Frame-to-Frame Attribution Instability in End-to-End Driving: Prevalence and Predictive Limits

Tian Yu
buc9hh@virginia.edu
University of Virginia
Charlottesville, VA, USA

Sebastian Elbaum
selbaum@virginia.edu
University of Virginia
Charlottesville, VA, USA

Abstract

Post-hoc visual attributions (saliency maps) are widely used to interpret end-to-end driving models, yet their behavior over time remains poorly understood. A natural assumption is that stable driving should produce stable explanations—if a vehicle proceeds smoothly, its attribution maps should not flicker erratically. We investigate this assumption by studying *frame-to-frame attribution instability* (FAI): large changes in attribution maps between consecutive frames despite minimal changes in the visual input and control output. Through experiments in the CARLA simulator, we find that FAI is surprisingly common during normal, collision-free driving and provides little predictive signal for upcoming failures. We also show that training more capable driving policies does not yield more stable attributions. Together, these findings reveal a *stability–safety disconnect*: post-hoc attributions can vary substantially without affecting driving performance, and their instability does not reliably indicate impending errors. This suggests that unregularized saliency methods are insufficient as runtime safety monitors for autonomous driving.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Neural networks**; • **Computer systems organization** → *Robotic autonomy*; • **Software and its engineering** → *Empirical software validation.*

Keywords

Autonomous Driving, Explainable AI, Saliency Maps, Attribution Stability, Safety Assurance

1 Introduction

End-to-end driving models map raw sensor inputs directly to control actions, achieving strong performance but offering limited insight into their decision-making. To interpret these black-box policies, researchers commonly use post-hoc visual attributions (saliency maps) as proxies for what the model "sees." A natural expectation is that stable driving should produce stable explanations: if a vehicle proceeds smoothly along a road, its attribution map should remain focused on consistent, task-relevant features rather than flickering erratically.

This expectation, however, is not guaranteed. In software testing, *coincidental correctness* occurs when a program produces the right output despite following a flawed internal path—the fault simply does not propagate to an observable failure. The same decoupling can occur in driving models: a vehicle may maintain smooth, collision-free control even while its internal decision evidence shifts erratically between redundant or spurious features. Behavioral success does not imply explanatory consistency.

To detect this hidden volatility, we must examine how attributions evolve over time. Yet despite driving being an inherently temporal task, most explainability studies evaluate single frames in isolation. This leaves a fundamental question unanswered: when consecutive frames and control outputs are nearly identical, should their attributions also be consistent? And if not, does attribution instability signal elevated safety risk?

We investigate these questions by studying **Frame-to-Frame Attribution Instability (FAI)**—large changes in attribution maps between consecutive timesteps despite minimal changes in visual input and control output. Through experiments in the CARLA simulator, we find that FAI is surprisingly common during normal, safe driving and carries little predictive signal for upcoming failures. Our results reveal a *stability–safety disconnect*: attribution instability is often benign, and behavioral correctness does not imply stable explanations. This challenges the use of unregularized saliency methods as runtime safety monitors and motivates more principled approaches to interpreting attribution dynamics in autonomous systems.

This report makes the following contributions:

(1) **Formalization of FAI:** We provide an operational definition for measuring frame-to-frame attribution instability under constraints of small input and control deltas.
(2) **Empirical Analysis of Monitoring Utility:** We quantify the prevalence of FAI in standard driving scenarios and show that attribution instability is frequently benign and weakly predictive of failures.
(3) **Capability–Interpretability Decoupling:** We demonstrate that improving driving capability (via increased training data or augmentation) does not consistently yield more temporally stable attributions.

2 Related Work

2.1 Foundations of Post-hoc Attribution

To understand the decision-making of black-box models, researchers rely on post-hoc attribution methods that assign importance scores to input features. Early gradient-based approaches, such as Saliency Maps [7] and Integrated Gradients [9], utilize backpropagation to estimate feature sensitivity. However, these methods can suffer from gradient saturation and noise. To provide a unified measure of feature importance, Lundberg and Lee introduced SHAP (SHapley Additive exPlanations) [3], which leverages game-theoretic Shapley values to guarantee axiomatic properties like local accuracy and consistency. In high-dimensional deep learning contexts, approximations like **GradientSHAP** are commonly used to estimate these values efficiently. While these methods are mathematically

