From MIRAGE to CLEAR: Component-Level Explainable Anomaly Reasoning for Autonomous Vehicle Perception Systems

David Fernandez, Pedram MohajerAnsari, Amir Salarpour, Cigdem Kokenoz, Bing Li, Mert D. Pesé
Clemson University, Clemson, SC, USA
{dferna3, pmohaje, asalarp, ckokeno, bli4, mpese}@clemson.edu

***Abstract*—Autonomous vehicle (AV) perception systems remain vulnerable to failures that current anomaly detectors can flag but cannot trace to a specific component; an attribution gap that impedes forensics and emerging transparency mandates like the EU AI Act. We introduce a unified framework comprising MIRAGE and CLEAR. MIRAGE (MUTCD-Informed Realistic Anomalous Generation Engine) integrates 48,022 real-world driving scenes from nuScenes, Waymo Open, and Argoverse 2 with structured rules from the Manual on Uniform Traffic Control Devices (MUTCD), generating 5,847 annotated anomalous scenarios with known module-level ground truth across four violation types: direct, subtle, contextual, and environmental. This data powers CLEAR, a hierarchical three-layer LLM pipeline that detects anomalies, classifies violations, and attributes failures to Traffic Sign Recognition (TSR), Automated Lane Centering (ALC), or Object Detection (OD) with interpretable justifications. Confidence-gated propagation and schema-constrained outputs prevent error cascades and minimize hallucinations. CLEAR achieves 95.2% detection accuracy, 62.3% classification accuracy, and 84% attribution accuracy on direct TSR violations. A Top-2 evaluation paired with per-module confidence analysis raises overall attribution to 74.6%, confirming that apparent misattributions largely reflect the multi-module nature of anomaly types rather than reasoning failures. These results show that grounding LLM reasoning in structured traffic regulations enables reliable, interpretable forensics for AV perception systems, offering a practical path toward auditable, regulation-compliant safety analysis. The code for this work is available at https://github.com/David-FR/MIRAGE_CLEAR.**

***Index Terms*—Autonomous vehicles, anomaly detection, component attribution, explainable AI, safety analysis, regulatory compliance.**

I. Introduction

Autonomous vehicles (AVs) depend on perception systems to interpret their environment and navigate safely [1]. Deep neural networks (DNNs) typically drive these systems, supporting essential functions like traffic sign recognition (TSR), automated lane centering (ALC), and object detection (OD) [2]. Although these networks perform well under normal conditions [3], they remain vulnerable to adversarial examples. These subtle perturbations, such as modified traffic signs or altered lane markings, are often imperceptible to humans but can result in critical perception failures [4], [5].

Recent accident data highlights the severity of these vulnerabilities. Statistics from the National Highway Traffic Safety Administration (NHTSA) *Standing General Order* show that crashes involving Level-2 Advanced Driver-Assistance Systems (ADAS) increased from 1,612 between July 2021 and July 2024 to 2,359 by April 2025 [6], [7]. While perception failures caused 17% of AV disengagements, only 4% included clearly identified causes [8]. This *attribution gap* impedes post-incident forensics, delays safety interventions, and obstructs systematic upgrades to AV perception stacks.

TABLE I: Comparison of existing approaches with CLEAR across four criteria: anomaly detection, component attribution, explainability, and regulatory compliance. CLEAR uniquely supports all dimensions in a unified pipeline.

Approach	Anomaly Detection	Component Attribution	Explainability	Regulatory Compliance
Statistical Methods [11]	●	○	○	○
Deep Learning AD [12]	●	○	○	○
Rule-based Systems [13]	●	◑	●	◑
Vision-Language Models [14]	●	○	●	○
Counterfactual Analysis [15]	●	●	○	○
CLEAR	●	●	●	●

Concurrently, the regulatory landscape is shifting toward strict accountability. Frameworks such as the EU AI Act (2024/1689) [9] and the NHTSA AV STEP proposal [10] mandate transparency in high-risk AI systems by emphasizing subsystem observability. Consequently, generic high-level alerts are insufficient. Regulators now require the precise identification of the failed component and the failure rationale. However, in modular AV stacks, observable symptoms rarely localize to a single module. A failure to stop might originate from OD localization errors, TSR misclassification, or ALC context misinterpretation. As summarized in Table I, existing statistical [11], deep learning [12], and rule-based [13] methods effectively detect anomalies but fail to isolate the responsible submodule or provide regulatory-aligned explanations.

This attribution gap is not a simple oversight in previous research; rather, it reflects a fundamental structural challenge. In modular AV perception systems, core subsystems such as TSR, ALC and OD are closely linked through shared sensor data and overlapping features. When an environmental condition such as fog degrades a scene, it simultaneously affects the TSR's contrast sensitivity, ALC's edge detection and OD's depth estimation. Because every component is affected, it is

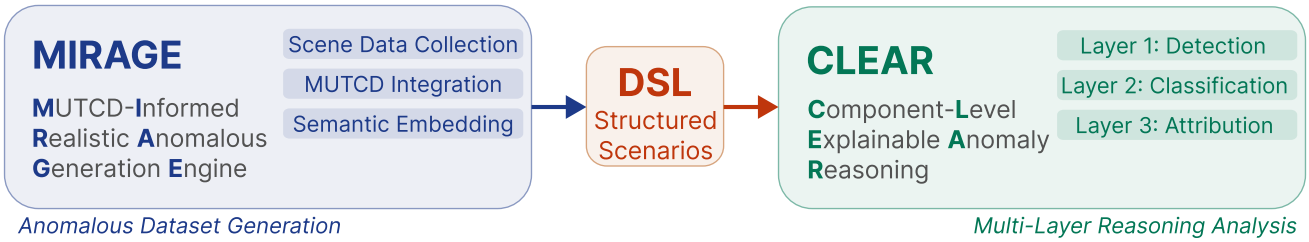

Fig. 1: Overview of our two-part system: MIRAGE constructs a policy-grounded dataset by injecting MUTCD-aligned anomalies into diverse driving scenes; CLEAR is a multi-layer reasoning pipeline that detects anomalies, classifies violation types, and attributes them to specific AV perception modules using interpretable LLM-generated reasoning.

inherently difficult to isolate a single primary source of failure.

