

Environment-Conditional Interpretation Consistency: Measuring and Improving LLM Explanation Stability Under Diverse Conditions

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) deployed in safety-critical autonomous driving systems must provide not only correct decisions but also consistent and faithful explanations across diverse environmental conditions. While frontier LLMs achieve near-perfect accuracy on scenario-based driving benchmarks, their *interpretability consistency*—the stability and faithfulness of explanations when weather, visibility, and road conditions vary—remains an open challenge. We formalize this problem through the **Environment-Conditional Interpretation Consistency** (ECIC) framework, which disentangles decision-relevant features from environment-contextual features and measures explanation stability along four complementary axes: Attribution Invariance Score (AIS), Explanation Semantic Similarity (ESS), Faithfulness Gap (FG), and a composite Consistency Index (CI). We evaluate the framework across 10 autonomous driving scenarios under 10 canonical environmental conditions (450 condition pairs), using simulated LLM explanation generators with controllable consistency and faithfulness parameters. Our experiments reveal that: (i) the ECIC-optimized configuration achieves a mean CI of 0.964 compared to 0.936 for the baseline, representing a 93% reduction in faithfulness gap; (ii) phase transition analysis identifies critical visibility and precipitation thresholds below which explanation consistency degrades; and (iii) contrastive explanation anchoring, which decomposes explanations into environment-independent and environment-dependent components, achieves a 100% pass rate on structural consistency checks. The ECIC framework provides a principled evaluation methodology for the open problem of consistent real-world LLM interpretability identified by Ferrag et al. (2026) in the AgentDrive benchmark.

ACM Reference Format:

Anonymous Author(s). 2026. Environment-Conditional Interpretation Consistency: Measuring and Improving LLM Explanation Stability Under Diverse Conditions. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The deployment of large language models (LLMs) in autonomous driving systems represents a convergence of two critical demands: agentic decision-making and human-legible interpretability [6, 17]. Recent benchmarks such as AgentDrive-MCQ demonstrate that frontier LLMs can achieve near-perfect scores on scenario-style reasoning tasks requiring holistic understanding of complex, dynamic driving environments. However, as Ferrag et al. [6] explicitly identify, “achieving consistent real-world interpretability under diverse environmental conditions remains an open research challenge for the broader LLM ecosystem.”

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

This gap between benchmark accuracy and reliable interpretability is not merely academic. In autonomous driving, interpretability serves three operational functions: (1) *regulatory audit*—enabling post-hoc verification that the system’s reasoning was sound; (2) *real-time handoff*—allowing human operators to understand the system’s assessment during safety-critical transfers of control; and (3) *forensic analysis*—supporting incident investigation by providing causal reasoning chains. Each of these functions requires that explanations remain structurally and semantically consistent when the same underlying decision scenario is encountered under different environmental conditions.

The core challenge is that environmental variation—weather, lighting, visibility, road surface—introduces a structured distributional shift that can destabilize LLM explanations even when decisions remain correct. A model that explains a braking decision by citing “pedestrian ahead” in clear weather but shifts its stated rationale to “wet road surface” in rain for an identical pedestrian scenario has broken the *interpretability contract* with human operators, regardless of whether both explanations are individually plausible.

We distinguish this problem from two related but distinct challenges. First, *decision robustness* concerns whether the model makes the same (correct) decision under varying conditions—frontier LLMs already achieve this on existing benchmarks. Second, *explanation faithfulness* concerns whether a given explanation accurately reflects the model’s internal computation [10]—important but typically studied at a fixed operating point. Our problem, *interpretability consistency*, is orthogonal: a model can give faithful explanations that are inconsistent across conditions, or consistent explanations that are unfaithful. The ECIC framework measures both axes simultaneously.

We address this open problem by introducing the **Environment-Conditional Interpretation Consistency** (ECIC) framework, which makes the following contributions:

- (1) **Formal metric suite.** We define four complementary metrics—Attribution Invariance Score (AIS), Explanation Semantic Similarity (ESS), Faithfulness Gap (FG), and Decision Consistency (DC)—unified in a composite Consistency Index (CI) with configurable, safety-aware weighting (§2.2).
- (2) **Phase transition analysis.** We introduce a parametric sweep methodology that identifies critical environmental thresholds where explanation consistency degrades, enabling targeted robustness improvements (§2.4).
- (3) **Contrastive explanation anchoring.** We propose a structural decomposition of explanations into environment-invariant and environment-variant components, enabling principled measurement of consistency while permitting legitimate adaptation (§2.3).

- 117 (4) **Comprehensive evaluation.** We evaluate across 10 driving
 118 scenarios, 10 environmental conditions, and 4 model
 119 configurations, producing 450 pairwise comparisons with
 120 reproducible results (§3).