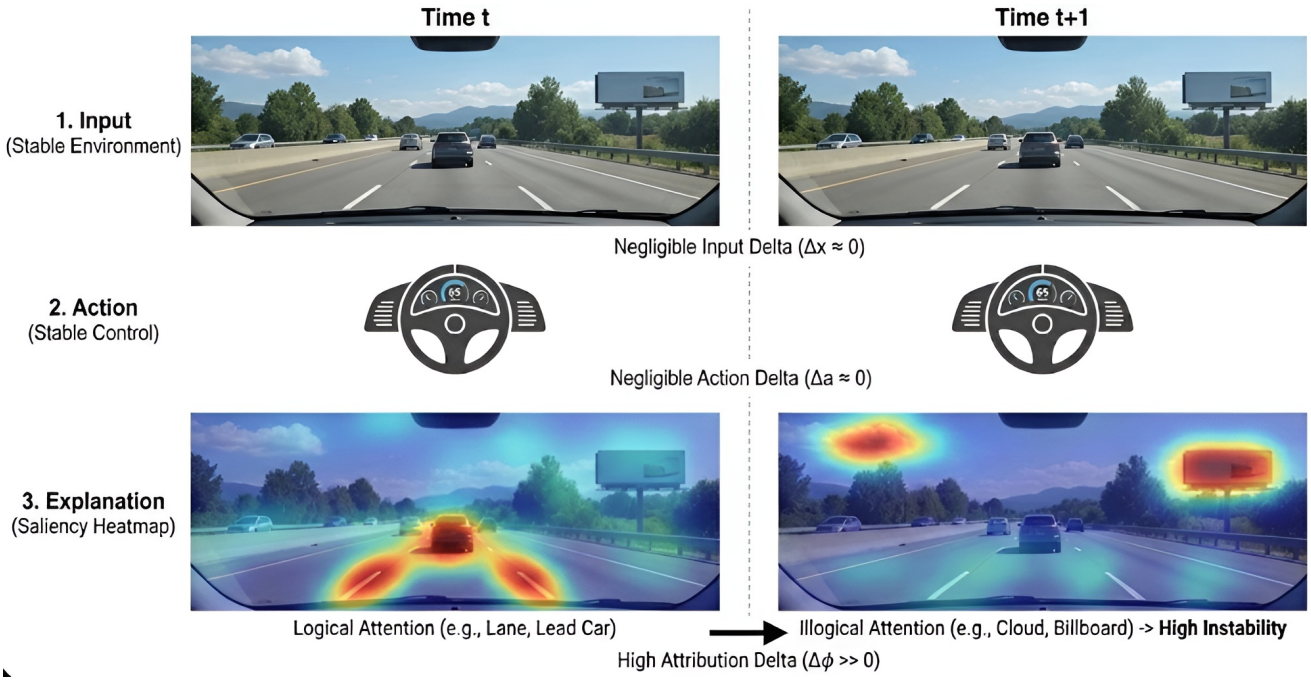

Figure 1: Frame-to-Frame Attribution Instability (FAI). Despite negligible changes in visual input ($\Delta x \approx 0$) and control action ($\Delta a \approx 0$) between consecutive timesteps, the post-hoc attribution map shifts drastically ($\Delta \phi \gg 0$) from task-relevant features (lane, lead car) to irrelevant background artifacts (clouds, billboard).

grounded, their application to safety-critical time-series data remains challenging, as independent per-frame estimates may lack the temporal coherence expected of a physical agent.

2.2 Visual Attribution in End-to-End Driving

In autonomous driving, attribution maps are widely used to qualitatively validate that end-to-end policies rely on task-relevant features. While pixel-level saliency provides a high-fidelity view of model sensitivity, recent work argues it lacks semantic meaning and human persuasibility. To address this, object-centric methods have been proposed, utilizing object masking [13] or objectification branches [12] to align explanations with human concepts. Zhang et al. demonstrated that such object-level abstractions generally offer a better balance of fidelity and user trust [11].

Moving beyond localization ("where"), parallel research has focused on semantic attribution ("what"). Shi et al. introduced methods leveraging traffic scene-informed hidden features [5] and semantic-informed Aumann-Shapley values [6]. Similarly, Sun et al. utilized Shapley-based counterfactuals to identify minimal semantic changes required to alter control decisions [8]. However, these approaches predominantly evaluate attribution quality on static, singular frames. Our work diverges by explicitly auditing the *temporal dynamics* of these attributions during continuous control.

2.3 Explanation Stability and the Oracle Problem

A core motivation for attribution monitoring is the detection of *coincidental correctness*—scenarios where a model produces the correct output (e.g., smooth steering) despite flawed or spurious internal logic. In software testing, this is a known oracle issue where observable failures are masked [10]. In machine learning, this corresponds to the "Right for the Right Reasons" objective [4], where models are penalized for relying on confounding distinct from the causal mechanism.

To serve as a reliable oracle, explanations must be stable. The DRIVE framework identifies "stable interpretability" as a core requirement, proposing ensemble methods to mitigate jitter [1], while Li et al. introduced the Shapley Value-Based Attribution Prior (SVAP) to regularize models against high-frequency sensitivity [2]. Yet, existing stability analyses are largely confined to *static robustness*—consistency under artificial perturbations of a single input. Our work addresses the gap between static robustness and *temporal stability*, evaluating whether frame-to-frame attribution consistency serves as a valid runtime proxy for safety in closed-loop settings.

3 Methodology

To investigate the prevalence and predictive utility of **Frame-to-Frame Attribution Instability (FAI)**, we first formalize the end-to-end driving setting and our explanation extraction pipeline. We then operationalize FAI by defining *Frame-to-Frame Attribution Stability (AS)*—a quantitative metric measuring the structural consistency of post-hoc attributions across consecutive timesteps. Low AS values serves as our indicator for high FAI. Finally, we describe our procedure for correlating these instability events with safety-critical anomalies in closed-loop simulation.

3.1 Preliminaries and Saliency Extraction

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote an end-to-end driving policy parameterized by θ. At time step t, the model maps a high-dimensional sensor input $x_t \in \mathbb{R}^{H\times W\times C}$ to a control action $y_t \in \mathbb{R}^k$ (e.g., steering angle, throttle).

To interpret the policy's decision evidence, we employ **SHAP (Shapley Additive Explanations)** to generate post-hoc visual attributions. We select SHAP for its additivity axiom, which ensures that feature attributions sum to the model output difference relative to a background baseline. For a given input x_t, the explanation function g produces an attribution map $S_t = g(f_\theta, x_t)$, where $S_t \in \mathbb{R}^{H\times W}$. Each pixel value in S_t represents the estimated contribution of that input feature to the predicted control output y_t. We specifically use **GradientSHAP** to approximate Shapley values efficiently for deep convolutional architectures.

