disclosure: to help develop and refine the conceptual framework and its terminology, to help formulate and edit the statements of the propositions and assumptions, to propose and refine hypotheses about the audit results, to provide feedback on experimental design (including the null calibration), to implement analysis scripts and job pipelines, to clean and reformat audit outputs, and to help interpret results. We did not use generative AI tools to generate any dataset: all constructed records (decorrelation records assembled from logged scenes and calibration spikes) were produced by rule-based scripts, and all audited trajectories were produced by the planners under test. Translation, qualitative or thematic analysis, and mathematical proofs beyond the proof sketches given in the paper are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing the manuscript for readability, suggesting its structure and title, identifying and verifying relevant literature, formatting references, editing code, and drafting initial layouts for Figures 1 and 2, which were redrawn by the authors. We have reviewed all AI-assisted work: every passage of AI-drafted text was rewritten or approved by the authors; AI-written code was executed end to end and its outputs were checked against the pre-specified decision gates and against independent reruns on a second cluster; every citation added with AI assistance was verified against the publisher's record. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Datasets, co-occurrence statistics, onset geometry, and cue-first scene counts are specified in §5.1 and App. A.2; onsets are fixed before any model runs. Model recipes, replica fleets, and the constructed control instances are described in App. B, and every single-instance verdict in Table 8 names the model instance it binds. Metric definitions and the six-tier verdict protocol are given in App. A.3, the screening-stage null calibration in App. D.1, and baseline configurations in App. A.4. The assumptions behind the identification results are stated explicitly (Assumptions 1 and 2), with proof sketches in the main text and App. C.4. The audit consumes only planned trajectories on replayed logs and was reproduced across two independent clusters; the null-calibration script is included with the released code. Code, the fine-tuning recipes, and the cached response series behind every audit table are available anonymously at `https://anonymous.4open.science/r/TREAT-5C14/README.md`.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, et al. LFM2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 136037–136083, 2024.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Howard S. Bloom. Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, 19(5):547–556, 1995.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2020.

Lloyd A. Currie. Limits for qualitative detection and quantitative determination: Application to radiochemistry. *Analytical Chemistry*, 40(3):586–593, 1968.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11693–11704, 2019.

Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*, pp. 1–16, 2017.

Johan Engström, Shu-Yuan Liu, Azadeh Dinparastdjadid, and Camelia Simoiu. Modeling road user response timing in naturalistic traffic conflicts: A surprise-based framework. *Accident Analysis & Prevention*, 198:107460, 2024.

Euro NCAP. Test protocol – AEB/LSS VRU systems, version 4.5.1. Technical report, European New Car Assessment Programme, February 2024. URL `https://cdn.euroncap.com/cars/assets/euro_ncap_aeb_lss_vru_test_protocol_v451_cb0d5dfd0a.pdf`.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Austin Bradford Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, 1965.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9734–9745, 2019.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023.

International Organization for Standardization. ISO 21448:2022 Road vehicles — Safety of the intended functionality. International Standard, 2022.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205, 2020.

Òscar Jordà. Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182, 2005.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Lauro Langosco di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 12004–12019, 2022.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024.

A. Craig MacKinlay. Event studies in economics and finance. *Journal of Economic Literature*, 35 (1):13–39, 1997.

Gustav Markkula, Johan Engström, Johan Lodin, Jonas Bärgman, and Trent Victor. A farewell to brake reaction times? Kinematics-dependent brake response in naturalistic rear-end emergencies. *Accident Analysis & Prevention*, 95:209–226, 2016.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3428–3448, 2019.

National Transportation Safety Board. Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018. Highway Accident Report NTSB/HAR-19/03, NTSB, 2019.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522, 2024.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017.

Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Causal imitation learning under temporally correlated noise. In *International Conference on Machine Learning (ICML)*, pp. 20877–20890, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Underwriters Laboratories. UL 4600: Standard for safety for the evaluation of autonomous products. Standard, 3rd edition, 2023.

United Nations. Convention on Road Traffic. Vienna, 8 November 1968, United Nations Treaty Series, vol. 1042, p. 17, 1968. URL https://unece.org/DAM/trans/conventn/Conv_road_traffic_EN.pdf.

Silvia F. Varotto, Girish Kumaar Srinivasan Ravi Kumar, Ruomei Liu, Matheus H. W. Stratermans, Reyzha Wikki Zaninshi, Eleonora Papadimitriou, and Meng Wang. What factors influence driver behaviour characteristics when pedestrians cross the road? Insights from the UDRIVE naturalistic driving study. *Transportation Research Interdisciplinary Perspectives*, 37:102013, 2026.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York, 1993.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

Kun-Feng Wu and Ya-Jin Lin. Exploring the effects of critical driving situations on driver perception time (PT) using SHRP2 naturalistic driving study data. *Accident Analysis & Prevention*, 128:94–102, 2019.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

A Implementation Details

A.1 VLM Selection

The audited panel is not a convenience sample. An instance-level claim about decision-cause alignment must be tested against the axes along which a planner's design could plausibly vary, so the panel spans four such axes and adds three boundary members. *(H1) Backbone operator*: LTC-hybrid (LFM2-VL, alternating convolution and attention) versus pure Transformer (Qwen2-VL, PaliGemma 2, LLaVA-1.6). *(H2) Decoder attention connectivity*: causal (LFM2-VL, Qwen2-VL, LLaVA-1.6) versus prefix-LM with bidirectional image attention (PaliGemma 2). *(H3) Scale and lineage*: within the LFM2 family, 450M/1.6B/3B parameters and the LFM2→LFM2.5 generation step, all trained under the flagged instance's own recipe, test whether scale or lineage inherits the misalignment. *(H4) Training exposure*: LLaVA-1.6-7B is trained at a deliberately lighter budget (2 epochs on 8k records) as a weak-but-aligned control, separating weakness from misalignment. Three further members fix the instrument's boundaries: a *constructed positive control* (Qwen2-VL-2B behavior-cloned onto "brake at crosswalks") supplies a known positive; *SmolVLM2-500M* exercises the admissibility prerequisite; and *UniAD* (Hu et al., 2023), a non-VLM end-to-end stack evaluated with its published checkpoints, tests whether the audit is VLM-specific. Table 4 summarizes the panel; every VLM member is behavior-cloned on the same task and image input (App. B), the text-recipe instance additionally receiving numeric ego state, so that differences in verdict cannot be attributed to the task.

Table 4: **Audited VLM panel.** The panel varies model scale, backbone, decoder connectivity, and training exposure while holding the trajectory-planning task fixed; exposure is training records $\times$ epochs, and replicas rerun the same training pipeline with fresh seeds (nuScenes / Argoverse 2). † text-recipe instance (App. B).

Planner	Params	Backbone	Decoder	Input	Exposure	Replicas
LFM2-VL (Amini et al., 2025)	450M	LTC-hybrid	causal	512^2	26.6k $\times 3$	—
LFM2-VL	1.6B	LTC-hybrid	causal	512^2	26.6k $\times 3$	29 / 2
LFM2.5-VL	1.6B / 3B	LTC-hybrid	causal	512^2	26.6k $\times 3$	—
Qwen2-VL (Wang et al., 2024)	2B	Transformer	causal	448^2	26.6k $\times 3$	13 / 13
Qwen2-VL†	7B	Transformer	causal	448^2	26.6k $\times 3$	—
PaliGemma 2 (Steiner et al., 2024)	3B	Transformer	prefix-LM	224^2	26.6k $\times 3$	4 / 4
LLaVA-1.6 (Liu et al., 2024)	7B	Transformer	causal	336^2	8k $\times 2$	9 / 9

A.2 Data and Hyperparameters

Detailed training and LOCATE/CERTIFY settings are summarized in Table 5; the CURE block gives the exact repair budgets used for the results in Table 2.