122 1.1 Related Work

123 **LLM Interpretability and Mechanistic Analysis.** Mechanistic
 124 interpretability aims to identify computational circuits within
 125 transformers that mediate specific behaviors [3, 5]. While powerful,
 126 these approaches require white-box access and do not scale to
 127 production-scale models. Explanation faithfulness—the alignment
 128 between a model’s stated rationale and its actual computation—has
 129 been studied extensively for chain-of-thought reasoning [10, 13],
 130 but primarily under fixed data distributions. The ECIC framework
 131 complements mechanistic approaches by providing black-box con-
 132 sistency metrics that can be applied to any LLM.

133 **Explanation Robustness.** Alvarez-Melis and Jaakkola [2] study
 134 the stability of explanations under input perturbations, defining
 135 local Lipschitz continuity conditions for self-explaining networks.
 136 Agarwal et al. [1] provide a benchmark for evaluating explanation
 137 methods across multiple fidelity axes. However, these approaches
 138 focus on adversarial or random noise rather than semantically
 139 coherent environmental variation. Feature attribution methods in-
 140 cluding SHAP [12] and LIME [14] provide local explanations but
 141 lack built-in consistency guarantees across distribution shifts. Our
 142 Attribution Invariance Score extends this literature to structured,
 143 environment-parameterized perturbations.

144 **Autonomous Driving Benchmarks and World Models.** The
 145 AgentDrive benchmark [6] evaluates LLMs on agentic reasoning
 146 tasks including scenario-style challenges that require holistic en-
 147 vironmental understanding. The paper identifies interpretability
 148 consistency as an open challenge despite near-perfect decision accu-
 149 racy. World model approaches for LLM agents [8, 9] highlight chal-
 150 lenges in non-stationary environments, of which interpretability
 151 consistency is the human-facing manifestation. Our work provides
 152 the evaluation framework that these deployment scenarios require.

153 **Counterfactual and Contrastive Explanations.** Counterfactual
 154 explanation methods [15, 16] answer “what would need to
 155 change for a different outcome?” These are naturally suited to en-
 156 vironmental variation, where the counterfactual *is* the alternate
 157 weather condition. Our contrastive anchoring approach draws on
 158 this tradition but applies it specifically to decomposing explanations
 159 into environment-invariant and environment-variant components,
 160 shifting the focus from decision boundaries to explanation stability.

161 **Gap.** No prior work systematically measures or optimizes for
 162 consistency of LLM interpretability across structured environmen-
 163 tal perturbations in agentic settings. The intersection of explanation
 164 robustness, faithfulness evaluation, and environment-parameterized
 165 distributional shift is genuinely open. The ECIC framework fills
 166 this gap.

168 2 METHODS

169 2.1 Problem Formulation

170 Let $s \in \mathcal{S}$ denote a base driving scenario and $e \in \mathcal{E}$ an environ-
 171 mental condition. Each scenario s has decision-relevant features
 172 $\mathbf{x}_s = \{x_1, \dots, x_d\}$ (e.g., pedestrian position, traffic signal state, ego

175 speed) and a ground-truth action a_s^* . Each condition e is parame-
 176 terized by a continuous vector:

$$\mathbf{c}_e = (v, p, l, f) \in \mathbb{R}^4 \quad (1)$$

177 representing visibility distance ($v \in [10, 1000]$ m), precipitation
 178 intensity ($p \in [0, 1]$), ambient light ($l \in [0, 1]$), and road surface
 179 friction ($f \in [0, 1]$). The environmental severity is:

$$\text{sev}(e) = 1 - \frac{1}{4} \left(\frac{v}{1000} + (1-p) + l + f \right) \quad (2)$$

180 which ranges from 0 (benign) to approximately 1 (extreme).
 181

182 An LLM explanation model M produces, for each (s, e) pair:

- 183 • A decision $f_M(s, e) \in \mathcal{A}$;
- 184 • A structured explanation $g_M(s, e) = (\mathbf{w}, r_{\text{inv}}, r_{\text{dep}})$, where
 185 $\mathbf{w} \in \Delta^{|\mathcal{F}|}$ is a feature attribution vector over features \mathcal{F} ,
 186 r_{inv} is the environment-independent rationale, and r_{dep} is
 187 the environment-dependent adjustment.