3.2 Frame-to-Frame Attribution Stability (AS)

The core hypothesis of this work is that reliable driving policies should exhibit temporally stable post-hoc attributions when the driving context is stable. We quantify *Frame-to-Frame Attribution Stability (AS)* as the similarity between explanations of consecutive frames, evaluating it under constraints where the input and control output change minimally.

Attribution stability. We use the **Structural Similarity Index Measure (SSIM)** to quantify the structural similarity between two attribution maps. For consecutive timesteps, we define:

$$\mathrm{AS}(t) \triangleq \mathrm{SSIM}(S_t, S_{t+1}), \tag{1}$$

where S_t and S_{t+1} are normalized to $[0, 1]$. Larger values indicate greater attribution stability.

Stable-regime constraints. To operationalize a "stable driving context," we restrict our analysis to frame pairs that satisfy both input and action consistency:

- **Stable action (lateral):** Since longitudinal speed is regulated by a fixed-target PID controller in our setting, the learned policy is responsible solely for lateral control. We therefore measure action stability using the normalized steering difference (mapping the action range $[-1, 1]$ to $[0, 1]$):

$$\Delta y_t \triangleq \frac{|y_{t+1}^{\mathrm{steer}} - y_t^{\mathrm{steer}}|}{2}, \qquad \mathrm{SteerSim}(t) \triangleq 1 - \Delta y_t, \tag{2}$$

  and require $\mathrm{SteerSim}(t) \geq \delta$.

Low-stability (high-instability) events. A timestep t is classified as a **low-stability event** if it satisfies:

$$\underbrace{\mathrm{ImageSSIM}(x_t, x_{t+1}) \geq \alpha}_{\text{Stable Input}} \wedge \underbrace{\mathrm{SteerSim}(t) \geq \delta}_{\text{Stable Control}} \wedge \underbrace{\mathrm{AS}(t) \leq \tau_{\mathrm{AS}}}_{\text{Low Attribution Stability}} . \tag{3}$$

Equivalently, this constitutes a high-instability event where $\mathrm{FAI}(t) \geq 1 - \tau_{\mathrm{AS}}$.

3.3 Correlating Attribution Stability with Agent Liveness

To assess the safety implications of temporal attribution behavior, we utilize a **liveness metric** based on velocity. In free-flow driving tasks, an unintended cessation of movement represents a critical failure mode.

We define a **failure event** E_t as:

$$E_t = \mathbb{I}(v_t < 0.1 \text{ km/h}), \tag{4}$$

where $\mathbb{I}$ is the indicator function. Since the PID controller attempts to maintain a constant target velocity, E_t serves as a proxy for:

- **Collisions:** Impact forcing the vehicle to a halt.
- **Entrapment:** A steering failure where the vehicle becomes stuck (e.g., against geometry) and cannot proceed despite throttle application.

We treat failure as a non-recoverable state; once $E_t = 1$, the episode is considered failed.

We hypothesize that **low attribution stability** (small $\mathrm{AS}(t)$, or high FAI) precedes these events. We quantify this relationship via time-lagged analysis, measuring the conditional probability of a failure occurring within a future window.

Stable-regime event. Let A_t denote the event that the driving context is locally stable:

$$A_t \triangleq [\mathrm{ImageSSIM}(x_t, x_{t+1}) \geq \alpha] \ \wedge \ [\mathrm{SteerSim}(t) \geq \delta] . \tag{5}$$

Low-stability event. Let L_t denote a low-attribution-stability event within this regime:

$$L_t \triangleq A_t \ \wedge \ [\mathrm{AS}(t) \leq \tau_{\mathrm{AS}}] . \tag{6}$$

Predictive test. Finally, we evaluate whether low stability within the stable regime predicts a failure at a future horizon k by computing

$$\mathbb{P}(E_{t+k} = 1 \mid L_t). \tag{7}$$

4 Experiments and Analysis

4.1 Experimental Setup

To evaluate the relationship between driving capability and temporal explanation stability, we generated a dataset of 110k frames in CARLA using a rule-based expert agent driving at a fixed velocity of 20 km/h across multiple towns. To ensure the driving task remained within the scope of a standard lane-following baseline, we disabled NPC traffic and set all traffic lights to green. We run CARLA in synchronous mode at 10 FPS, so each timestep corresponds to 0.1s; throughout, lookahead horizons of K steps translate to $K/10$ seconds. We trained a baseline DAVE-2 end-to-end driving policy on this dataset for 50 epochs using the AdamW optimizer ($lr = 10^{-3}$, batch size 256) to minimize Mean Squared Error (MSE) on 320×180 images. At inference time, the DAVE-2 model provides lateral control (steering), while a separate PID controller provides longitudinal control to maintain the target velocity of 20 km/h.