To bridge this gap, we introduce a unified framework, illustrated in Figure 1, that integrates policy-grounded dataset generation with reasoning-based attribution:

Dataset Generation (MIRAGE). We develop the **M**UTCD-**I**nformed **R**ealistic **A**nomalous **G**eneration **E**ngine. MIRAGE synthesizes 48,022 scenes by injecting targeted anomalies into real-world data (nuScenes [16], Waymo [17], Argoverse2 [18]). It aligns unstructured visual scenes with structured rules from the Manual on Uniform Traffic Control Devices (MUTCD) [19] to ensure anomalies have a traceable ground-truth mapping to specific perception modules (see Figure 2).

Reasoning Framework (CLEAR). We present **C**omponent-**L**evel **E**xplainable **A**nomaly **R**easoning. CLEAR (Figure 3) is a three-stage LLM-based pipeline that detects anomalies, classifies violation types, and attributes failures to specific modules (TSR, ALC, or OD). While individual layers utilize established chain-of-thought prompting [20], our technical contribution lies in the architecture rather than the prompting technique. Unlike single-stage methods [15] that merge detection and attribution, our hierarchical design introduces confidence-gated propagation, which halts reasoning when confidence drops to prevent error cascades, and task-specific schema constraints, which restrict outputs to fixed taxonomies to minimize hallucinations in reasoning descriptions.

We evaluated our system across 5,847 annotated anomalies. **CLEAR achieves** 95.2% **detection accuracy and** 62.3% **violation classification accuracy.** For component attribution, which no prior method achieves, **CLEAR reaches** 84% **accuracy on direct TSR violations**, the most relevant category based on the dataset. While strict single-module attribution accuracy is 41.7%, a Top-2 evaluation paired with per-module confidence analysis raises this to 74.6% overall and 75.0% for environmental violations. The correlation between confidence distribution and recovery rate across violation types confirms that apparent misattributions largely reflect the multi-module nature of anomaly types rather than reasoning failures. CLEAR is designed for offline forensic applications, including post-incident analysis, validation pipelines, and regulatory audits, therefore interpretability and regulatory traceability are prioritized over real-time latency. Our contributions are as follows:k

- **Policy-anchored dataset construction.** We propose **MIRAGE**, a novel pipeline that generates driving scenarios with annotated, component-specific perception anomalies grounded in MUTCD rules and real-world sensor data.
- **Attribution-focused reasoning architecture.** We design **CLEAR**, a hierarchical reasoning pipeline that decomposes anomaly analysis into detection, regulatory classification, and component attribution.
- **Empirical validation and deployment analysis.** We demonstrate that CLEAR satisfies detection, attribution, explainability, and compliance criteria where existing prompt-based and counterfactual methods fall short.

II. Related Work

The safety, robustness, and explainability of autonomous vehicle (AV) perception systems have been extensively studied. We examine AV perception safety across four critical dimensions: adversarial robustness, anomaly detection, neuro-symbolic reasoning, and regulatory compliance.

Adversarial Vulnerabilities in Perception. Perception modules (TSR, ALC, OD) remain susceptible to adversarial perturbations ranging from physical stickers [4], [21], [22] and lane spoofing [23] to spatial-temporal camouflage [24]. While defensive strategies like adversarial training exist [25], they are predominantly reactive. Crucially, these defenses lack the diagnostic granularity to identify specific failing subsystems which is a prerequisite for regulatory compliance.

Anomaly Detection and the Attribution Gap. Current anomaly detection relies on statistical deviations [11], learned representations [26], or rule-based logic [12]. However, these approaches exhibit a fundamental *attribution gap*. Statistical methods treat the perception stack as a monolithic black box, while deep learning representations mix features across modules which renders attribution impossible. Similarly, heuristic rules fail to generalize across combinatorial real-world scenarios [27]. Consequently, existing pipelines detect *that* a failure occurred but fail to localize *where*. Because of this gap, existing anomaly detection methods can only be compared using binary detection metrics. No prior work provides component-level attribution results (see Table I); therefore, CLEAR's contribution of identifying the specific perception module that failed cannot be benchmarked against existing methods.

Component-Level Attribution. Recent work [15] targets component-level attribution via counterfactual causality analysis. This approach requires idealized ground-truth components

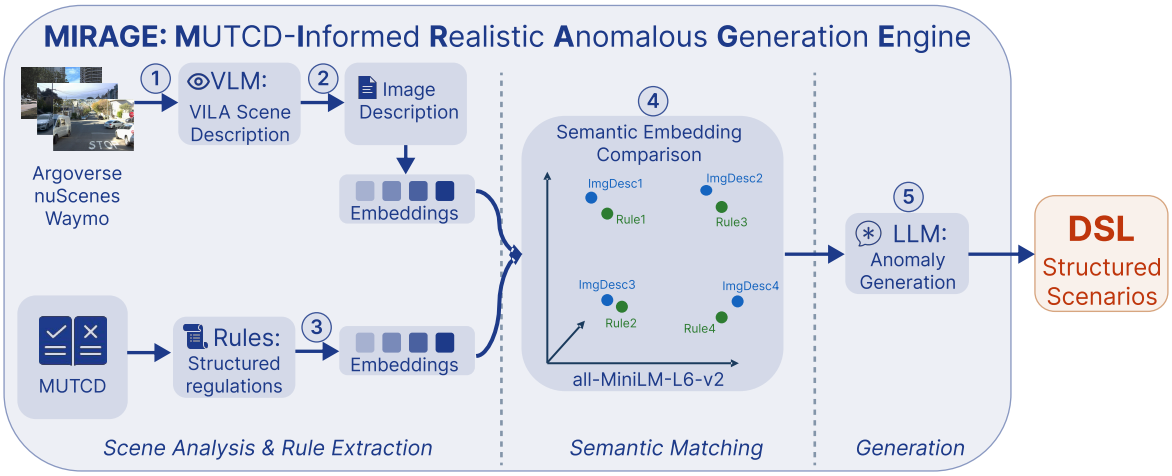

Fig. 2: MIRAGE framework: From real-world AV datasets (nuScenes, Waymo, Argoverse2), MIRAGE extracts scenes and annotates them with MUTCD-aligned regulatory violations using semantic embeddings and vision-language models (VLMs). Each scene is paired with structured DSL outputs and known component-level anomaly ground truth.