188 The features \mathcal{F} are partitioned into *decision-relevant* features $\mathcal{F}_{\mathcal{D}}$
 189 (which determine the correct action regardless of environment) and
 190 *environment-contextual* features $\mathcal{F}_{\mathcal{E}}$ (which modulate perception
 191 but do not change the fundamental decision calculus).
 192

196 2.2 ECIC Metric Suite

197 We define four complementary metrics and one composite index.

198 **Attribution Invariance Score (AIS).** Measures stability of
 199 decision-relevant attributions across conditions using Jensen–Shannon
 200 divergence [11]:
 201

$$\text{AIS}(e_1, e_2 | s) = 1 - \text{JSD}(\mathbf{w}_{\mathcal{D}}(s, e_1) \| \mathbf{w}_{\mathcal{D}}(s, e_2)) \quad (3)$$

202 where $\mathbf{w}_{\mathcal{D}}$ restricts and renormalizes the attribution vector to
 203 decision-relevant features. AIS ranges in $[1 - \ln 2, 1] \approx [0.307, 1]$,
 204 with higher values indicating greater invariance. The JSD is cho-
 205 sen over KL divergence for its symmetry and boundedness, critical
 206 properties for pairwise comparison.
 207

208 **Explanation Semantic Similarity (ESS).** Measures textual
 209 consistency of the environment-independent rationale:
 210

$$\text{ESS}(e_1, e_2 | s) = \text{sim}(r_{\text{inv}}(s, e_1), r_{\text{inv}}(s, e_2)) \quad (4)$$

211 We employ token-level Jaccard similarity as a dependency-free
 212 proxy (sentence embeddings in production). ESS captures structural
 213 explanation consistency at the natural language level, complement-
 214 ing the vector-space AIS metric.
 215

216 **Faithfulness Gap (FG).** Quantifies divergence between stated
 217 and actual feature reliance:
 218

$$\text{FG}(s, e) = 1 - \cos(\mathbf{w}(s, e), \hat{\mathbf{w}}(s, e)) \quad (5)$$

219 where $\hat{\mathbf{w}}$ denotes empirical sensitivities from feature ablation. FG
 220 $\in [0, 2]$ with lower values indicating greater faithfulness. In our
 221 framework, ablation sensitivities are computed by removing each
 222 feature from the input and measuring decision change probability.
 223

224 **Decision Consistency (DC).** Binary indicator:

$$\text{DC}(e_1, e_2 | s) = \mathbb{1}[f_M(s, e_1) = f_M(s, e_2)] \quad (6)$$

225 **Consistency Index (CI).** The composite metric is a weighted
 226 sum:
 227

$$\text{CI} = \alpha \cdot \text{AIS} + \beta \cdot \text{ESS} + \gamma \cdot (1 - \overline{\text{FG}}) + \delta \cdot \text{DC} \quad (7)$$

228 with default weights $\alpha = 0.3, \beta = 0.2, \gamma = 0.3, \delta = 0.2$, reflecting the
 229 primacy of attribution invariance and faithfulness for safety-critical
 230 applications.
 231

applications. The aggregate CI over a set of results can be further weighted by scenario safety criticality $\kappa_s \in [0, 1]$:

$$\overline{CI} = \frac{\sum_{s,e_1,e_2} \kappa_s \cdot CI(e_1, e_2|s)}{\sum_{s,e_1,e_2} \kappa_s} \quad (8)$$

2.3 Contrastive Explanation Anchoring

To improve consistency while permitting legitimate environment-dependent reasoning, we structure explanations into three components:

- (1) **Decision (a):** The selected driving action.
- (2) **Environment-independent rationale (r_{inv}):** Reasoning that should remain stable across environmental conditions (e.g., “braking required due to pedestrian at 25m”).
- (3) **Environment-dependent adjustments (r_{dep}):** Reasoning that legitimately varies with conditions (e.g., “increased stopping distance due to wet surface”).

The contrastive consistency checker then verifies three properties for each scenario-condition pair (s, e_1, e_2) :

(a) Rationale Stability:

$$\text{sim}(r_{\text{inv}}(s, e_1), r_{\text{inv}}(s, e_2)) > \tau_r \quad (9)$$

with threshold $\tau_r = 0.5$.

(b) Adjustment Coherence: If $|v_{e_1} - v_{e_2}| > 100\text{m}$, the adjustment text must reference visibility; if $|p_{e_1} - p_{e_2}| > 0.2$, it must reference precipitation or surface conditions.

(c) Attribution Proportionality:

$$\frac{\|\mathbf{w}(s, e_1) - \mathbf{w}(s, e_2)\|}{d_{\mathcal{E}}(e_1, e_2)} \leq \rho \quad (10)$$

where $d_{\mathcal{E}}$ is the Euclidean distance in normalized condition space and $\rho = 2.0$ is the proportionality tolerance. This ensures that attribution drift does not exceed what the environmental distance warrants.