Table 5: **Implementation settings for TREAT.** Detailed training, audit, repair, and calibration configurations used in the experiments. Dataset statistics are reported separately in Table 6.

Category	Setting
Hardware Configuration	
GPU	NVIDIA H200 (Slurm, one GPU per job, 120 GB host memory) and NVIDIA A100
Precision	bfloat16
Software stack	*transformers* + *peft*; nuScenes devkit; Argoverse 2 API
Main-Task Fine-Tuning (all VLM planners)	
Input	3 consecutive CAM_FRONT frames (Δt=0.5 s), task-only prompt, no numeric text (except the text-recipe instance)
Output	6 future waypoints (t+0.5 s to t+3.0 s, ego meters) as text tokens
Objective	causal language-modeling loss on the trajectory tokens
Adapter	LoRA, rank 16, α=32, dropout 0.05, attention and MLP projections
Optimizer / LR	AdamW, 2×10^{-5} (LLaVA: 1×10^{-5})
Weight decay / warmup ratio	0.01 / 0.10
Effective batch size	16
Epochs / records	3 / 26,595 (LLaVA: 2 / 8,000)
Image resolution	native per VLM (Table 4)
Replica seeds	seed only (data order, initialization); pipeline bit-for-bit fixed
LOCATE configuration	
Danger zone d_X	20 m
Cue thresholds d_Z	crosswalk 30 / intersection 35 / stop line 25 / traffic light 40 m
Response r_t	planned mean speed over the 3 s horizon
Pre-onset baseline / post-onset window	3 frames / 5 frames (cue-first window truncated at τ_X)
Minimum cue-first scenes	5
Screening level α	0.05, two-sided paired t with the cue-ward direction required (one-sided 0.025)
Human reference	logged ego speed on the same scenes and windows
CURE configuration	
Decorrelation records (nuScenes)	2,700 (900 unique, oversampled 3$\times$); targets are logged human futures
Cell composition (unique records)	cue-only 300, event-only 150, both 150, empty 300
Main-task mixture / epochs	20,000 records / 2 epochs, base recipe otherwise
Compute-matched control	same 20,000 records and schedule, no decorrelation records
Audit hygiene	every audited scene excluded from the repair and control mixtures (original training overlaps 8/11 audit scenes; App. D.4)
Argoverse 2 arms	$\approx$1,350 records per decorrelating cell; control add-on from the other cells; budgets 1$\times$ (LFM2-VL-1.6B) and 1$\times$, 2$\times$ (LFM2-VL-450M)
CERTIFY calibration	
Defect injection	$r_t \leftarrow (1-s)\, r_t$ on cue-present frames of each cleared planner's own response series (flagged instances inherit their cleared peers')
Strength grid s	$\{0,\ 0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.30,\ 0.40,\ 0.50\}$
Bootstrap	1,000 resamples over scenes per s
Null calibration	20,000 study-level replications over the 122-cell grid (App. D.1)

Datasets and splits. Table 6 summarizes the datasets and audit sets. All scene inclusion rules and event thresholds are fixed before model evaluation.

Table 6: **Datasets and audit sets.** Cue-first counts refer to the crosswalk cue used in the main audit.

Dataset	Training data	Audit data	Cue-first	Purpose
nuScenes (Caesar et al., 2020)	26,595 records	40 forced-braking scenes	11	Main audit / repair
Argoverse 2 Sensor (Wilson et al., 2021)	16,744 records	496 val records	17	Independent-domain audit

Event definitions. Keyframes are sampled at 2 Hz. The human-relevant event X fires when an agent (pedestrian or vehicle ground-truth track) enters the danger zone within $d_X{=}20$ m of the ego path; a cue Z fires when a map object enters view within its per-cue threshold (d_Z: crosswalk 30 m, intersection 35 m, stop line 25 m, traffic light 40 m). On the retained cue-first scenes the triggering agent is a vehicle in $10/11$ (nuScenes) and $15/17$ (Argoverse 2) cases and a pedestrian in the remainder. Onsets are the first threshold crossing; a scene qualifies for cue Z when $\tau_Z < \tau_X$ (Prop. 2). Scene inclusion is purely geometric and fixed before any model runs. For each onset, the response is baselined on the 3 frames preceding it and averaged over the 5 frames following it ($\ell_{\min}$ of one frame, a 2.5 s window); the cue-first window is truncated at the onset of X so that none of its frames carries a contribution from X. A cue is tested on an instance only when at least five cue-first scenes are available.

Co-occurrence measurement. For each cue, the four-cell contingency (braking $\times$ cue-present) is computed over the full training manifest with braking defined as a planned speed drop exceeding 1.0 m/s over the 3 s horizon; the cue–braking co-occurrence is $P(\text{cue}\,|\,\text{braking})$. Because $P(\text{braking}\,|\,\text{cue}) < P(\text{braking}\,|\,\text{no cue})$ on nuScenes, this quantity is a statement about co-occurrence during braking demonstrations, not about the cue raising braking probability; Argoverse 2 is the exception where the crosswalk positively predicts braking (0.172 vs. 0.135).

A.3 Evaluation Metrics

Let i index audited scenes, $E \in \{X, Z\}$ an event with onset $\tau_{E,i}$, and $r_{i,t}$ the planner's response at frame t, the mean planned speed over the 3 s horizon computed from the six predicted waypoints (keyframes at 2 Hz). Windows follow App. C.1: W_i^X is the $h{=}5$ frames after the agent onset and $W_i^Z \equiv W_i$ the cue-first window truncated at $\tau_{X,i}$, each referenced to the $b{=}3$ frames before its onset. The post-onset braking magnitude is

$$\overline{\Delta v}_{i,\text{post-}E} = \frac{1}{b}\sum_{k=1}^{b} r_{i,\tau_{E,i}-k} \;-\; \frac{1}{|W_i^E|}\sum_{t\in W_i^E} r_{i,t}, \tag{12}$$

with larger values denoting stronger braking.

Locate. The scene-level locking contrast is $L_i(Z)$ of Eq. (9), positive when the response is locked more strongly to the cue than to X. Across the N_Z cue-first scenes,

$$\begin{aligned} n_Z &= \sum_{i=1}^{N_Z} \mathbf{1}[L_i(Z) > 0], \qquad \bar{L}_Z = \frac{1}{N_Z}\sum_{i=1}^{N_Z} L_i(Z), \\ t_Z &= \frac{\bar{L}_Z}{s_Z/\sqrt{N_Z}}, \qquad p_Z = 2\big[1 - F_{N_Z-1}(|t_Z|)\big], \end{aligned} \tag{13}$$

where s_Z is the sample standard deviation of $\{L_i(Z)\}$ and F_{N_Z-1} the Student-t distribution function with $N_Z - 1$ degrees of freedom; n_Z/N_Z is the fraction of scenes leaning toward the cue and $\bar{L}_Z$ the mean contrast (m/s). A cue is flagged when $p_Z < 0.05$ and $\bar{L}_Z > 0$ (one-sided at 0.025); a significant contrast in the X-aligned direction is reported descriptively as locking toward X, never as a flag. The human reference applies the same statistic to the logged ego speed on the same scenes and windows.