substitutes. For example, these methods might replace a real TSR module with a perfect module to determine if system failures persist. However, this substitutes are unavailable in real-world perception data from production AV systems. CLEAR instead operates on observational data that does not require a simulation environment, making it suitable for forensic analysis of actual deployments.

LLMs and Regulatory Alignment. Frameworks such as the EU AI Act [9] and NHTSA's AV STEP [10] mandate subsystem observability and traceability. While LLMs demonstrate utility in transparency-critical domains via chain-of-thought reasoning [20], current Vision-Language Models (VLMs) describe scenes without structured diagnostic capabilities [28]–[30]. Furthermore, most XAI methods explain decisions without localizing failure sources [31]. **CLEAR** addresses this by leveraging the **MIRAGE** dataset to embed MUTCD-aligned regulatory logic into the diagnostic process, unifying detection, attribution, and compliance in a single framework.

III. Background

To support interpretable attribution of perception anomalies, we review the architecture of AV perception systems, regulatory standards that govern their behavior, and emerging reasoning techniques for structured, safety-critical diagnostics.

AV Perception Architecture. Modern AVs use a modular perception stack, where specialized subsystems process sensor data to construct a coherent understanding of the driving environment. We focus on three perception modules [32]:

- **Traffic Sign Recognition (TSR):** identifies regulatory signage and triggers control actions such as braking.
- **Automated Lane Centering (ALC):** delineates lane boundaries via semantic segmentation for lateral control.
- **Object Detection (OD):** localizes obstacles by fusing sensor data to enable collision avoidance.

While efficient, this modularity introduces vulnerabilities. Recent work has shown that anomaly detection systems can be compromised through sophisticated attacks [33]. Thus, component-level attribution is essential for identifying the source of perception failures and enabling effective diagnostics, safety validation, and regulatory compliance.

Regulatory Ground Truth: MUTCD Rule Taxonomy. The Manual on Uniform Traffic Control Devices (MUTCD) defines the national standard for traffic signs, signals, and road markings in the US [19]. MUTCD rules fall into four categories: Standard rules are legally mandatory and must always be followed (e.g., stopping at a red light); Guidance rules represent recommended practices that encourage safety or consistency, subject to engineering judgment; Option rules permit flexible actions based on situational context (e.g., placement of warning signs); and Support rules provide non-binding contextual information with no regulatory force. These categories serve as the foundation for our framework, allowing violations to be interpreted in a policy-aligned context.

LLMs for Structured Reasoning. LLMs such as GPT-4 and Phi-4 demonstrate state-of-the-art capabilities in contextual reasoning, logical inference, and explanation generation. Unlike conventional DNNs, which often produce opaque outputs, LLMs are capable of articulating their decision-making processes. This makes them particularly well-suited for post-hoc diagnostics in safety-critical domains. Techniques such as Chain-of-Thought (CoT) prompting [20] and self-consistency decoding [34] significantly enhance LLM reasoning fidelity. In autonomous driving, CoT-style decomposition has been adopted for real-time decision-making: DriveCoT [15] decomposes perception into sub-tasks whose intermediate outputs feed a final planning decision, while DriveGPT4 [35] employs multimodal LLMs to provide step-by-step justifications grounded in visual input. CLEAR extends this principle to offline forensic analysis, where each reasoning layer (detection $\rightarrow$ classification $\rightarrow$ attribution) produces structured outputs grounded in regulatory standards. These capabilities are leveraged in our framework to generate structured reasoning chains

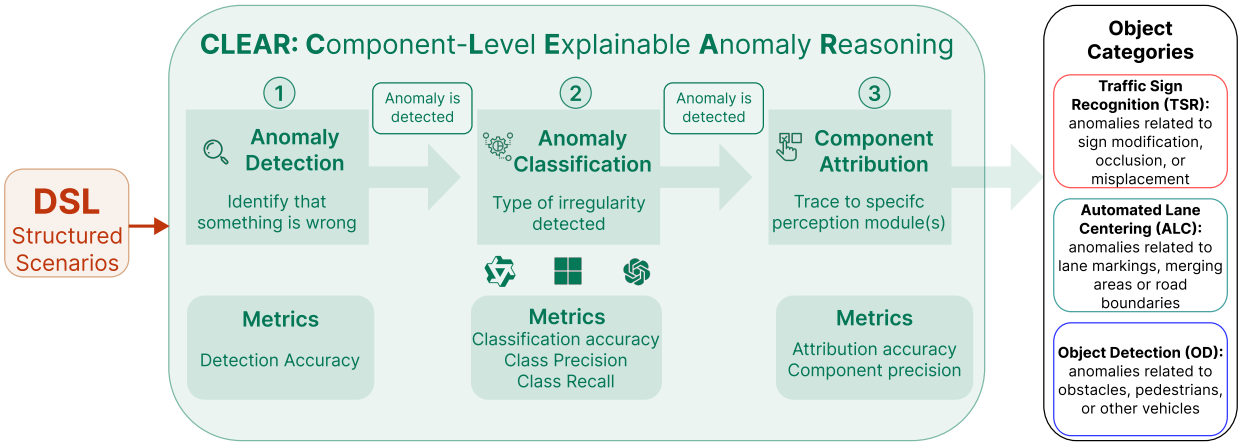

Fig. 3: CLEAR's hierarchical reasoning framework: Layer 1 detects potential anomalies, Layer 2 classifies the type of violation, and Layer 3 attributes the anomaly to specific AV perception components (TSR, ALC, OD). Each step provides interpretable reasoning with confidence scores.

that explain both the nature and source of anomalies.