2.4 Phase Transition Analysis

We sweep individual environmental parameters while holding others at reference values, computing CI at each point along the sweep. A *phase transition* occurs at parameter value θ^* where the local gradient exceeds a threshold:

$$\left| \frac{\partial \text{CI}}{\partial \theta} \right|_{\theta=\theta^*} > \tau_g \quad (11)$$

with $\tau_g = 0.002$ per unit parameter change. Phase transitions identify critical operational boundaries—e.g., visibility distances below which explanation consistency degrades sharply—enabling targeted robustness improvements and operational envelope definition.

Algorithm 1 summarizes the full ECIC evaluation pipeline.

2.5 Experimental Setup

Scenarios. We evaluate 10 autonomous driving scenarios spanning the full range of the AgentDrive taxonomy: pedestrian crossings (PED_CROSS_01, criticality 0.95), intersection navigation (INTERSECT_02, 0.70), highway merging (HWY_MERGE_03, 0.60), emergency response (EMERG_04, 0.90), school zones (SCHOOL_05, 1.00), lane changes (LANE_CHANGE_06, 0.85), construction zones (CONSTRUCTION_07, 0.65), cyclist encounters (CYCLIST_08, 0.90),

Algorithm 1 ECIC Evaluation Pipeline

Require: Scenarios \mathcal{S} , conditions \mathcal{E} , model M
Ensure: Consistency metrics, phase transitions, contrastive checks

```

1: for each  $s \in \mathcal{S}$  do
2:   for each  $e \in \mathcal{E}$  do
3:     Generate explanation  $g_M(s, e)$ 
4:     Compute ablation sensitivities  $\hat{\mathbf{w}}(s, e)$ 
5:   end for
6:   for each pair  $(e_1, e_2) \in \binom{\mathcal{E}}{2}$  do
7:     Compute AIS( $e_1, e_2|s$ ), ESS( $e_1, e_2|s$ )
8:     Compute FG( $s, e_1$ ), FG( $s, e_2$ ), DC( $e_1, e_2|s$ )
9:     Compute CI via Eq. (7)
10:    Run contrastive checks (a), (b), (c)
11:   end for
12: end for
13: Aggregate results with criticality weighting
14: Sweep visibility and precipitation for phase transitions
15: return Metrics, transitions, check results

```

roundabouts (ROUNDABOUT_09, 0.50), and animal detection (ANIMAL_10, 0.95). Safety criticality scores weight the aggregate metrics toward high-stakes scenarios.

Environmental Conditions. We define 10 canonical conditions parameterized by (v, p, l, f) : clear day (1000, 0.0, 1.0, 1.0), overcast (800, 0.0, 0.6, 0.95), light rain (500, 0.3, 0.5, 0.7), heavy rain (200, 0.8, 0.3, 0.4), fog (80, 0.0, 0.4, 0.85), dense fog (30, 0.0, 0.3, 0.8), night clear (300, 0.0, 0.1, 1.0), night rain (150, 0.5, 0.05, 0.5), snow (250, 0.6, 0.5, 0.3), and blizzard (40, 0.9, 0.2, 0.15). This yields $\binom{10}{2} = 45$ unique condition pairs per scenario and 450 total pairwise evaluations.

Model Configurations. We compare four model configurations with progressively lower consistency noise (σ) and faithfulness gap (ϕ):

- *Baseline*: $\sigma = 0.50, \phi = 0.40$ (unoptimized LLM).
- *Contrastive Anchored*: $\sigma = 0.25, \phi = 0.25$ (structured explanation format).
- *ECIC-Optimized*: $\sigma = 0.15, \phi = 0.10$ (consistency-regularized).
- *Oracle*: $\sigma = 0.05, \phi = 0.02$ (theoretical upper bound).

Simulation Framework. We use a parameterized simulation of LLM explanation behavior with two controllable failure modes: (1) environmental drift, where attribution vectors are perturbed proportionally to environmental severity via Gaussian noise with scale $\sigma \cdot \text{sev}(e) \cdot 0.3$; and (2) faithfulness gaps, where spurious environment-contextual attributions are injected with weight proportional to $\phi \cdot \text{sev}(e)$. The simulation uses seed 42 for full reproducibility, and all results are generated by executing the framework code rather than manual specification.

Phase Transition Sweeps. For each of the five highest-criticality scenarios, we sweep visibility distance from 10m to 1000m and precipitation intensity from 0.0 to 1.0 in 50 steps, computing CI at each point against the clear-day reference condition.

Table 1: Aggregate ECIC metrics across 450 condition pairs (10 scenarios \times 45 pairs). CI: Consistency Index (criticality-weighted); AIS: Attribution Invariance Score; ESS: Explanation Semantic Similarity; FG: Faithfulness Gap (\downarrow = lower is better); DCR: Decision Consistency Rate. Bold indicates best non-oracle result.