To extract post-hoc attributions, we generate explanation maps using **GradientSHAP**. We construct the SHAP background distribution via K-Means clustering ($K = 64$) over up to 5,000 frames uniformly sampled from the evaluation sequence. This batch-mode baseline constitutes a threat to external validity because it leverages

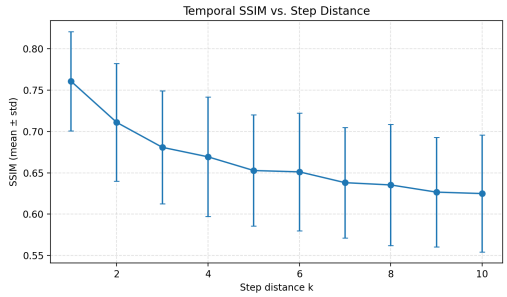

Figure 2: Temporal Decay of Attribution Similarity. Average Attribution SSIM (mean ± std) as a function of time-step distance k. Data was collected using a baseline DAVE-2 model driving on Town01 over 1,500 steps (150s). The monotonic decrease confirms that SSIM effectively captures the temporal divergence of attribution maps as the driving context evolves.

distributional information that would be unavailable to a strict runtime monitor. We adopt this idealized baseline to reduce attribution variance due to baseline mismatch and environmental distribution shift, so that the measured frame-to-frame attribution instability more directly reflects model/explanation dynamics rather than artifacts of baseline selection. We compute SSIM on grayscale saliency magnitude maps

$$M_t(h, w) \triangleq \sum_{c=1}^{C} |S_t(c, h, w)|, \tag{8}$$

where $S_t \in \mathbb{R}^{C\times H\times W}$ denotes the per-channel GradientSHAP attribution at time t, and $M_t \in \mathbb{R}^{H\times W}$ aggregates attribution strength by summing absolute contributions across channels to obtain a single nonnegative saliency intensity map. SSIM is computed on the saved grayscale M_t images after per-frame robust normalization using the 99th-percentile magnitude (for visualization).

4.2 Preliminary: Validation of the Similarity Metric

Before analyzing the relationship between control stability and attribution stability, we validated the sensitivity of our similarity metric (SSIM) to temporal changes. Specifically, we computed the mean **Attribution SSIM** between explanation maps separated by increasing time lags k (i.e., comparing S_t to S_{t+k}). This step ensures that the metric captures meaningful structural divergence rather than being dominated by noise or normalization artifacts.

Figure 2 shows a clear monotonic decay in similarity as the temporal distance increases, confirming that SSIM is sensitive to meaningful temporal changes in attribution structure.

4.3 Result 1: Input-Dominance of Frame-to-Frame Attribution Instability

We first establish the baseline relationship between scene dynamics and explanation stability using the **open-loop training dataset** (~110k frames). **Analyzing the training distribution allows us to isolate intrinsic attribution instability from generalization errors that might occur on unseen data.** For each adjacent frame pair, we compute Image SSIM (input similarity) and Attribution SSIM (explanation similarity). We interpret *low* Attribution SSIM as *high* frame-to-frame attribution instability (FAI), and ask whether observed instability is primarily a byproduct of visual motion (e.g., ego-motion/optical flow) or persists even when the input changes minimally.

4.3.1 Scene Motion Explains the Trend, Not the Dispersion. Figure 3 shows a **moderate positive correlation** between Image SSIM and Attribution SSIM ($r = 0.528$). This indicates that attribution maps tend to become more similar as the underlying frames become more similar, consistent with an *input-dominance* effect: a nontrivial fraction of attribution variation is aligned with natural scene evolution.

However, the plot also exhibits pronounced **conditional dispersion**. In the high-similarity regime (Image SSIM ≥ 0.9), Attribution SSIM spans a wide range (approximately 0.3 to 0.9), rather than concentrating near 1.0. In other words, even when two frames are nearly identical, the corresponding attributions can be either highly consistent or substantially different. This vertical spread is the empirical signature of FAI beyond what is explained by scene motion alone.

A second notable pattern is **heteroscedasticity**: the range of Attribution SSIM values expands as Image SSIM increases. When Image SSIM is low, Attribution SSIM is also typically low and tightly bounded; when Image SSIM is high, Attribution SSIM varies widely. This wedge-shaped geometry suggests that input similarity is a necessary but insufficient condition for attribution stability, leaving substantial variance attributable to factors other than raw pixel-level change (e.g., feature redundancy, sensitivity of the explainer, or small semantic changes that are poorly captured by Image SSIM).

4.3.2 Stable Control Does Not Imply Stable Attributions. Orange points in Figure 3 highlight the top 10% most control-consistent pairs (highest `steer_sim`). If action invariance constrained explanation variability, these points would concentrate into a tight cluster at high Attribution SSIM. Instead, they largely **overlap** the full distribution and retain substantial vertical dispersion, including within the Image SSIM ≥ 0.9 region. This indicates that holding the steering command nearly constant does not, by itself, enforce stable post-hoc decision evidence.

Red triangles mark representative bottom-right outliers (top 20% high Image SSIM, bottom 20% low Attribution SSIM), which correspond to high-FAI events under near-identical inputs. Although these outliers are not the dominant mass of the distribution, their presence demonstrates that large attribution changes can occur even when both the input and the control signal are effectively stable. Overall, the figure supports a one-way dependence: input dynamics explain part of attribution dynamics, but neither stable inputs nor stable control are sufficient to guarantee stable explanations.