Cure. The human-relevant event response is the mean braking magnitude after the agent onset over the same N_Z cue-first scenes,

$$R_X = \frac{1}{N_Z}\sum_{i=1}^{N_Z} \overline{\Delta v}_{i,\text{post-}X} \quad \text{(m/s)}. \tag{14}$$

Braking reproduction compares the planner (π) with the logged human driver (H) on the dataset-specific evaluation scenes $\mathcal{S}_F^{(d)}$ in which the human brakes for the agent:

$$\mathrm{BR}_d = 100 \times \frac{\sum_{i \in \mathcal{S}_F^{(d)}} \overline{\Delta v}^{\,\pi}_{i,\text{post-}X}}{\sum_{i \in \mathcal{S}_F^{(d)}} \overline{\Delta v}^{\,H}_{i,\text{post-}X}} \quad (\%), \tag{15}$$

with a bootstrap 95% confidence interval over scenes. For nuScenes, $\mathcal{S}_F^{(\mathrm{nu})}$ is the 40 forced-braking scenes; for Argoverse 2, $\mathcal{S}_F^{(\mathrm{AV2})}$ is the fixed Argoverse 2 evaluation subset used for the Table 2 entries. Task preservation is the average displacement error on the held-out planning split,

$$\mathrm{ADE} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\| \hat{\mathbf{y}}_{i,k} - \mathbf{y}_{i,k} \right\|_2, \qquad K = 6, \tag{16}$$

where $\hat{\mathbf{y}}_{i,k}$ and $\mathbf{y}_{i,k}$ are the predicted and logged waypoints (meters) of record i.

CERTIFY. For injected strength s, the injected series is $r_{i,t} \leftarrow (1 - s)\, r_{i,t}$ on cue-present frames of a cleared planner's own series (§4.3), and LOCATE is rerun on B=1,000 scene bootstraps:

$$\begin{aligned} \hat{D}(s) &= \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\big[\text{LOCATE flags bootstrap } b \text{ at strength } s\big], \\ \mathrm{FP} &= \hat{D}(0), \qquad \mathrm{LOD}_p = \min\{s : \hat{D}(s) \geq p\}, \end{aligned} \tag{17}$$

the empirical form of Eq. (11), with p=0.8 and $s \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$ in Table 1 (5% steps in Figure 3(b)). Both quantities are instance-specific; a flagged instance inherits the values of the same-family cleared instance closest in scale on the same dataset (otherwise the pooled null of all cleared planners), measured at the reference cue configuration (crosswalk; N_Z=11 on nuScenes, 17 on Argoverse 2) and rescaled by $\sqrt{N_Z^{\mathrm{ref}}/N_Z}$ for other scene counts (Prop. 3). To express a detection limit in the units of Figure 3(a), we inject $s = \mathrm{LOD}_{0.8}$ into the planner's own series and report the resulting mean locking contrast: on nuScenes it lies at $+1.5$ to $+3.0$ m/s across the cleared planners whose limit is reached, and on Argoverse 2 at $+1.9$ to $+2.4$ m/s for PaliGemma 2, LLaVA-1.6, and the two LFM2.5-VL planners, whereas the unusually sensitive Argoverse 2 Qwen2-VL-2B series reaches its 5% limit at only $+0.3$ m/s; these ranges are the shaded bands of Figures 3(a) and 4(a). The screening-stage null calibration (App. D.1) gives the number of nominal flags the whole study should expect under a global null; the corroboration chain beyond the flag is not null-calibrated.

Population-level tilt. Tilt is the fraction of decidable cue-first windows leaning cue ($L_i(Z) > 0$) pooled over a replica fleet, compared with a null pooled from the seven cleared native nuScenes VLM instances (one training run each); fleet replicas do not enter the null pool. The comparison uses Welch's t on per-run lock fractions, the run being the unit of replication, with a permutation test over run labels; tilt is reported only at the population level and does not determine any instance-level verdict.

Verdict protocol (pre-specified). p_Z<0.05 with $\bar{L}_Z$>0 flags a cue; confirmation additionally requires the corroboration chain of §4.3: the human reference favors X, the flag is cue-specific, and repair specificity holds, i.e., CURE returns the reading to not flagged while retaining at most half of the cue-ward contrast retained by the compute-matched control; replication of the reading on an independent replay is a prerequisite rather than a chain element. Tiers: *confirmed misaligned*, *persistent* (no arm, at either budget, reduces the cue-ward contrast below the original reading), *transient* (disappears under both arms without a CURE-specific reduction), *unadjudicated* (flagged but outside the three descriptions), *cleared* (bounded negative certificate with stated miss risk), *inadmissible* (no audited cue has $N_Z \geq 5$ usable cue-first scenes). No nominal p-value serves as a family-wise decision criterion at any point; the null calibration of App. D.1 concerns the screening stage only.

A.4 BASELINES

We compare against four detection baselines and four repair baselines, and run three controls, one per stage. All baselines run on the identical planners and audit scenes, so a disagreement with LOCATE isolates the diagnostic rather than the data.

(1) Detection baselines.

- **Open-loop evaluation (ADE)** (Hu et al., 2023). The standard open-loop trajectory score, computed on the same scenes; the two planners with the worst ADE on each dataset are flagged.
- **Linear probing** (Alain & Bengio, 2016; Belinkov, 2022). Layerwise linear decodability of the cue from frozen, pooled hidden features; reports whether the cue is *encoded*.
- **Pixel-deletion attribution** (Zeiler & Fergus, 2014). Inpainting-based removal of the agent or the cue, controlled by a sky-region null edit (Table 1: agent removal raises ADE for every planner, LFM2-VL-1.6B by $+1.55$ m and Qwen2-VL-7B by $+0.37$ m, so both respond to the agent while the ranking is inverted).
- **Cue-conflict factorial scoring** (Geirhos et al., 2019). A static cue-present/absent $\times$ X-present/absent grid (100 frames per cell), which alarms on context-appropriate caution.

None of the four reports a calibrated error rate; their validation is interpretive: held-out trajectory error on i.i.d. data for open-loop evaluation, control tasks and chance-level comparison for probing (Hewitt & Liang, 2019), sanity checks and null edits for deletion attribution (Adebayo et al., 2018; Hooker et al., 2019), and agreement with human observers for cue-conflict scoring (Geirhos et al., 2019).

(2) Repair baselines. All arms start from the flagged weights and are re-audited with the identical instrument.

- **Activation steering** (Turner et al., 2023; Rimsky et al., 2024). Injection of the cue direction at the readout and mid-stack.
- **Direction ablation** (Ravfogel et al., 2020; Arditi et al., 2024). Projecting out the cue direction in the mid layers.
- **Targeted behavior cloning** (Kaushik et al., 2020). LoRA fine-tuning on the 900 unique decorrelation records at the audited locus (1 epoch, LR 1×10^{-5}).
- **Contrastive likelihood margin** (Welleck et al., 2020). A hinge on the teacher-forced NLL of the logged future with the agent present vs. inpainted out (margin 0.3, weight 1.0, 2 epochs); active in $17/54$ pairs from an initial gap of -0.008.

(3) Experimental controls.

- **Constructed positive control** (CERTIFY). Qwen2-VL-2B with a LoRA behavior-cloned onto "brake at crosswalks" records: a known misaligned instance that the audit must flag, used to check instrument sensitivity ($11/11$, p=0.002).
- **Compute-matched control** (CURE). The same schedule and main-task records as CURE, without the decorrelation content; the two-arm contrast (CURE vs. control) doubles as the transient/persistent classification of App. D.3.
- **Human reference** (LOCATE). The locking statistic computed on the logged human trajectories of the same scenes and windows, which checks that the audited scenes do not themselves favor the cue (nuScenes $0/11$ cue-locked, $\bar{L}_Z$=-2.67 m/s, p<0.001: the drivers brake for the agent; Argoverse 2 $5/17$, -0.05 m/s, p=0.90: no locking either way; Table 1).