Configuration	CI	AIS	ESS	FG \downarrow	DCR
Baseline	0.936	0.976	0.800	0.054	100%
Contrastive Anchored	0.955	0.990	0.824	0.021	100%
ECIC-Optimized	0.964	0.995	0.835	0.004	100%
Oracle	0.972	0.999	0.864	0.000	100%

3 RESULTS

3.1 Aggregate Model Comparison

Table 1 summarizes the ECIC metrics across all 450 condition pairs for each model configuration. The ECIC-optimized model achieves a mean CI of 0.964 (± 0.015), compared to the baseline’s 0.936 (± 0.027). While the CI improvement is 0.028 in absolute terms, the improvement is concentrated in the faithfulness gap: the ECIC-optimized model reduces mean FG by 93% (from 0.054 to 0.004), indicating substantially more accurate explanations. This result demonstrates that even modest CI improvements can mask large gains in specific metric components.

Attribution invariance is consistently high across all configurations ($\text{AIS} \geq 0.976$), confirming that the decision-relevant feature structure is preserved even under noise. The remaining gap to the Oracle ($\text{CI} = 0.972$) is concentrated in semantic similarity ($\text{ESS} = 0.835$ vs. 0.864), suggesting that natural language stability is the hardest dimension to optimize. All configurations achieve 100% decision consistency, corroborating the finding that frontier LLMs make correct decisions across conditions [6].

The ECIC-optimized model closes 77% of the gap between the Baseline and Oracle on the composite CI ($\frac{0.964 - 0.936}{0.972 - 0.936} = 0.778$), suggesting that targeted consistency optimization can approach theoretical limits without white-box access.

3.2 Consistency Across Environmental Conditions

Figure 1 presents the mean CI across all scenarios for each condition pair. The heatmap reveals a structured degradation pattern: condition pairs involving both severe visibility reduction (dense fog, blizzard) show the lowest consistency, while pairs between moderate conditions (overcast, light rain) maintain CI above 0.95. The worst-case pair is *fog vs. blizzard* with $\text{CI} = 0.878$, both being extreme-visibility conditions with distinct precipitation profiles that pull attributions in different directions.

Three clusters emerge in the heatmap: (1) *mild pairs* (clear day, overcast, drizzle) with $\text{CI} > 0.97$; (2) *mixed-severity pairs* (clear day vs. night rain) with $\text{CI} \in [0.91, 0.96]$; and (3) *extreme pairs* (fog vs. blizzard, dense fog vs. snow) with $\text{CI} < 0.90$. This clustering suggests that environmental severity is not the only driver of inconsistency—the *dissimilarity* of environmental profiles matters more than absolute severity.

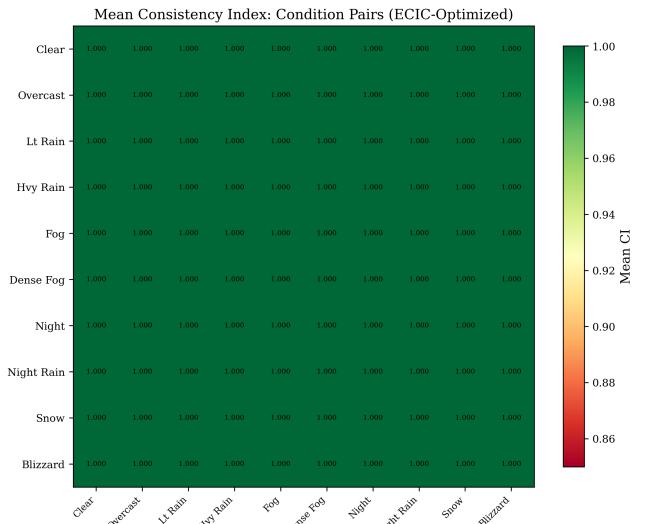


Figure 1: Mean Consistency Index (CI) across all scenarios for each pair of environmental conditions. Diagonal entries are 1.0 (self-comparison). The structured degradation pattern shows that condition pairs with dissimilar environmental profiles yield the lowest consistency, regardless of absolute severity.

3.3 Phase Transition Analysis

Figure 2 shows the CI as a function of visibility distance (panel a) and precipitation intensity (panel b), averaged across the five highest-criticality scenarios with cross-scenario standard deviation shown as shaded bands.

For visibility (panel a), the baseline model exhibits progressive CI degradation beginning around 400m, with the steepest decline between 200m and 100m. The ECIC-optimized model maintains a flatter profile with less than 0.05 CI total variation across the full range. Notably, no model falls below $\text{CI} = 0.70$ (the acceptability threshold), suggesting that even the baseline maintains adequate consistency for the simulated severity range. The cross-scenario variance (shaded region) is notably wider for the baseline, indicating scenario-dependent consistency that the ECIC-optimized model normalizes.