Semantic Embedding for Rule Matching. To align textual scene descriptions with relevant MUTCD rules, our framework employs semantic embedding techniques. Semantic embedding models enable efficient comparison of textual content based on meaning rather than surface-level keyword overlap. These models map both scene narratives and regulatory texts into high-dimensional vector spaces, preserving meaning through geometric similarity. Cosine similarity is used to identify the most relevant rules for a given scene. We leverage a combination of pretrained sentence encoders (e.g., `all-MiniLM-L6-v2` [36]), keyword-extraction methods (e.g., KeyBERT [37]), and LLM-derived semantic anchors, i.e., textual templates or keyphrase vectors generated by prompting LLMs, to enhance context-aware alignment. Scene descriptions are extracted using a VLM such as `GPT-4V`, which translates multimodal input (e.g., a traffic scene image) into structured, natural language narratives. This semantic alignment enables precise mapping from scene understanding to rule violations and component-level fault tracing.

IV. Threat Model

We consider a realistic adversarial setting in which attackers induce failures in AV perception systems by modifying elements of the driving environment. The attacker has no access to the vehicle's internal software stack, perception models, or sensor configurations, but can manipulate external stimuli to exploit vulnerabilities in specific perception modules.

Attacker Goals. The adversary aims to cause specific failures that meet four conditions: *1) Stealth.* Modifications must be unnoticeable to humans to avoid detection. *2) Targeted Misperception.* The attack causes a specific module (TSR, ALC, or OD) to make incorrect predictions. *3) Cross-Module Evasion.* The attack affects only the targeted module without impacting others. *4) Regulatory Ambiguity.* The attacker uses legally valid but confusing designs, such as unusual fonts, that comply with regulations but confuse the perception system.

Adversarial Capabilities. We assume the attacker operates with the following limits: *1) Physical Access.* The attacker modifies physical objects like signs or lane markings using stickers or paint. We assume physical changes remain a viable threat even if digital patches are detectable. *2) Semantic Crafting.* Modifications follow MUTCD standards for shape and color but include small deviations that degrade machine performance. *3) Black-Box Knowledge.* The attacker does not know specific model parameters but understands the general system architecture and its weaknesses.

Assumptions and Constraints. To ensure realism, we apply the following conditions: *1) Regulatory Plausibility.* All modifications must appear to be valid traffic control devices to a human inspector. *2) Observational Independence.* We treat TSR, ALC, and OD as separate observational units. We attribute errors based on the visible output of each module rather than analyzing internal feature dependencies. This aligns with regulations that focus on observable system behavior rather than internal architecture. *3) Sensor Visibility.* Modifications must be within the standard field of view and resolution of the sensors. *4) Immutable Reasoning.* The attacker cannot tamper with the LLM reasoning engine or the external regulatory database. *5) Temporal Stability.* Modifications must last long enough to be processed by the vehicle's sensor fusion system.

V. MIRAGE and CLEAR: Attribution-Ready Dataset and Reasoning Framework

This section introduces a unified two-phase framework for interpretable, component-level failure attribution in AV perception. We combine regulatory alignment with structured reasoning for traceable, policy-aware analysis. **MIRAGE** is a dataset construction pipeline of driving scenarios annotated with policy-grounded traffic control violations. Each scenario is semantically linked to the MUTCD and organized for targeted analysis of perception module behavior, under realistic conditions. **CLEAR** is a hierarchical reasoning framework powered by LLMs, designed to detect anomalies, classify

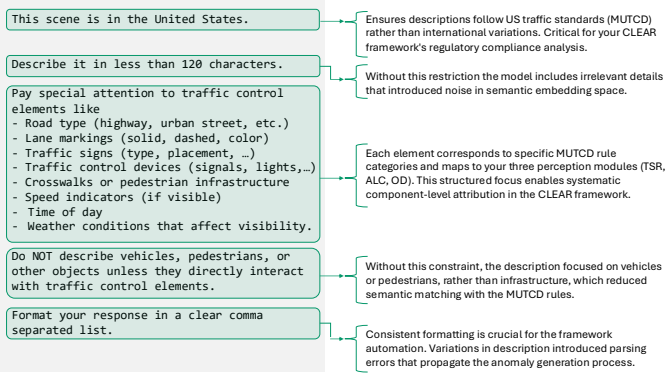

Fig. 4: Prompt used with the VILA VLM to extract concise, regulation-relevant descriptions of AV driving scenes. Focus is on infrastructure elements affecting traffic compliance.

violation types, and attribute perception errors to specific AV subsystems with interpretable justifications.

Together, **MIRAGE** and **CLEAR** form the first integrated system for regulatory-aligned, component-level diagnostic reasoning in AV perception. Critically, the LLM serves merely as a reasoning engine within our architecture, similar to how CNNs serve as feature extractors in object detection. The system's value emerges from the end-to-end framework. Direct prompting without our semantic alignment and dataset generation would not achieve the same attribution accuracy, demonstrating that component attribution requires architectural innovation beyond prompt engineering.

A. MIRAGE: Dataset Generation and Regulatory Grounding

MIRAGE is a purpose-built dataset designed to enable traceable and interpretable attribution of AV perception failures. As illustrated in Figure 2, its construction follows a four-phase pipeline: (1) scene collection and filtering, (2) regulatory rule extraction, (3) semantic matching, and (4) violation grounding and annotation.