B TRAINING OF THE AUDITED PLANNERS

Fine-tuning protocol. Every VLM planner is behavior-cloned in two stages with LoRA (rank 16, α=32, dropout 0.05; AdamW 2×10^{-5} (LLaVA: 1×10^{-5}), weight decay 0.01, warmup 0.10, effective batch 16, bfloat16). The first stage adapts the base model to driving imagery with decision question answering on CAM_FRONT frames and is merged into the weights. The second stage produces the audited planner: the input is the task instruction and three consecutive CAM_FRONT

frames (2 Hz) with every numeric text feature removed from the first epoch on, because ego speed and lead distance given as text predict the trajectory so well that the model otherwise ignores the images; the target is the logged ego future as six waypoints ($t{+}0.5$ to $t{+}3.0$ s, ego meters) serialized as text tokens under the causal language-modeling loss, for 3 epochs on the 26,595 nuScenes records (LLaVA: 2 epochs on 8,000, a deliberately lighter budget that makes it the weak-but-aligned control). Each model uses its native resolution and prompt format (LFM2 lineage 512^2; Qwen2-VL two-stage 448^2; PaliGemma 2 224^2 raw prefix; LLaVA 336^2 chat template). The adapter is merged, and the merged checkpoint is the model instance that is evaluated, audited, and, where applicable, repaired; replicas rerun this pipeline with only the seed changed. The confirmed instance is the reference recipe's own output, and every LFM2 scale and lineage control (LFM2-VL-450M, LFM2.5-VL-1.6B, LFM2.5-VL-3B) is trained under that identical recipe; Argoverse 2 instances follow the same recipe on the Argoverse 2 train split.

Pre-audit evaluation and admissibility. Each merged instance passes three checks before the audit. (i) On 300 held-out records disjoint from training, every output must parse into six waypoints, and ADE/FDE are computed against the logged futures. (ii) The same evaluation is repeated with the images blanked: a planner that reads the images loses most of its accuracy (LFM2-VL-1.6B $1.62 \rightarrow 8.15$ m, PaliGemma 2 $1.85 \rightarrow 8.01$ m), whereas a planner whose text channel answers for it does not, which is why text-recipe instances (the *text recipe* keeps numeric ego-state features in the prompt) are audited only with the caveat marked †. (iii) The instance is replayed on the 40 nuScenes forced-braking scenes to obtain its response series and its braking reproduction, the event-aligned post-onset deceleration as a percentage of the human driver's, with a bootstrap 95% CI. Frames whose output collapses to a constant or zero trajectory are treated as missing; a scene is usable only if at least five finite frames remain, and a cue is tested only if at least five usable cue-first scenes remain. An instance that fails this on every cue is inadmissible: SmolVLM2-500M, with 72% degenerate frames, is the one case. Admission never depends on the audit outcome.

Replica fleets. Replicas rerun the identical pipeline with the random seed as the only change (data order and initialization), with code and data bit-for-bit fixed: LFM2-VL-1.6B $\times 29$, Qwen2-VL-2B $\times 13$, LLaVA-1.6-7B $\times 9$, PaliGemma 2 $\times 4$ (55 replicas). On Argoverse 2 the fleets are LFM2-VL-1.6B $\times 2$, Qwen2-VL-2B $\times 13$, PaliGemma 2 $\times 4$, and LLaVA-1.6-7B $\times 9$ (28 replicas). No replica is selected or discarded on its audit outcome; every trained replica enters Table 8.

Constructed, repaired, and induced instances. The *positive control* is Qwen2-VL-2B with a LoRA behavior-cloned onto "brake at crosswalks" records, a known positive used to check instrument sensitivity. The *repaired instance* (π_{Θ^+}) continues the flagged weights for 2 epochs on 20,000 main-task records mixed with 2,700 decorrelation records (900 unique, oversampled $3\times$; cue-only, event-only, both, and empty cells in the proportions of Table 5); every target is a logged human future, never a fabricated trajectory, and every audited scene is excluded from the mixture. The *compute-matched control* receives the same 20,000 records for the same 2 epochs without the decorrelation records. On Argoverse 2, the three seeds under the confirmed instance's recipe and the LFM2-VL-450M instance are trained on the training split; the repair and control add-ons ($\approx 1{,}350$ records per decorrelating cell, and the same count from the other cells) are applied at $1\times$ and $2\times$ budget. *Injected instances* M_s (App. A.3) are synthetic modulations of a cleared planner's response series and involve no training.

Non-VLM planner. UniAD is evaluated with its published stack and checkpoints, replayed on the same scenes; its planning trajectory supplies the response scalar, and its degenerate-frame rate (7%) and train–audit overlap are disclosed with its verdict.

C Method Details

C.1 Locate: Windows, Statistic, and Cue Scan

Onset-aligned windows. For each candidate cue $Z \in V$, compute the event onsets $\tau_{Z,i}$ and $\tau_{X,i}$ from tracks and map layers, and retain cue-first scenes satisfying $\tau_{Z,i} < \tau_{X,i}$. For each retained scene i, construct the cue-first window $W_i = [\, \tau_{Z,i} + \ell_{\min},\ \tau_{X,i} \,)$, truncated to at most h frames, the conservative subset of the window of Prop. 2 on which the cue can already act while X cannot yet,

and the event window $W_i^X = [\,\tau_{X,i} + \ell_{\min},\ \tau_{X,i} + \ell_{\min} + h\,)$ of h frames after the event onset; in our study $\ell_{\min}$ is one frame and $h = 5$ frames (2.5 s at 2 Hz). Windows are clipped at the end of the log, and a scene whose cue-first window is empty is not retained. Let N_Z denote the number of retained scenes for cue Z.

Response and statistic. Replay π_Θ on each retained scene and extract its braking response r_t, the planned mean speed over the planning horizon; the post-onset braking magnitudes Eq. (12), the locking contrast Eq. (9), and the aggregate $(n_Z, \bar{L}_Z, t_Z, p_Z)$ of Eq. (13) are defined in App. A.3. LOCATE tests relative locking rather than estimating β_Z directly: a planner whose X-aligned response dominates a smaller cue-aligned response may therefore be cleared at the audit's calibrated resolution despite nonzero cue dependence. The audit likewise does not estimate β_X directly; braking reproduction is reported to show that the audited response channel is behaviorally expressed, but it is not an admissibility criterion. LOCATE uses Prop. 2 only to construct the contrast, treating scenes as the unit of replication. The same statistic applied to the logged human trajectories on the same scenes provides the human reference, and p_Z is a screening statistic rather than a planner-level verdict; its realized false-positive rate and detection limit are calibrated by CERTIFY.

Cue scan and flag rule. The procedure is repeated for every $Z \in V$, one test per cue; in our study $V = \{\text{crosswalk, intersection, stop line, traffic light}\}$, and a cue is tested only when $N_Z \geq 5$. For each tested cue, LOCATE returns the reading $(N_Z, n_Z, \bar{L}_Z, p_Z)$. A cue is flagged when $p_Z < \alpha$ and $\bar{L}_Z > 0$, where α is a screening level (0.05 in our study) rather than an adjudication threshold; because the cue-ward direction is required, the rule is one-sided at level $\alpha/2$. The flagged cues form

$$\mathcal{O}(\pi_\Theta, r) = \{\, Z \in V :\ p_Z < \alpha,\ \bar{L}_Z > 0 \,\}. \tag{18}$$

Cues with fewer than five retained scenes receive no reading and are recorded as untested. A co-onset event that would violate Assumption 2 is added to V and audited as its own cue, as in §3. A cue-first scene is usable when the planner yields a well-formed response series with sufficient finite frames for the onset-aligned windows; a cue is audited only when $N_Z \geq 5$ usable cue-first scenes remain, and an instance is marked inadmissible only if no audited cue meets this requirement. Braking reproduction is reported as a behavioral metric but is not used as an admissibility criterion.