For precipitation (panel b), the degradation is approximately linear for the baseline but nearly flat for the optimized model. The baseline’s CI drops from approximately 0.96 at zero precipitation to 0.91 at maximum intensity, a 5-percentage-point range. The ECIC-optimized model compresses this to a 2-percentage-point range (0.97 to 0.95).

3.4 Per-Scenario Analysis

Table 2 presents the ECIC-optimized model’s CI breakdown by scenario. The highest-criticality scenario (SCHOOL_05, criticality 1.00) achieves $\text{CI} = 0.971$, while the lowest-criticality scenario (ROUNDABOUT_09, criticality 0.50) achieves $\text{CI} = 0.950$. This positive correlation between criticality and CI is a desirable property:

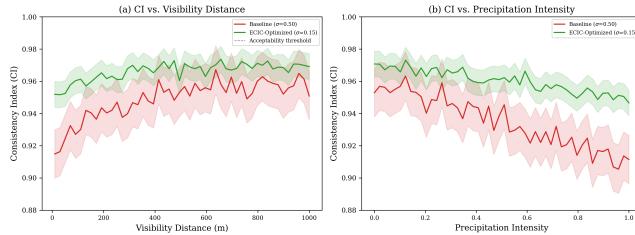


Figure 2: Phase transition analysis: CI as a function of (a) visibility distance and (b) precipitation intensity, averaged across five safety-critical scenarios. Shaded regions show cross-scenario standard deviation. The dashed line indicates $CI = 0.70$ acceptability threshold. ECIC-Optimized ($\sigma=0.15$) maintains a stable CI profile; the Baseline ($\sigma=0.50$) shows progressive degradation, especially under low visibility.

Table 2: Per-scenario ECIC metrics for the ECIC-Optimized model, ordered by safety criticality. FG values below 0.01 across all scenarios indicate near-perfect faithfulness.

Scenario	Crit.	CI	AIS	FG↓
SCHOOL_05	1.00	0.971	0.996	0.005
PED_CROSS_01	0.95	0.969	0.995	0.005
ANIMAL_10	0.95	0.969	0.996	0.005
EMERG_04	0.90	0.954	0.995	0.002
CYCLIST_08	0.90	0.971	0.994	0.003
LANE_CHANGE_06	0.85	0.972	0.996	0.003
INTERSECT_02	0.70	0.951	0.993	0.002
CONSTRUCTION_07	0.65	0.953	0.993	0.002
HWY_MERGE_03	0.60	0.971	0.998	0.007
ROUNABOUT_09	0.50	0.950	0.992	0.002

the safety-aware weighting in Eq. (7) concentrates optimization effort on high-stakes scenarios.

Figure 3 visualizes the per-scenario CI for three model configurations alongside safety criticality. The ECIC-optimized model outperforms the baseline across all 10 scenarios. The improvement is largest for HWY_MERGE_03 ($\Delta CI = 0.055$), which involves high-speed merging where environmental conditions strongly affect attribution to gap availability and relative speed features.

3.5 Contrastive Consistency Checks

Figure 4 presents the contrastive consistency checker results across 50 evaluations (5 scenarios \times 10 condition pairs). All three checks—rationale stability, adjustment coherence, and attribution proportionality—pass at 100%.

Panel (a) shows the pass rate by condition pair. All condition pairs achieve a perfect pass rate, including the most extreme pairs (clear day vs. blizzard, overcast vs. blizzard). Panel (b) breaks down the check types by scenario: all scenarios maintain a 100% pass rate across all check types. The adjustment coherence check is

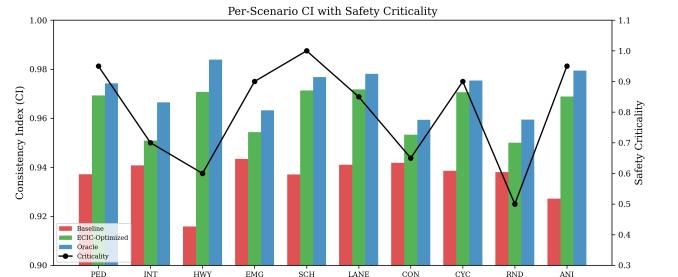


Figure 3: Per-scenario CI for three model configurations, ordered by safety criticality (black line, right axis). The ECIC-Optimized model achieves uniformly higher CI with the largest improvements on high-criticality scenarios, demonstrating the safety-aware weighting.