❶ **Scene Collection and Filtering.** We extract driving scenes from three diverse AV datasets: nuScenes [16], Waymo Open [17], and Argoverse 2 [18], chosen for their geographic, infrastructural, and regulatory variance:

- **nuScenes**: Urban scenes from Boston and Singapore, capturing dense intersections (TSR-heavy).
- **Waymo**: Urban/suburban settings from Phoenix, San Francisco, and Mountain View.
- **Argoverse 2**: Complex road networks across Detroit, Miami, and Pittsburgh, relevant to OD and ALC.

To ensure compatibility with the U.S. MUTCD, only U.S.-based scenes are retained. We downsample sequences from 10–20 Hz to 5 Hz to reduce temporal redundancy while preserving key semantic changes. The final dataset contains 48,022 images: 3,728 from nuScenes, 36,290 from Waymo, and 8,004 from Argoverse 2.

❷ **Scene Description via Vision-Language Models.** We employ VILA [38] VLM, to extract textual descriptions of each scene focused on regulatory-relevant features (e.g., signs, road markings). As illustrated in Figure 4, a prompt is crafted to ignore non-essential elements in the scene and include all elements that can be referenced in the MUTCD.

❸ **Regulatory Rule Extraction.** To support policy-grounded diagnostics, we convert the MUTCD sections into a structured JSON rule base (shown in Listing 1) through a multi-step extraction process illustrated in Figure 5. First, we parse section headers and rule boundaries to isolate individual regulations; an example (Rule 2B.04) is shown in Figure 6. Each rule is then classified by type (*Standard*, *Guidance*, etc.), with priority given to safety-critical constraints. Atomic rules are extracted from numbered paragraphs and augmented with semantic keywords using both KeyBERT [37] and LLM-based keywording (`Phi-4`) to support embedding and retrieval:

The result is a searchable regulatory knowledge base compatible with natural language scene descriptions.

```
{
 "rule_id": "2B.04-S01",
 "section": "2B.04 STOP Sign (R1-1) and ALL-WAY
      Plaque (R1-3P)",
 "type": "Standard",
 "paragraph_num": "01",
 "text": "A full stop is always required on an
     approach to an intersection, a STOP (R1-1)
     sign (Figure 2B-1) shall be used.",
 "keywords": ["stop", "sign", "intersection"],
 "keywords_embeddings": [0.73, 0.12, 0.09, ...]
}
```

Listing 1: Structured JSON representation of MUTCD section.

❹ **Semantic Matching.** We compute scene-to-rule alignment using cosine similarity between embeddings generated via `all-MiniLM-L6-v2`. We evaluate different representations (full-text, KeyBERT keywords, LLM keywords) and select the top-k most relevant rules per scene that exceed a similarity threshold. To determine the optimal threshold, we conducted an ablation study across values from 0.50 to 0.80 (Table II). We found that lower thresholds (0.50–0.60) introduce significant noise (up to 9.3 rules per scene), whereas higher thresholds (0.70–0.80) cause substantial coverage loss. A threshold of 0.65 achieves the optimal trade-off: scenes receive an average of 2.4 matched rules with 0.712 average similarity, while only 18.6% of scenes lack matches. This configuration ensures sufficient rule coverage for the CLEAR reasoning pipeline while maintaining strong semantic alignment between scene descriptions and regulatory text.

❺ **Anomaly Generation MIRAGE** integrates regulatory semantics with diverse driving contexts to enable structured attribution. Each annotated violation is explicitly linked to the affected perception subsystem, e.g., sign visibility for TSR, lane integrity for ALC, and occlusions for OD. This systematic injection of MUTCD-grounded anomalies creates a benchmark where ground-truth attribution is verifiable against legal standards rather than subjective labels, establishing a new paradigm for safety-critical AI evaluation.

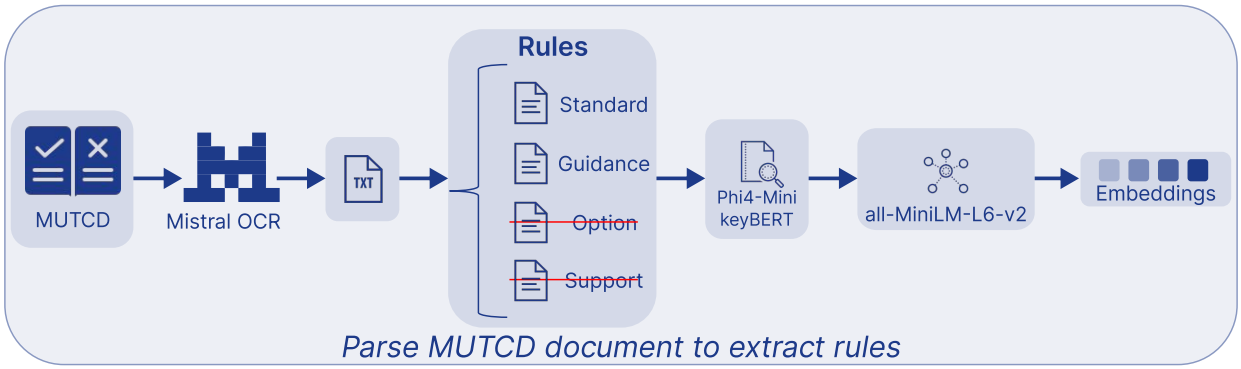

Fig. 5: Extraction of MUTCD rule sections: The structured regulatory knowledge base is built by parsing MUTCD documents, enriching rules with LLM and keyword-based semantic anchors, and embedding them for rule-scene matching.