4.4 Result 2: The Stability–Safety Disconnect

We extend the analysis to closed-loop control to test whether temporal explanation instability provides predictive signal for driving failures. We evaluate the baseline DAVE-2 policy in active CARLA

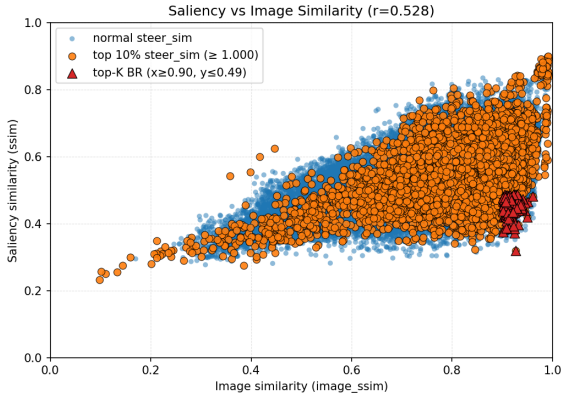

Figure 3: Attribution stability is primarily driven by input dynamics. Scatter plot of Image SSIM vs. Attribution SSIM for consecutive frames in the **open-loop training set ($r = 0.528$). Blue points show all frame pairs; orange points highlight the top 10% most control-consistent pairs (highest `steer_sim`). Although Attribution SSIM increases on average with Image SSIM, substantial dispersion persists in the high-similarity regime (Image SSIM ≥ 0.9): even under near-identical inputs and stable control, explanations can vary sharply. Red triangles mark representative bottom-right outliers (top 20% high Image SSIM, bottom 20% low Attribution SSIM) illustrating frame-to-frame attribution instability (FAI).**

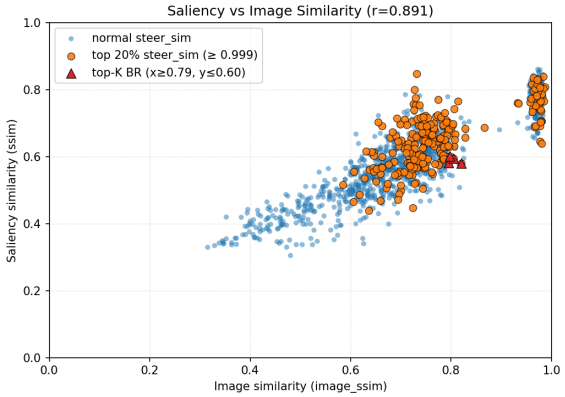

Figure 4: Input–Attribution Coupling in Closed-Loop Control. Scatter plot of Image SSIM vs. Attribution SSIM for a 1,200-step evaluation sequence on Town10HD. The dense cluster near $(1.0, > 0.9)$ corresponds to stationary segments (post-impact), where the vehicle is not moving and both inputs and attributions become trivially stable. Excluding the stationary segments, we observe a strong positive correlation ($r \approx 0.802$): attribution similarity decreases as input similarity decreases.

simulation. To isolate FAI from deliberate control maneuvers, we restrict our analysis to the **stable regime**: frame pairs exhibiting high input similarity (top 30%) and high control consistency (top 20%).

Within this stable regime, we identify **high-FAI events** (outliers with low Attribution SSIM, bottom 30%). We relax the strict quantile cutoffs used in the open-loop analysis (Result 1) to ensure a sufficient sample size of instability events for robust failure correlation, as closed-loop rollouts naturally exhibit higher variance in ego-motion than the expert training data.

4.4.1 Stationary (Post-Impact) Artifact. In the raw closed-loop scatter, we observe a dense cluster of points with near-perfect similarity (Image SSIM ≈ 1.0, Attribution SSIM > 0.9) in the top-right region of Figure 4. Qualitative inspection shows these correspond to **stationary segments**, all after a collision or when the vehicle becomes stuck. In this regime, the camera view changes minimally from frame to frame, and both the input and attribution maps trivially stabilize. Because these points reflect a degenerate "no-motion" condition rather than informative stability during active driving, we exclude stationary frames from subsequent safety analyses.

4.4.2 Coupling Between Input Similarity and Attribution Similarity. After filtering out these stationary artifacts, we focus exclusively on active driving timesteps. Figure 4 reveals a strong positive association between Image SSIM and Attribution SSIM ($r \approx 0.802$) in this regime. Compared to the open-loop setting, the coupling is notably tighter: attribution stability is largely dominated by input similarity. Consequently, high-FAI events (high Image SSIM but low Attribution SSIM) appear as rare outliers rather than a structural feature of the distribution.This suggests that, during active closed-loop operation, attribution stability provides limited incremental signal beyond scene dynamics: when the visual input changes (due to ego-motion or scene evolution), attribution maps tends to degrade at a commensurate rate.

4.4.3 Temporal Dynamics and Limited Predictive Signal. To further examine whether attribution instability anticipates failures, we study the temporal evolution of the metrics during active driving. Figure 5 plots Image SSIM, Attribution SSIM, and steering similarity over a representative 1,200-step segment. Attribution SSIM closely tracks Image SSIM: drops in input similarity are consistently accompanied by drops in attribution similarity. Meanwhile, steering similarity remains high across long intervals yet does not prevent attribution fluctuations. This supports the conclusion that attribution dynamics are primarily coupled to the sensor stream rather than being governed by the smoothness of the control signal.

To connect the time-series view back to the scatter analysis, the **red shaded regions** in Figure 5 correspond to the same class of bottom-right outliers highlighted as **red triangles** in Figure 4: transitions with relatively high input similarity and near-identical steering (high Image SSIM and high `steer_sim`) but unexpectedly low attribution similarity. These segments isolate candidate "high-FAI" events under an approximately stable input/action regime, enabling direct inspection of whether such spikes precede failures.