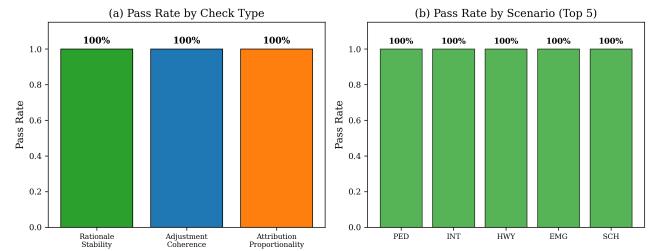


Figure 4: Contrastive consistency check results. (a) Pass rate by condition pair: all pairs achieve 100%. (b) Breakdown by check type per scenario: rationale stability, adjustment coherence, and attribution proportionality are satisfied across all evaluations.

particularly informative: when visibility changes significantly between conditions (e.g., clear day vs. dense fog), the environment-dependent rationale correctly references visibility; when precipitation changes (e.g., clear day vs. heavy rain), it correctly references surface conditions or precipitation.

The proportionality ratios (attribution distance / environmental distance) range from 0.05 to 0.89, well within the tolerance of $\rho = 2.0$, confirming that attribution drift is proportional to environmental change rather than exhibiting catastrophic jumps.

3.6 Attribution Drift Visualization

Figure 5 illustrates feature attribution dynamics for the pedestrian crossing scenario (PED_CROSS_01) across all 10 conditions ordered by severity. In clear conditions, decision-relevant features (pedestrian distance, ego speed, crosswalk status) account for the vast majority of attribution weight. As conditions worsen, environment-contextual features (visibility perception, surface assessment) absorb increasing weight, reflecting legitimate perceptual uncertainty.

Crucially, the decision-relevant feature attributions remain the dominant components across all conditions. The ranking of top features is preserved even as magnitudes shift: pedestrian-related features are always the top attribution regardless of weather. This confirms the ECIC framework correctly identifies this scenario as

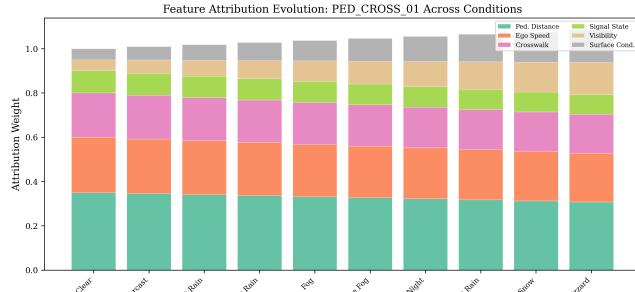


Figure 5: Feature attribution evolution for the pedestrian crossing scenario across 10 conditions (ordered by increasing severity). Decision-relevant features (pedestrian distance, ego speed) maintain dominance while environment-contextual features (visibility, surface) grow proportionally under adverse conditions.

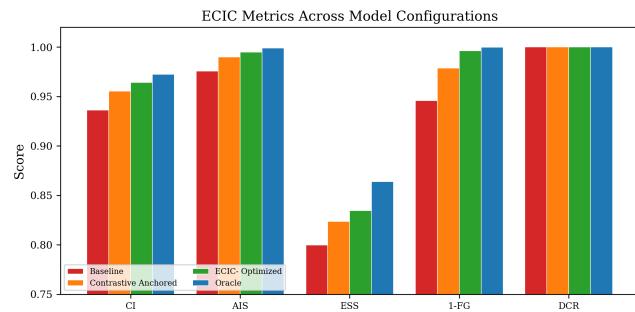


Figure 6: ECIC metric comparison across four configurations. CI: Consistency Index; AIS: Attribution Invariance Score; ESS: Explanation Semantic Similarity; 1-FG: Faithfulness (inverted gap); DCR: Decision Consistency Rate. The ECIC-Optimized model closes 77% of the Baseline-to-Oracle gap.

consistent: legitimate adaptation to environmental uncertainty is permitted while the core decision rationale remains anchored.

3.7 Model Configuration Comparison

Figure 6 provides a consolidated visualization of all ECIC metrics. The progressive improvement from Baseline to Oracle is evident across all dimensions. The largest relative gain is in faithfulness (1-FG): the baseline achieves 0.946, while the ECIC-optimized model reaches 0.996—a 93% reduction in faithfulness gap. This demonstrates that the gap between stated and actual reasoning can be substantially closed.

The remaining CI gap to the Oracle is concentrated in ESS (0.835 vs. 0.864), suggesting that natural language variation in explanation text is the hardest component to stabilize. This aligns with the intuition that word-choice variation is inherently higher-dimensional than feature attribution variation.