B. CLEAR: Layered LLM-Based Reasoning Pipeline

CLEAR is a hierarchical, interpretable reasoning system designed to attribute anomalies in AV perception systems to specific components. Drawing inspiration from expert diagnostic processes, **CLEAR** decomposes complex attribution tasks into three sequential reasoning layers (illustrated in Figure 3):

1. **Layer 1 – Anomaly Detection:** Determine whether the scene contains a deviation from regulatory norms.
2. **Layer 2 – Violation Classification:** Categorize violation type (e.g., direct, subtle, contextual, or environmental).
3. **Layer 3 – Component Attribution:** Identify which perception module (TSR, ALC, OD) impacted and why CLEAR's Layer 3 output includes a ranked list of affected modules with per-module confidence scores, enabling both strict single-module and relaxed Top-2 evaluation.

Traditional anomaly detectors operate as black boxes, often yielding opaque decisions with limited interpretability. **CLEAR** addresses this limitation by producing structured outputs, grounded justifications, and confidence scores at every stage, supporting both technical debugging and regulatory auditability. **CLEAR** adheres to three core design goals: **Hierarchical Decomposition:** Attribution is broken into logically dependent subtasks, simplifying complexity. **Explainability:** Each layer outputs interpretable natural-language rationales alongside structured results. **Confidence-Gated Reasoning:** Reasoning halts if a layer's confidence is below threshold, mitigating error propagation.

To assess the sensitivity of CLEAR to specific prompt formulations, we evaluated five variations: the baseline (Listings 2, 3, and 4), along with minimal, verbose, reordered, and paraphrased versions. The results are summarized in Table III.

TABLE II: Ablation study on cosine similarity threshold for semantic matching between scene descriptions and MUTCD rules using `all-MiniLM-L6-v2` embeddings.

Threshold	Avg. Matches	No Match (%)	<5 Matches (%)	Avg. Sim.
0.50	9.3	2.1	12.4	0.587
0.55	6.1	5.8	24.7	0.621
0.60	3.8	12.3	41.2	0.668
0.65*	**2.4**	**18.6**	**52.8**	**0.712**
0.70	1.3	38.4	71.5	0.756
0.75	0.6	61.2	86.3	0.798
0.80	0.2	82.7	94.1	0.836

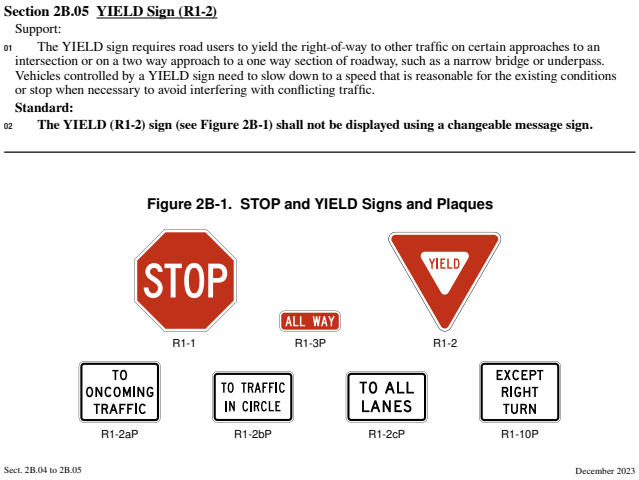

Fig. 6: Visual excerpt from MUTCD Section 2B.04: Regulatory specifications for STOP and YIELD signs including their physical standards, placement, and optional supplements.

```
Scene Description: {scene_description}
Additional Context:
- Road Type: {road_type}
- Lane Markings: {lane_markings}
- Traffic Signs: {traffic_signs}
- Traffic Control Devices: {traffic_control_devices}
- Time of Day: {time_of_day}
- Weather: {weather}
Analyze the scene and provide your response in the
  following JSON format: {{ ... }}
```

Listing 2: CLEAR Layer 1 prompt: identifies regulatory anomalies within VILA-generated scene descriptions.

Performance remained stable within $\pm 1.5\%$ across prompt variations that preserved semantics. This indicates that the effectiveness of **CLEAR** derives from its architecture, rather than precise phrasing. Consequently, users can adapt prompt wording to their needs without performance loss.

Layer 1: Anomaly Detection. The LLM analyzes VILA-generated scene descriptions alongside matched regulatory context to determine whether a perception-disrupting anomaly is present. It outputs a binary label (anomaly / no anomaly), a confidence score, and a rationale. The prompt used for this layer is shown in Listing 2.

Layer 2: Violation Classification. If an anomaly is detected, it is categorized into one of four violation classes:

- **Direct:** Explicit MUTCD violation, for example, a missing STOP sign or improper placement/mounting.
- **Subtle:** Low-visibility, such as faded centerlines, worn crosswalk paint, or partially occluded signs
- **Contextual:** Violations dependent on dynamic or environmental context, for example, school-zone timing or temporary detours that lack required advance warning
- **Environmental:** Conditions that impair perception, including sun glare, fog, heavy rain, or deep shadows

This categorization informs downstream module attribution by establishing the nature of perception failure. Listing 3 depicts the detailed prompt used for classification.

Layer 3: Component Attribution. This layer identifies the affected perception module (TSR, ALC, or OD) based on how the classified violation would affect sensor interpretation or downstream decision-making. The output includes the primary affected module, detailed reasoning with scene evidence, and descriptions of expected failure modes for each module.

Confidence Thresholds. Each layer outputs a confidence score $c \in [0, 1]$, and reasoning propagates only if the score exceeds a layer-specific threshold. We employ progressively higher thresholds, $\tau_1 = 0.70$ (detection), $\tau_2 = 0.75$ (classification), and $\tau_3 = 0.80$ (attribution), to reflect increasing consequence severity. The progressive threshold design, requiring higher confidence at each successive layer ($0.70 \rightarrow 0.75 \rightarrow 0.80$), aligns with safety-critical standards that mandate confidence ≥ 0.80 for high-consequence decisions.