Visual inspection of these outliers reinforces our findings. Figure 6 exemplifies one such event, where attribution similarity drops substantially (Attribution SSIM ≈ 0.58) despite relatively high input similarity (Image SSIM ≈ 0.82) and near-identical steering (Steer Sim > 0.99). Importantly, this instability does not precede a

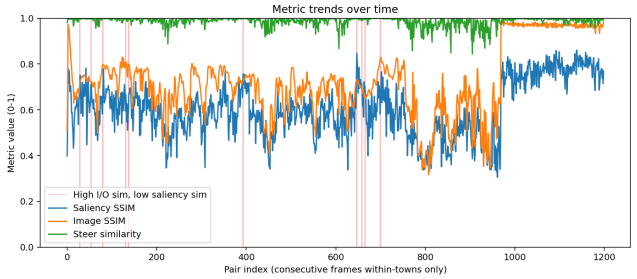

Figure 5: Temporal Dynamics of Closed-Loop Stability. A representative 1,200-step sequence showing Image SSIM (orange), Attribution SSIM (blue), and Steering Similarity (green). Attribution SSIM largely tracks Image SSIM: as the visual scene evolves, attribution similarity changes at a similar rate. Periods of high steering similarity do not necessarily correspond to stable attributions. Red shaded regions highlight candidate instability spikes: transitions with *high input/output similarity* (high Image SSIM and high SteerSim; i.e., stable scene and nearly unchanged steering) but *low attribution similarity* (low Attribution SSIM), indicating sharp changes in decision evidence despite minimal perceptual and control change.

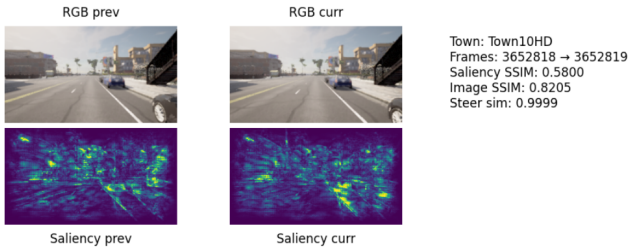

Figure 6: Large Attribution Changes Under Stable Input and Control. A qualitative outlier from the closed-loop evaluation. Despite high input similarity (Image SSIM ≈ 0.8205) and near-identical steering (Steer Sim ≈ 0.9999), the attribution map changes substantially (Attribution SSIM ≈ 0.5800). The vehicle continues safely through the road, illustrating that high frame-to-frame attribution instability (FAI) is not necessarily a precursor to failure.

collision or infraction; the vehicle successfully traverses the subsequent junction. This example illustrates that large frame-to-frame attribution changes can be benign, likely reflecting redundancy in the visual evidence supporting the action rather than an impending loss of control.

4.4.4 Instantiating the Predictability Test in the Time-Series Example. Equation (7) makes the notion of "predictive signal" explicit by evaluating the conditional event rate $\mathbb{P}(E_{t+k} = 1 \mid L_t)$. Here, L_t (Eq. (6)) identifies timesteps that fall within a locally stable driving regime A_t (Eq. (5)) yet exhibit low attribution stability ($\mathrm{AS}(t) \leq \tau_{\mathrm{AS}}$)—the high-FAI events.

In Figure 5, the **red shaded regions** visually instantiate L_t: each region marks a transition with high input/output similarity (satisfying A_t) but low attribution similarity. Under Eq. (7), the onset of each shaded region defines a binary trial: we ask whether the failure indicator $E_{t+k} = 1$ is active at horizon k. Given the assumption that failure is non-recoverable, this is equivalent to checking if a failure occurred within the first k steps.

To quantify the predictive utility, we track the *first-hit time* $k^*(t) = \min\{k \geq 1 : E_{t+k} = 1\}$ following an instability event. In the depicted rollout, the nearest downstream failure occurs at $k^* \approx 300$ steps (30s) after the highlighted instability. Consequently, for operational monitoring horizons (e.g., $K = 10 \ldots 50$ steps, corresponding to 1–5 seconds), these high-FAI spikes register as **false positives** (non-events). They would only be considered predictive if the horizon K were expanded to tens of seconds, at which point the causal link becomes tenuous due to intervening scene changes.

This example illustrates the practical manifestation of the stability–safety disconnect: even after conditioning on stable input/control (A_t), sharp attribution shifts are not reliably followed by near-term failures. The lack of imminent consequences suggests that attribution instability, in isolation, offers limited actionable predictive lift beyond standard scene dynamics.

Overall, these results indicate a fundamental decoupling: while attributions are frequently unstable, this variability is largely benign or explained by input flux, failing to serve as a reliable precursor for safety-critical failures under our liveness and collision criteria.

4.5 Result 3: Capability–Interpretability Decoupling

Hypothesis of Competence. A natural counter-argument to our findings is that attribution instability might simply be a symptom of model incompetence—a confused policy yielding flickering evidence. If this were true, one would expect attribution stability to improve monotonically as the driving policy becomes more capable.

To test this, we isolate the effect of model competence by training DAVE-2 variants on nested subsets of the training data (12.5%, 25%, 50%, 100%), corresponding to $N \in \{13{,}750,\ 27{,}500,\ 55{,}000,\ 110{,}000\}$ frames.[1] All models are trained using the standard protocol defined in Section 4.1. We compare a baseline training pipeline against two performance-oriented regimes:

- **Augmented:** Standard visual perturbations, including horizontal flipping ($p = 0.5$ with steering sign inversion), ColorJitter (strength=0.2), and additive Gaussian noise ($\sigma = 0.01$).
- **Aug Balanced:** The Augmented pipeline plus a weighted sampler that counteracts the imbalance of high-curvature turns by enforcing 50% of each batch to have steering magnitude $> |0.5|$ (sampling with replacement; maximum oversampling factor 3).