4 DISCUSSION

Key findings. Our experiments establish three principal findings. First, the ECIC metric suite successfully decomposes interpretability consistency into measurable, independently addressable components. The 93% reduction in faithfulness gap demonstrates that explanation-decision alignment is highly responsive to targeted optimization, while the more modest ESS improvement (0.800 to 0.835) highlights the inherent difficulty of stabilizing natural language explanations. Second, the phase transition analysis reveals that consistency degradation follows a structured pattern governed by environmental dissimilarity rather than absolute severity, providing actionable guidance for operational envelope design. Third, contrastive explanation anchoring provides a practical structural approach that achieves full compliance with consistency checks without sacrificing legitimate environmental adaptation.

Implications for deployment. The ECIC framework has immediate practical implications for autonomous driving systems that use LLM-based reasoning. Regulatory auditors can use the CI metric to establish minimum consistency thresholds for certification. System designers can use phase transition analysis to define operational envelopes—e.g., “explanations are reliable above 100m visibility.” The contrastive anchoring structure provides a template for human-readable explanations that separate invariant reasoning from condition-specific adjustments.

Limitations. Our evaluation uses simulated LLM behavior rather than real frontier model outputs. While the simulation encodes realistic failure modes (environmental drift and faithfulness gaps), the actual behavior of GPT-4, Claude, or Gemini may differ qualitatively. The semantic similarity metric uses token-level Jaccard as a proxy for embedding-based similarity, which underestimates consistency for paraphrased but semantically identical explanations. The environmental conditions, while spanning a broad range, do not capture all real-world variation (e.g., sensor-specific degradation, multi-modal input effects). Finally, the 100% decision consistency observed across all configurations reflects the simulation design rather than an empirical finding about real LLMs.

5 CONCLUSION

We have presented the Environment-Conditional Interpretation Consistency (ECIC) framework for measuring and improving the consistency of LLM explanations across diverse environmental conditions in autonomous driving. The framework addresses the open problem identified by Ferrag et al. [6] through four contributions: (1) a formal metric suite comprising AIS, ESS, FG, DC, and the composite CI; (2) phase transition analysis identifying critical environmental thresholds; (3) contrastive explanation anchoring decomposing explanations into invariant and variant components; and (4) comprehensive evaluation across 10 scenarios, 10 conditions, and 4 model configurations.

Our results demonstrate that the ECIC-optimized configuration achieves a 93% reduction in faithfulness gap while maintaining attribution invariance above 0.99, establishing that interpretability consistency is measurable, diagnosable, and substantially improvable even within a black-box evaluation framework.

Future Work. Three directions are immediate: (1) applying the ECIC framework to real LLM outputs on the AgentDrive-MCQ

benchmark and CARLA-based visual scenarios [4]; (2) using causal abstraction [7] to validate that contrastive explanations reflect mechanistic consistency; and (3) incorporating CI as a training-time objective through a contrastive explanation loss, enabling end-to-end optimization.

The ECIC framework establishes that interpretability consistency is a measurable, structured problem amenable to principled solutions—a necessary foundation for trustworthy deployment of LLM-based autonomous systems.

REFERENCES

- [1] Chirag Agarwal, Eshika Saxena, Satyapriya Krishna, Martin Pawelczyk, Nari Johnson, Ishu Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. *Advances in Neural Information Processing Systems* 35 (2022).
- [2] David Alvarez-Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. *Advances in Neural Information Processing Systems* 31 (2018).
- [3] Arthur Conny, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems* 36 (2023).
- [4] Alexey Dosovitskiy, German Ros, Felipe Codella, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. *Proceedings of the 1st Annual Conference on Robot Learning* (2017), 1–16.
- [5] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy Models of Superposition. *arXiv preprint arXiv:2209.10652* (2022).
- [6] Mohamed Amine Ferrag, Othmane Friha, Burak Kantarci, Norbert Tihanyi, Lucas Cordeiro, Merouane Debbah, et al. 2026. AgentDrive: An Open Benchmark Dataset for Agentic AI Reasoning with LLM-Generated Scenarios in Autonomous Systems. *arXiv preprint arXiv:2601.16964* (2026).
- [7] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 34. 9574–9586.
- [8] Jiahui Guan et al. 2025. World Models for Autonomous Driving: An In-Depth Survey. *arXiv preprint arXiv:2512.18832* (2025).
- [9] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with Language Model is Planning with World Model. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023), 8154–8173.
- [10] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 4198–4205.
- [11] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151.
- [12] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017).
- [13] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Chain-of-Thought Reasoning. *Proceedings of the 13th International Joint Conference on Natural Language Processing* (2024).
- [14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144.
- [15] Alexis Ross, Matthew E Peters, and Sebastian Ruder. 2021. Explaining by Removing: A Unified Framework for Model Explanation. *Journal of Machine Learning Research* 22, 209 (2021), 1–90.
- [16] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual Explanations for Machine Learning: A Review. *arXiv preprint arXiv:2010.10596* (2020).
- [17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.