While low-confidence detections can be filtered downstream, incorrect attributions compromise regulatory reporting. These thresholds produce layer-specific rejection rates. Layer 1 filters 4.2% of outputs, typically scenes with ambiguous regulatory compliance. Layer 2 filters 8.7% of classifications, most commonly for violations that span multiple categories. Layer 3 filters 12.1% of attributions, mostly in multi-module scenarios where the model cannot isolate a single primary component. This upward trend in rejection rates is expected because, as reasoning complexity grows, the model more frequently signals low confidence rather than producing an unreliable output. Since CLEAR is intended for offline deployment, these filtered cases are routed to human analysts for manual review, which ensures full scenario coverage without sacrificing diagnostic reliability. To verify our threshold selection, we ran 1,000 rounds of testing on randomly drawn subsets. Across all rounds, performance varied by no more than ± 0.02 F1 points, confirming that the chosen thresholds are stable.

TABLE III: Per-layer accuracies under different prompt variants and deviation from the baseline (in %).

Prompt Variant	Layer 1	Layer 2	Layer 3	Δ from Baseline
Baseline	95.2	62.3	41.7	—
Minimal	91.8	54.6	35.2	−6.5 avg
Verbose	94.7	63.1	42.3	+0.3 avg
Reordered	94.9	61.8	40.9	−0.5 avg
Paraphrased	95.0	61.5	41.2	−0.5 avg

CLEAR, when paired with **MIRAGE**, forms a comprehensive system for anomaly detection, classification, and regulatory-aligned attribution in AV perception. It enables interpretable diagnostics at the component level, addressing safety, transparency, and compliance in a unified framework. We provide the full component attribution prompt in Listing 4.

VI. Experimental Evaluation

We evaluate the **CLEAR+MIRAGE** framework across the three mentioned core tasks in AV perception. Specifically, we address the technical capability of our framework to correctly identify which perception module is affected when an anomaly occurs, and whether the reasoning chains it produces can be effectively interpreted for system diagnostics and remediation strategies. To evaluate these capabilities, we structure our experimental analysis around the following research questions:

- **RQ1 – Attribution Accuracy:** How reliably can LLMs localize anomalies to specific AV perception modules?
- **RQ2 – Reasoning Quality:** Do LLM-generated reasoning chains offer interpretable, regulation-aligned justifications suitable for safety-critical decision-making?

```
Given that an anomaly has been detected in the
  following scenario, classify the type of anomaly
  present.
Scene Description:{anomalous_description}
Anomaly Detection Results: {detection_results}
Violated Rules (if known): {violated_rules}
Specific Changes: {specific_changes}
Classify the anomaly into one of the following
  categories:
1. Direct_violation: Clear violation of traffic
  control standards (e.g., missing required signs,
  improper placement)
2. Subtle_violation: Minor deviations that could
  affect perception (e.g., faded markings, partially
   obscured signs)
3. Contextual_violation:Context-dependent violations
   (e.g., school zone signs without proper timing)
4. Environmental_violation: Environmental conditions
   creating perception challenges (e.g., glare,
  shadows, weather)
Provide your response in the following JSON format:
{{ ... }}
```

Listing 3: CLEAR Layer 2 prompt: Classifies detected anomalies into one of four types (direct, subtle, contextual, environmental) based on regulatory and visual cues.

```
You are an expert in autonomous vehicle perception
  systems analyzing which specific perception
  modules are affected by detected anomalies.

Given the classified anomaly in the following
  scenario, determine which perception module(s) are
   primarily affected.

Scene Description: {anomalous_description}
Anomaly Classification: {anomaly_classification}
Classification Reasoning: {classification_reasoning}

Perception Modules:
- TSR (Traffic Sign Recognition): Processes and
  interprets traffic signs
- ALC (Automated Lane Centering): Detects and tracks
   lane markings, road boundaries
- OD (Object Detection): Identifies and tracks
  obstacles, vehicles, pedestrians

Analyze which module(s) would be most affected by
  this anomaly and provide your response in the
  following JSON format: {{ ... }}
```

Listing 4: CLEAR Layer 3 prompt: attributes anomalies to the most affected AV perception modules.

A. Dataset Configuration

We curate a dataset of anomalous driving scenes from an initial pool of 48,022 real-world frames sourced from nuScenes, Waymo Open, and Argoverse 2. To ensure representative sampling across diverse driving conditions and dataset characteristics, we implement a systematic image selection methodology. Our sampling process operates across three dataset-specific directories containing pre-filtered images. The selection algorithm distributes the target sample size roughly equally across the three datasets, with any remainder allocated to maintain balance. The final image list is shuffled to eliminate any dataset-specific ordering bias.

Each anomaly is labeled by both its violation type (direct, subtle, contextual, environmental) and its impact on perception modules: TSR, ALC, OD. The evaluation dataset consists of two primary components. For Layer 1 (anomaly detection), we constructed a balanced set of 11,694 scenes. This includes 5,847 anomalous scenarios generated by MIRAGE paired with 5,847 unmodified scenes randomly sampled from the original 48,022 real-world frames. In contrast, evaluation for Layers 2 and 3 (classification and attribution) operates exclusively on the 5,847 anomalous scenarios (Table IV), as normal scenes contain no violations to analyze.

TABLE IV: Distribution of 5,847 annotated anomalies across violation types and perception modules.

Anomaly Type	**TSR**	**ALC**	**OD**	**Total**
Direct Violation	1,245	892	567	2,704
Subtle Violation	456	1,023	234	1,713
Contextual Violation	234	145	456	835
Environmental Violation	123	234	238	595
Total	2,058	2,294	1,495	5,847

TABLE V: Performance of Layer 1 anomaly detection in CLEAR, showing high accuracy and reliability.