Capability metric. We quantify driving capability using **average liveness duration**, defined as the average time (in seconds) that the ego-vehicle maintains velocity above a stagnation threshold ($v_t > 0.1$ km/h) before failure or episode timeout.

We evaluate each agent across 8 CARLA towns using seeds 0–10, with a maximum episode length of 1,800 steps (180s). Across four

[1]These counts reflect aggregation across all training towns; each subset is a strict superset of the smaller subsets to reduce sampling variability.

Table 1: Impact of Training Regime on Agent Liveness. Average duration (seconds) that the ego-vehicle maintains velocity above the liveness threshold before stagnation.

Training Variant	Training Set Size (N frames)			
	13,750	**27,500**	**55,000**	**110,000**
Baseline	53.9	78.0	69.4	90.1
Augmented	**67.2**	66.2	**107.3**	**146.0**
Aug Balanced	65.9	**95.2**	102.0	100.4

dataset sizes and three training variants, augmentation consistently improves driving capability, achieving the best liveness in three of the four dataset sizes (Table 1). Based on these results, we use the **Augmented** regime as the representative high-performance model family for the subsequent stability analysis.

4.5.1 Orthogonality of Capability and Attribution Stability. To test whether improved driving capability is accompanied by more temporally stable explanations, we compare models trained under identical augmentation but with increasing training-set size ($N \in \{13{,}750,\ 27{,}500,\ 55{,}000,\ 110{,}000\}$ frames). We evaluate each model in the same closed-loop Town10HD setting and quantify temporal attribution stability using **Attribution SSIM** between consecutive attribution maps.

Stability evaluation under a matched regime. A central confound in temporal explainability is that attribution maps naturally change as the visual scene changes. To reduce the influence of scene dynamics and isolate intrinsic instability, we restrict our analysis to frame pairs that satisfy the **stable-regime constraints** defined in Section 3 (i.e., high input similarity and high control consistency).

We further restrict analysis to the **first 30 seconds** of each rollout to eliminate post-impact stationary segments and late-horizon route divergence effects that can introduce degenerate "no-motion" stability. Under these constraints, we compute the mean and median Attribution SSIM over the remaining eligible frame pairs for each model.

Observed trend. Table 2 reports stability summaries for $\alpha = 0.6$ and $\delta = 0.999$. We selected a strict steering threshold (δ) to ensure control invariance, while the moderate visual threshold (α) balances the need for scene consistency against the requirement to retain sufficient sample sizes across varying policy behaviors. Attribution stability does not improve monotonically with training-set size; in fact, we observe the opposite trend at the extremes. The smallest-data model ($N = 13{,}750$) exhibits the highest median Attribution SSIM (0.823), while the largest-data model ($N = 110{,}000$)—which achieved the highest liveness capability—exhibits the lowest stability (0.616). Overall, despite capability gains from more data and augmentation (Table 1), temporal attribution stability remains variable and fails to track capability improvements. This supports a strong **decoupling** between behavioral performance and post-hoc explanation stability: a model can become more capable in closed-loop control while producing significantly less stable attributions under matched input/action regimes.

Table 2: Attribution stability across training set sizes. Mean/median Attribution SSIM over the first 30 seconds of Town10HD, restricted to pairs with Image SSIM ≥ 0.6 and SteerSim ≥ 0.999. N_{sel} is the number of eligible frame pairs after filtering.

Training size (N)	N_{sel}	Mean	Median
13,750	106	0.809	0.823
27,500	165	0.704	0.722
55,000	111	0.771	0.786
110,000	143	0.609	0.616

Limitations. This analysis is narrow and should be interpreted as indicative rather than definitive. First, it considers a single town and a short evaluation window; stability may vary substantially across routes, towns, weather, and seeds. Second, strict thresholded filtering (Image SSIM ≥ 0.6, SteerSim ≥ 0.999) **prioritizes experimental control over sample size**; varying α or δ changes both the count and distribution of eligible pairs, and these counts differ across models due to policy-dependent trajectory divergences. Third, summarizing stability via mean/median Attribution SSIM may obscure tail behavior that is more safety-relevant (e.g., the frequency of extreme instability events). Extending this comparison across multiple towns and seeds, utilizing matched input-similarity strata and tail-focused metrics, is a key next step.

4.5.2 Predictive signal check across trained agents. As a final check, we evaluated whether the high-FAI pattern identified in Result 2 transfers to the agents trained in Result 3. We replicate the closed-loop evaluation setting on Town10HD (one rollout per model) and mark candidate "high-FAI" events using the same definition as Figure 4: frame pairs with top 20% steering similarity (`steer_sim`), top 30% Image SSIM, and bottom 30% Attribution SSIM.

Within this evaluation, the agents trained on 12.5% and 50.0% of the data crashed. Notably, in the segments preceding these crashes, *no* timesteps satisfied the high-FAI criteria—a false negative result where instability failed to anticipate the error. In contrast, the 25% and 100% models completed the rollout safely, yet exhibited multiple instances of high Image/Steering similarity accompanied by low attribution similarity—a false positive result.

Although limited in scope, this comparison reinforces the stability–safety disconnect suggested by Result 2: the presence of high-FAI events is not a reliable precursor to failure, and their absence does not guarantee safety.