C.2 CURE: DECORRELATION DATA AND OBJECTIVE

CURE acts only on cues flagged by LOCATE; if $\mathcal{O}(\pi_\Theta, r) = \varnothing$, no repair is applied. Otherwise, guided by the generation–likelihood dissociation (App. C.4), the flagged cues are handled jointly by changing the training evidence at those cues through targeted fine-tuning.

Decorrelation data. For each flagged cue $Z \in \mathcal{O}(\pi_\Theta, r)$, the decorrelation set $\mathcal{D}_{\text{dec}}(Z) = \mathcal{D}_{\text{off}}(Z) \cup \mathcal{D}_{\text{diag}}(Z)$ contains all four cue–event configurations. The two off-diagonal cells $\mathcal{D}_{\text{off}}(Z)$ provide the identifying evidence that breaks cue–event confounding,

$$b^\star = \begin{cases} 0, & (Z, X) = (1, 0), \\ b_X^\star, & (Z, X) = (0, 1), \end{cases} \qquad b_X^\star > 0, \tag{19}$$

where $b^\star$ is the target post-onset braking magnitude, oriented as in Eq. (12) so that larger values mean stronger braking, and $b_X^\star$ is the desired braking for the human-relevant event: the cue alone should not trigger braking, whereas X should still do so without the cue. The matched on-diagonal records $\mathcal{D}_{\text{diag}}(Z)$, $(Z, X) \in \{(1, 1), (0, 0)\}$, retain the original contexts and stabilize task behavior. On nuScenes we use 300 cue-only, 150 event-only, 150 both-present, and 300 neither-present unique records, so the off-diagonal and on-diagonal groups each contribute 450 records before $3\times$ oversampling (Table 5). For all flagged cues, pool the records as

$$\mathcal{D}_{\text{dec}}(\mathcal{O}) = \bigcup_{Z \in \mathcal{O}(\pi_\Theta, r)} \mathcal{D}_{\text{dec}}(Z). \tag{20}$$

Fine-tuning. Starting from the audited weights Θ, the pooled decorrelation data $\mathcal{D}_{\text{dec}}(\mathcal{O})$ are mixed with the original training data and the main-task objective is continued with the mixture objective of Eq. (10), the decorrelation records oversampled $3\times$. The first term preserves the original planning task, while the second incorporates the targeted evidence that separates X from the flagged cues; task regression is not constrained during training and is reported afterwards as held-out ADE (Table 2).

A compute-matched control starts from the same Θ and receives the same additional optimization budget but excludes $\mathcal{D}_{\mathrm{dec}}(\mathcal{O})$, isolating the effect of the targeted evidence from that of continued training alone.

Re-audit. Both π_{Θ^+} and the compute-matched control are returned to LOCATE for a full-vocabulary re-audit and evaluated on the main task. This checks whether the flagged signal is specifically reduced by the targeted repair, whether it migrates to another cue, whether continued training alone produces the same change, and whether task quality regresses. The post-repair readings and task errors are passed to CERTIFY.

C.3 CERTIFY: CALIBRATION, ADJUDICATION, AND CERTIFICATE

Calibration. Injected instances M_s are constructed by injecting cue-locked braking of known strength $s \in [0, 1]$ into a cleared planner's response series on cue-present frames, so that the calibration truth is known by construction rather than taken from the audit's own reading; running LOCATE on each M_s gives the detection curve, FP, and LOD_p of Eq. (11). The calibration is instance-specific: the defects are injected into each cleared planner's own response series, so the $s{=}0$ endpoint is that planner's unmodified behavior, and any residual cue-ward tilt in it inflates FP, the conservative direction. A flagged instance, whose own series cannot serve as a null, inherits the calibration of its cleared peers: the same-family cleared instance closest in scale on the same dataset when one exists, otherwise the pooled null of all cleared planners. Both quantities are measured at the reference cue configuration (the crosswalk cue at its N_Z cue-first scenes) and transferred to other cues by Prop. 3.

Adjudication. A cue that is not flagged is cleared at the level

$$\delta_Z = \mathrm{LOD}_p(N_Z) \approx \mathrm{LOD}_p \sqrt{N_Z^{\mathrm{ref}}/N_Z}, \tag{21}$$

the detection limit rescaled from the reference count ($N_Z^{\mathrm{ref}}{=}11$ on nuScenes, 17 on Argoverse 2) to the cue's own scene count by Prop. 3, with miss risk at most $1 - p$ for dependence of strength at least δ_Z. A flagged cue is confirmed misaligned only if it survives the corroboration chain, patterned on Hill's viewpoints for weighing observational evidence (Hill, 1965): the human reference favors X, the signal is cue-specific, and repair specificity holds: CURE returns the reading to not flagged while retaining at most half of the cue-ward contrast retained by the compute-matched control. The chain corroborates the contribution identified in Prop. 2; it adds no identification of its own. This is a magnitude criterion rather than a significance test: on the confirmed nuScenes instance the residual contrasts are $+0.04$ (CURE) versus $+0.60$ (control), whereas the scene-paired difference between the two arms is not itself significant at $N_Z{=}11$ ($+0.56$ m/s, paired t, $p{=}0.16$). Flagged cues that fail this chain receive no certified risk and are recorded descriptively (§4.3): transient when the reading disappears under both arms without a CURE-specific reduction, persistent when no arm at either budget reduces the cue-ward contrast below the original reading; a flagged cue that falls outside these categories (none occurred in this study) is reported as flagged but unadjudicated, without a certified verdict, in which case detection stands and repair failed. Because confirmation requires a flag,

$$\Pr(\textsc{confirmed} \mid \beta_Z = 0) \le \mathrm{FP}, \tag{22}$$

the chain is corroborative: apart from the flag, its elements are not null-calibrated (App. D.1 calibrates the screening stage only).

Certificate. CERTIFY issues the instance-level certificate of §4.3,

$$\begin{aligned} \mathcal{C}(\pi_\Theta, r) &= \big(d,\ \mathrm{FP},\ \mathrm{LOD}_p,\ V\big), \qquad d = (d_Z)_{Z \in V}, \\ d_Z &\in \{\textsc{cleared},\ \textsc{confirmed},\ \textsc{transient},\\ &\qquad \textsc{persistent},\ \textsc{unadjudicated},\ \textsc{untested}\}, \end{aligned} \tag{23}$$

where V is the audited cue set, FP and LOD_p are the instance's calibrated false-positive rate and detection limit at the reference cue configuration, and each cleared cue carries its clearance level δ_Z of Eq. (21) at its coverage N_Z. The risk attached to each entry follows from the calibration: a cleared cue is missed with probability at most $1 - p$ for dependence of strength at least δ_Z, and, under $\beta_Z = 0$, a confirmed verdict arises with probability at most FP because confirmation requires a flag Eq. (22); transient and persistent entries carry no certified risk. The certificate is bound to the

weights actually audited, because re-running an identical pipeline yields instances whose verdicts differ (App. D.2). After CURE, the repaired weights Θ^+ define a new model instance and must be re-audited before receiving a new certificate. Untested cues are recorded as such, and inadmissible instances receive no certificate; because N_Z grows as deployment accumulates cue-first scenes, untested cues are re-audited as the fleet's own diversity enlarges the instrument.