5 Future Work

This work provides an initial characterization of frame-to-frame attribution behavior in end-to-end driving. To generalize these findings, a primary next step is to expand coverage across *more diverse driving contexts*—additional towns/routes, weather and illumination, traffic density, and longer rollouts—with substantially more seeds and repeated runs per model. This will enable stronger quantitative analysis beyond descriptive plots, including stratified evaluations under matched input-similarity regimes, tail-focused statistics (e.g., rate of low-AS / high-FAI events), and predictive tests

Figure 7: Shadow-induced saliency shift (motivating concept). *Concept illustration* depicting how a physically plausible illumination change can induce a substantial redistribution of attribution mass (yellow) despite minimal change in overall scene semantics. This motivates future work in targeted physical/adversarial perturbations that preserve near-invariant control while maximizing attribution instability under perceptually small, semantically meaningful changes.

(e.g., AUC/lift) relative to baselines that use only input similarity and control smoothness.

A second direction is to broaden the measurement toolkit. SSIM is a convenient structural metric but does not fully capture perceptual or semantic changes in driving scenes, especially under lighting effects. We will therefore evaluate additional similarity metrics such as FSIM and learned perceptual distances (LPIPS), and replicate the stability analyses across multiple attribution methods beyond GradientSHAP (e.g., Grad-CAM/CAM-family variants and gradient-based saliency) to determine whether the observed stability–safety disconnect is robust to both metric and explainer choice.

Finally, naturally occurring frame pairs may under-sample the most informative regime: *semantically subtle yet attribution-disruptive* changes. We will move from passive observation to *targeted adversarial physical generation* of controlled perturbations that optimize a multi-objective criterion: (i) minimal perceptual change (constrained under SSIM/FSIM/LPIPS), (ii) semantic plausibility (physically realizable changes), (iii) small output deviation (near-invariant control), and (iv) large attribution instability. Figure 7 illustrates the motivation: a small change in cast *shadow geometry* induces a pronounced shift in attribution focus, despite the scene remaining largely unchanged. Targeted generation of such cases would better ground the study in semantically meaningful perturbations and provide a systematic way to probe when and why saliency shifts occur.

6 Conclusion

We investigated the viability of post-hoc visual attributions as runtime safety monitors for end-to-end driving. Our analysis reveals a fundamental disconnect between explanation stability and behavioral safety: high attribution instability frequently occurs during benign driving, while actual failures often lack distinctive attribution shifts. Furthermore, we find that increasing model capability does not guarantee improved attribution stability, suggesting a decoupling between performance and interpretability. These results demonstrate that unregularized post-hoc attributions are currently insufficient for direct safety assurance. Future work must bridge this "oracle gap" through rigorous quantitative benchmarks and the development of stability-regularized explanation methods.

References

[1] Songning Lai, Tianlang Xue, Hongru Xiao, Lijie Hu, Jiemin Wu, Ninghui Feng, Runwei Guan, Haicheng Liao, Zhenning Li, and Yutao Yue. 2024. DRIVE: Dependable Robust Interpretable Visionary Ensemble Framework in Autonomous Driving. *arXiv preprint arXiv:2409.10330* (2024). https://arxiv.org/abs/2409.10330

[2] Meng Li, Hengyang Sun, Zhihao Cui, Yanjun Huang, and Hong Chen. 2025. SVAP: Shapley Value Guided Attribution Prior for Neural Network-Based Autonomous Driving. *IEEE Transactions on Vehicular Technology* (2025). doi:10.1109/TVT.2025.3575760 Early Access.

[3] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. 4765–4774.

[4] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. 2662–2670. doi:10.24963/ijcai.2017/371

[5] Rui Shi, Tianxing Li, Yasushi Yamaguchi, and Liguo Zhang. 2025. Traffic Scene-Informed Attribution of Autonomous Driving Decisions. *IEEE Transactions on Intelligent Transportation Systems* 26, 7 (2025), 9175–9186. doi:10.1109/TITS.2025.3547879

[6] Rui Shi, Tianxing Li, Yasushi Yamaguchi, and Liguo Zhang. 2025. Understanding Decision-Making of Autonomous Driving via Semantic Attribution. *IEEE Transactions on Intelligent Transportation Systems* 26, 1 (2025), 283–294. doi:10.1109/TITS.2024.3483810

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[8] Hengyang Sun, Meng Li, Zhihao Cui, Yanjun Huang, and Hong Chen. 2025. Semantic Shapley-Based Counterfactual Explanations for End-to-End Autonomous Driving. *Engineering Applications of Artificial Intelligence* 159 (2025), 111638. doi:10.1016/j.engappai.2025.111638

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, 3319–3328.

[10] Jeffrey M. Voas. 1992. PIE: A dynamic failure-based technique. *IEEE Transactions on Software Engineering* 18, 8 (1992), 717–727. doi:10.1109/32.153381

[11] Chenkai Zhang, Daisuke Deguchi, Jialei Chen, and Hiroshi Murase. 2024. Comprehensive Evaluation of End-to-End Driving Model Explanations for Autonomous Vehicles. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*. SciTePress, 509–518. doi:10.5220/0012365900003660

[12] Chenkai Zhang, Daisuke Deguchi, Jialei Chen, and Hiroshi Murase. 2024. Toward Explainable End-to-End Driving Models via Simplified Objectification Constraints. *IEEE Transactions on Intelligent Transportation Systems* 25, 10 (2024), 14521–14534. doi:10.1109/TITS.2024.3385754

[13] Chenkai Zhang, Daisuke Deguchi, Yuki Okafuji, and Hiroshi Murase. 2023. More Persuasive Explanation Method for End-to-End Driving Models. *IEEE Access* 11 (2023), 4270–4282. doi:10.1109/ACCESS.2023.3235739