C.4 TRAINING-SUPPORT INSUFFICIENCY (GENERATION–LIKELIHOOD DISSOCIATION)

Props. 1–3 concern what a planner's behavior reveals about β_Z; Prop. 4 concerns what its training objective reveals, which is what any repair method acts on, and is the basis for the decorrelation data of CURE (§4.2). Its full statement and proof sketch follow.

Full statement of Prop. 4 (generation–likelihood dissociation). Let $\pi_\Theta(y \mid x)$ be the planner as an autoregressive policy over trajectory tokens y given observation x, trained on demonstrations $(x, y^\star) \sim \mathcal{D}_{\mathrm{tr}}$ with the teacher-forced likelihood $L(\Theta) = -\mathbb{E}_{(x,y^\star)\sim\mathcal{D}_{\mathrm{tr}}} \sum_t \log \pi_\Theta(y_t^\star \mid x, y_{<t}^\star)$, and let $\hat{y}_t = \arg\max_a \pi_\Theta(a \mid x, \hat{y}_{<t})$ be its generated rollout at deployment. If the cue and the human-relevant event co-occur on the training support ($Z \equiv X$, the static case of Prop. 1), then there exist Θ_1, Θ_2 with $L(\Theta_1) = L(\Theta_2)$ whose rollouts differ on cue-only inputs $(Z{=}1, X{=}0)$. Consequently, any repair objective that is a functional of L on the training support is invariant on this pair: it cannot distinguish the misaligned member from the aligned one, and cannot guarantee removal of a β_Z-type misalignment; whether an optimizer lands on the right member is left to implicit bias, not to the objective. Here, $y^\star$ is the demonstrated trajectory and $y_{<t}^\star$ its prefix, $\mathcal{D}_{\mathrm{tr}}$ is the training distribution, and the scope is the likelihood-functional family (behavior-cloning patches, contrastive margins); activation-level edits (steering, ablation) fall outside the proposition and are ruled out empirically instead (§5.2).

Proof sketch. L constrains π_Θ only along reference continuations $y_{<t}^\star$ and only on the support of $\mathcal{D}_{\mathrm{tr}}$, where $Z \equiv X$. The $\arg\max$ decision boundary on the unvisited cell $(Z{=}1, X{=}0)$ is not pinned by L; two L-equal networks may place it differently. □

D ADDITIONAL EXPERIMENTS

This section collects the screening-stage null calibration, the sensitivity analyses, the case studies, and the complete readings with generalization checks that support the main results of §5.

D.1 NULL CALIBRATION OF THE SCREENING STAGE

Using a Westfall–Young-style study-level resampling scheme (Westfall & Young, 1993), we run $20{,}000$ global-null replications by centering the per-scene paired differences pooled from the four cleared nuScenes instances and resampling them at every cell of the study grid (122 cells at their observed scene counts: 65 nuScenes crosswalk audits including all fleet replicas, 38 Argoverse 2-native panels including all fleet replicas and the positive control, 7 zero-shot panels, and the 4×3 extra-cue vocabulary cells). Cues are screened with a one-sided cue-ward paired t-test at 0.05, a more liberal rule than the released two-sided-with-direction rule (equivalent to one-sided 0.025); the resulting null flag counts therefore provide an upper bound on the count expected under the released rule. Under this null, the study expects 7.6 ± 2.7 nominal flags (median 7, 90% interval 3–12); the per-cell flag rate is 6.2%, slightly above the nominal 5%, consistent with heavy-tailed scene-level responses. Among the non-constructed panels of this grid, the real study produced four nominal flags (three native and one zero-shot), plus two removed by quality control; the four observed nominal flags fall within the 90% interval under this more liberal null screen, and the two constructed positive controls flag as intended. A nominal flag is therefore not evidence by itself: the confirmed instance is supported by its location within the fleet distribution (App. D.2) together with the pre-specified corroboration chain (App. C.3), whose elements beyond the screening flag are descriptive rather than null-calibrated.

D.2 Sensitivity Study

Instrument sensitivity and identification. LOCATE responds monotonically to known injected defects: on the Qwen2-VL-7B reference series, detection rises from about 1% at s=0 to about 60% at s=0.30 and 100% at s=0.50 (Figure 3(b)). Across cleared planners, FP ranges from 0.1–14% and $\mathrm{LOD}_{0.8}$ from 5–>50% (Table 1); the Argoverse 2 curves are shown in Figure 4(b). Thus, the audit has graded sensitivity rather than a binary operating point, while the variation across planners motivates instance-specific calibration. Identification quality also matters: expanding the crosswalk set from 11 cue-first scenes to 21 onset-diluted scenes preserves the cue-ward direction (14/21) but removes significance (p=0.16; Table 7). Adding scenes therefore does not increase power when those scenes violate the temporal separation that supplies identification, consistent with Prop. 3.

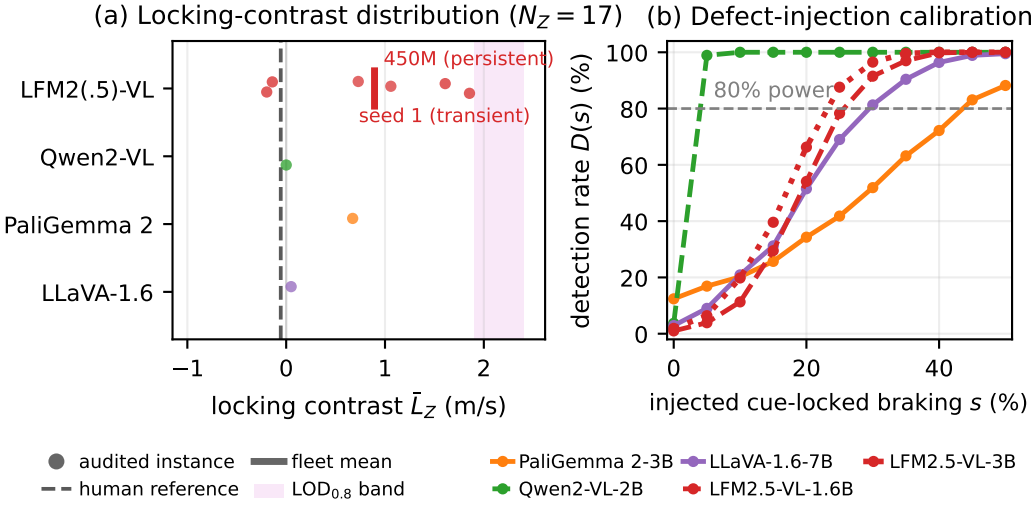


Figure 4: **Population structure and audit calibration on Argoverse 2.** (a) Locking contrast $\bar{L}_Z$ of the native-trained instances by family, conventions as in Figure 3(a): all instances (dots) with the mean of the LFM2-VL-1.6B seed-replica fleet (bar), including the transient seed-1 and persistent LFM2-VL-450M readings, the human reference (dashed, −0.05 m/s), and the calibrated detection limits in contrast units (shaded). (b) Detection rate $D(s)$ for the five planners calibrated on their own Argoverse 2 series; $D(0)$ gives FP and the 80% crossing gives $\mathrm{LOD}_{0.8}$.

Audit coverage. The audit is informative only where the cue onset precedes the onset of X. Figure 5 shows that the number of cue-first crosswalk windows increases as deployment environments accumulate, turning an initially untestable cue into an admissible audit.

Co-occurrence does not determine the cue. Within the nuScenes training world, the confirmed cue is not the most correlated one: intersection has higher cue–braking co-occurrence than crosswalk (0.841 vs. 0.659), yet only crosswalk produces a cue-ward flag (Table 3). Hence cue–braking co-occurrence does not by itself determine which cue drives an individual planner.

Population susceptibility is graded; the instance is thresholded. Replica fleets occupy different positions on the locking-contrast axis (Figure 3(a); Argoverse 2: Figure 4(a)): on nuScenes the LLaVA-1.6-7B fleet locks entirely toward the agent (mean $\bar{L}_Z = -1.04$ m/s, 0/9 replicas cue-ward), PaliGemma 2-3B and Qwen2-VL-2B are mildly cue-ward (+0.14, 3/4; +0.37, 13/13), and LFM2-VL-1.6B is shifted furthest (+0.49, 25/29), with the pooled tilt tests of Table 7 placing the same fleets at the two extremes, LFM2-VL highest and LLaVA lowest. The confirmed LFM2-VL-1.6B seed lies in this fleet's high-dependence tail together with two replicas having larger point estimates but wider intervals; none of the 29 replicas yields an additional confirmed instance. Fleet position therefore characterizes population susceptibility, whereas the certificate must bind the realized weights: at N_Z=11, scene-level uncertainty remains large (typical 95% CIs of roughly ±1.5 m/s), so the ordering of nearby checkpoints is only weakly resolved.

Robustness to analysis choices and multiplicity. The verdict is not driven by a single favorable geometric threshold. Across a 3×3 grid of (d_X, d_Z) and alternative alignment windows, the four

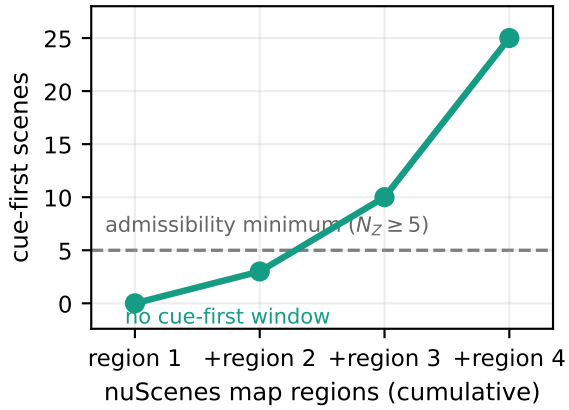


Figure 5: **Coverage grows with deployment diversity.** Cue-first windows for the crosswalk cue across cumulative nuScenes map regions, counted over the 96 harvested scenes. In a single region the audit has no cue-first window for this cue; the count passes the admissibility minimum ($N_Z \geq 5$) as regions accumulate.

cleared planners produce no flags over 44 tested cells, whereas the confirmed instance retains the cue-ward signal most strongly for the tighter danger-zone settings and attenuates as the cue-first window is shortened (Table 7). This pattern is consistent with Prop. 3: advancing the human-relevant event reduces the interval in which the cue can be separately identified. Finally, the screening-stage null calibration over the study's 122-cell grid (App. D.1) shows that several nominal cue-level flags are expected under multiplicity; confirmation therefore rests on fleet position and corroboration rather than on an isolated p-value.

Table 7: **Sensitivity checks.** Controlled variations of identification, architecture, and audit geometry.

Check	Variation	Result
Identification	11 cue-first → 21 onset-diluted scenes	Direction preserved (14/21), but p=0.16
Population tilt	Identical task across replica fleets	LFM2-VL 61.4% (p=6×10^{-4}); PaliGemma 2 59.1% (p=0.029); Qwen2-VL 57.3% (p=0.0029); LLaVA 37.4% (n.s.)
Audit geometry	3×3 (d_X, d_Z) grid + alternative windows	Cleared planners: 0/44 flags (min p=0.09); confirmed instance is strongest for $d_X \leq 20$ m and attenuates as the cue-first window narrows

D.3 Case Studies

From a flag to a confirmed misalignment. The nuScenes LFM2-VL-1.6B instance illustrates how the evidence accumulates for a single trained planner. LOCATE first detects cue-ward braking on the crosswalk audit (9/11, $\bar{L}_Z$=+1.22 m/s, p=0.049; Table 1), while the logged human reference leans strongly toward the agent (0/11, $\bar{L}_Z$= − 2.67 m/s). The signal is specific to the crosswalk rather than simply the most correlated training cue (Table 3), and reproduces for the fixed weights on an independent replay. Most importantly, continued training with the same compute budget does not selectively remove the diagnosed dependence, whereas targeted decorrelation with CURE nearly eliminates the locking contrast while restoring the human-relevant event response (Table 2). The verdict therefore depends on a chain of mutually constraining observations, rather than on the nominal p-value of the original flag alone.

The same flag can imply different repair regimes. The Argoverse 2 cases show why detection and repair must be separated. For the LFM2-VL-1.6B seed, the initial flag disappears after either CURE or compute-matched continued training without a CURE-specific reduction; we therefore classify it as *transient*. By contrast, the LFM2-VL-450M instance remains cue-ward under both CURE budgets and their matched controls and is classified as *persistent* (Table 2). Together with the confirmed nuScenes instance, these cases expose three qualitatively different outcomes—targeted reversibility,

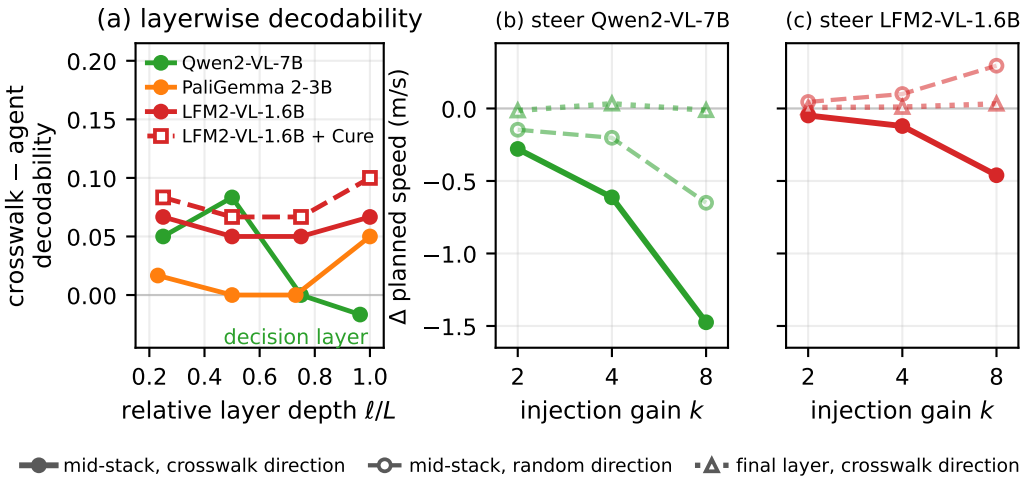


Figure 6: **Mechanistic anatomy of aligned and misaligned planners.** (a) Layerwise decodability of crosswalk versus agent information for three audited planners and the repaired instance (LFM2-VL-1.6B + CURE). (b,c) Activation steering along the crosswalk direction at mid-stack and final layers, against a random direction: cue-responsive pathways exist in both aligned and misaligned planners, and their behavioral use, rather than mere presence, differs.

nonspecific disappearance, and resistance to the current repair—that cannot be distinguished from the initial flag alone.

What representation analysis can—and cannot—explain. Figure 6 helps interpret these behavioral verdicts but does not replace them. Crosswalk information is highly decodable in all examined planners (0.92–1.00), including planners that are behaviorally aligned. The aligned Qwen2-VL-7B suppresses the cue relative to the agent near its decision layer, whereas the flagged instance remains cue-dominant toward the output; PaliGemma 2-3B, however, is aligned without exhibiting the same intermediate reversal. Activation steering likewise reveals cue-responsive pathways in both aligned and misaligned planners. Thus, representation-level evidence can localize how a diagnosed dependence may be implemented, but the presence or decodability of a cue is neither necessary nor sufficient for a decision-cause verdict.

Beyond VLMs. Applying the identical audit to UniAD yields a cleared verdict ($6/11$, p=0.99; Table 1). This case indicates that the measurement principle is not tied to language-model decoding or VLM-specific representations: it operates on the planner's temporal response and therefore extends to a non-VLM trajectory planner.

D.4 FULL LOCATE READINGS AND GENERALIZATION CHECKS

Table 8 reports the complete native-training LOCATE results for every admissible planner, control, and replica fleet on nuScenes and Argoverse 2. Main-text repair outcomes are reported separately in Table 2. The full panel confirms the main-text pattern: most model instances are cleared, while the few flagged cases differ in their subsequent repair response. Population-level fleet statistics are summarized separately in App. D.2.

Reproducibility and training replication. For fixed weights, the nuScenes flag reproduces across two independent replays on separate clusters ($9/11$ in both runs, p=0.048 and 0.049), showing that the reading is not a replay artifact. By contrast, retraining the identical LFM2-VL-1.6B pipeline with the random seed as the only change yields no additional flagged instance among 29 replicas (Table 8); on Argoverse 2, the replica fleets of Qwen2-VL-2B, PaliGemma 2-3B, and LLaVA-1.6-7B (13, 4, and 9 replicas) and the two LFM2-VL-1.6B replicas likewise contain no flagged member. Thus, the measurement is reproducible for a fixed planner, while the misalignment is not determined by the training pipeline.

Table 8: **Complete native-training LOCATE readings.** For each model instance and dataset, we report the number of cue-first scenes N_Z, lock count n_Z/N_Z, mean locking contrast $\bar{L}_Z$ (m/s), paired-t p_Z, and the resulting verdict. Replica rows summarize the fleet by mean lock count, mean contrast, minimum p_Z, and number of flagged members. Fleet sizes are given as $\times n$ / $\times m$ for nuScenes / Argoverse 2; a single $\times n$ applies to both datasets. Repair outcomes are reported in Table 2; zero-shot transfer is reported in Table 9. — = not applicable (human reference). †Text-recipe instance (numeric ego state in the prompt); its cleared verdict certifies the audited response, not visual alignment. *Instrument check with a constructed dependence; no repair arms, hence no certified verdict.

Group	Instance	nuScenes N_Z	nuScenes n_Z/N_Z	nuScenes $\bar{L}_Z$	nuScenes p_Z	nuScenes verdict	Argoverse 2 N_Z	Argoverse 2 n_Z/N_Z	Argoverse 2 $\bar{L}_Z$	Argoverse 2 p_Z	Argoverse 2 verdict
Models	**LFM2-VL-1.6B** (confirmed instance, seed 1)	11	9/11	+1.22	0.049	**confirmed**	17	12/17	+1.61	0.049	transient
	Qwen2-VL-2B	11	5/11	+0.62	0.259	cleared	17	8/17	+0.00	0.995	cleared
	Qwen2-VL-7B (text recipe)†	11	5/11	-0.19	0.800	cleared	17	9/17	+0.01	0.901	cleared
	PaliGemma 2-3B	11	7/11	+0.32	0.736	cleared	17	10/17	+0.67	0.474	cleared
	LLaVA-1.6-7B (lighter budget)	11	5/11	-0.15	0.852	cleared	17	10/17	+0.05	0.948	cleared
	UniAD (non-VLM)	11	6/11	-0.01	0.990	cleared	17	10/17	+0.12	0.980	cleared
	SmolVLM2-500M	inadmissible (72% degenerate frames)					inadmissible (72% degenerate frames)				
Variants	LFM2-VL-450M (same recipe)	11	5/11	-0.74	0.553	cleared	17	12/17	+1.85	0.014	**persistent**
	LFM2.5-VL-1.6B (same recipe)	11	5/11	-0.60	0.295	cleared	17	6/17	-0.20	0.770	cleared
	LFM2.5-VL-3B (same recipe)	11	4/11	-0.73	0.134	cleared	17	10/17	-0.14	0.828	cleared
	LFM2-VL-1.6B replicas ($\times$29 / $\times$2)	11	6.8/11	+0.49	$\geq$0.065	cleared (0/29)	17	10.0/17	+0.89	$\geq$0.090	cleared (0/2)
	Qwen2-VL-2B replicas ($\times$13)	11	6.3/11	+0.37	$\geq$0.116	cleared (0/13)	17	11.1/17	+0.12	$\geq$0.100	cleared (0/13)
	PaliGemma 2-3B replicas ($\times$4)	11	6.5/11	+0.14	$\geq$0.771	cleared (0/4)	17	12.0/17	+0.72	$\geq$0.131	cleared (0/4)
	LLaVA-1.6-7B replicas ($\times$9)	11	4.1/11	-1.04	$\geq$0.072	cleared (0/9)	17	8.1/17	+0.02	$\geq$0.124	cleared (0/9)
Controls	Positive control (Qwen2-VL-2B + LoRA)	11	11/11	+2.66	0.002	flagged*	17	17/17	+2.89	0.011	flagged*
	Human reference (logged drivers)	11	0/11	-2.67	<0.001	—	17	5/17	-0.05	0.902	—

Leakage and output quality. On nuScenes, $8/11$ cue-first audit scenes overlap the training manifest; a separate 32-scene strictly unseen analysis confirms that admissible planners still reproduce the required braking behavior. Argoverse 2 has disjoint training and audit splits by construction. Degenerate outputs are treated as missing rather than zero-speed responses; this removes two nominal flags driven by degenerate outputs, and the admissibility criterion prevents a verdict when output quality is insufficient (SmolVLM2-500M: 72% degenerate frames).

Table 9: **Zero-shot transfer from nuScenes to Argoverse 2.** nuScenes-trained weights are replayed on the Argoverse 2 audit without adaptation. The final column denotes the uncorrected per-model LOCATE flag; the single nominal flag does not survive Bonferroni adjustment over the seven panels ($0.024 \times 7 = 0.17$). Qwen2-VL-2B retains 15 of the 17 scenes; two are dropped for degenerate output.

nuScenes weights	N_Z	n_Z/N_Z	$\bar{L}_Z$	p_Z	nominal flag
LFM2-VL-1.6B (confirmed)	17	10/17	-0.22	0.739	no
Qwen2-VL-2B	15	11/15	+0.75	0.068	no
LLaVA-1.6-7B	17	10/17	+1.49	0.102	no
LFM2-VL-450M	17	11/17	+0.33	0.592	no
LFM2.5-VL-1.6B	17	12/17	+1.19	0.024	yes
LFM2.5-VL-3B	17	11/17	+0.57	0.256	no
LFM2-VL-1.6B replica	17	11/17	+0.43	0.205	no

Zero-shot transfer. Table 9 replays nuScenes-trained weights directly on the Argoverse 2 audit without adaptation. The confirmed nuScenes LFM2-VL-1.6B instance is not flagged after transfer ($10/17$, $\bar{L}_Z = -0.22$, p=0.739). Across the seven transferred planners, one model yields a nominal p<0.05 result (LFM2.5-VL-1.6B, p=0.024), which does not survive Bonferroni adjustment over the seven transferred panels ($0.024 \times 7 = 0.17$). We therefore find no evidence that the confirmed nuScenes dependence persists under zero-shot transfer to Argoverse 2.