Metric	**Value**	**Metric**	**Value**
Accuracy	95.2%	F1-Score	95.2%
Precision	93.8%	Avg. Confidence	0.89
Recall	96.7%	Avg. Processing Time	2.3s

B. Model Configuration

To assess reasoning capabilities, we benchmark three LLM configurations. We used `GPT-4o-mini` as our primary model to process textual and visual inputs jointly. We selected GPT-4o-mini as our primary model to balance reasoning accuracy with cost-efficiency at scale. This choice aligns with recent findings that smaller models can match larger ones on structured reasoning tasks at significantly reduced cost [39]. At evaluation scale, processing 5,847 anomalous scenes across three reasoning layers requires 17,541 API calls, making frontier reasoning models (e.g., GPT-4o, o1) 10–30× more expensive. We also evaluate `Phi-4` and `Phi-4-mini`, both optimized for structured reasoning but differing in computational footprint. All models are run with temperature = 0.7 and top-p = 0.9 to balance determinism and diversity. Top-p limits the model's word choices to only the most likely options whose combined probability reaches the top-p threshold, filtering out unlikely outputs while still allowing enough variety to produce natural, well-reasoned responses.

CLEAR integrates intermediate representations such as keyphrases and structured JSON outputs to link detection, classification, and attribution. **MIRAGE** builds upon these representations to generate compliant outputs aligned with MUTCD standards, ensuring that each anomaly explanation is traceable and auditable by regulatory stakeholders.

C. Evaluation Metrics

We use quantitative and qualitative metrics tailored to each layer of the framework. For Layer 1 (Anomaly Detection), we use binary classification metrics, including accuracy, precision, recall, and F1-score. For Layer 2 (Violation Classification), we assess multiclass performance across four categories: direct, subtle, contextual, and environmental. Precision and recall are calculated for each class by treating each violation as positive with the remaining three as negative. These values are then averaged across all four categories. For layer 3 (Component Attribution), we report accuracy based on both the perception module and violation type to analyze where attribution succeeds and where overlapping module responsibilities create ambiguity. For all layers, we evaluate only outputs that pass both confidence gating and schema validation.

In addition, we assess reasoning quality using four human evaluated dimensions. Completeness measures whether the reasoning output includes all required elements specified in each layer's prompt template, such as anomaly indicators for Layer 1, violation justifications for Layer 2, and failure mode descriptions for Layer 3. We calculate this as the percentage

of responses that contain all mandatory elements. Consistency evaluates whether the reasoning supports the predicted output without contradictions, scored as the percentage of reasoning chains where explanations align with predictions and evidence supports the conclusions. MUTCD Citation Accuracy (only for Layers 2 and 3) verifies that regulatory references are correct and relevant to the scenario, calculated as the ratio of accurate citations to total citations made. Explanation clarity provides a qualitative assessment of reasoning comprehensibility, rated as High, Medium, or Low based on logical flow, appropriate technical terminology, and absence of ambiguity.

D. Layer 1 Results: Anomaly Detection

CLEAR demonstrates strong performance in detecting anomalies across a wide variety of driving scenes. As shown in Table V, the model achieves an overall accuracy of 95.2%, with a precision of 93.8% and recall of 96.7%.

```
{
 "scene_id": "waymo_sf_0234",
 "anomaly_detected": true,
 "confidence": 0.92,
 "reasoning": "The scene shows a STOP sign
 placed at ground level near the intersection,
 violating MUTCD section 2B.04 which requires
 proper mounting height of 7 feet.",
 "anomaly_indicators": ["improper sign
     placement", "height violation"]
}
```

Listing 5: Layer 1 output example (waymo_sf_0234)

The F1-score of 95.2% confirms the robustness of **CLEAR**'s binary detection layer. The system outputs a mean confidence score of 0.89, suggesting high certainty in its predictions. Listing 5 illustrates a typical Layer 1 output, where the model identifies a STOP sign height violation with 92% confidence and provides specific regulatory justification citing MUTCD section 2B.04. While Layer 1 achieves a strong detection accuracy, this is not the primary contribution of our work. CLEAR's contribution begins where these existing methods stop. Layer 1 serves as the entry point to a reasoning pipeline that provides violation classification in Layer 2 and component attribution with regulatory justification in Layer 3, which are capabilities no existing method currently offers.

E. Layer 2 Results: Violation Classification

In the second layer, **CLEAR** classifies detected anomalies into four regulatory categories. As summarized in the confusion matrix in Table VI, the system performs best on direct violations (DV), achieving 69.2% of correct predictions. These anomalies typically feature clear and explicit rule violations such as sign placement, which the model identifies with high fidelity. Overall, Layer-2 multi-class accuracy is 62.3%. We noted that most residual errors arise in context-dependent cases, such as temporary work or legacy sign designs.

Subtle violations (SV), including occluded signage, are also reliably classified, with 61.3% of correct predictions, showing the model's capability to detect safety-relevant degradations.

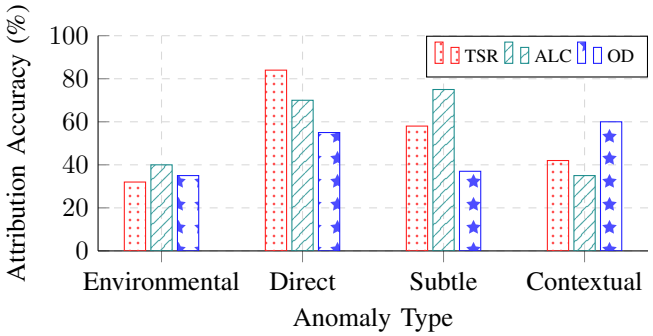

Fig. 7: Attribution accuracy across anomaly types and perception modules: Direct violations yield the highest attribution performance for TSR modules (84%). In contrast, environmental anomalies are more ambiguous and affect multiple modules, leading to the lowest per-module accuracy.

Contextual violations (CV) show moderate performance, with 54.6% of correctly classified cases. These often involve conditional reasoning, such as temporal signage (e.g., school zones) that depends on time-of-day or traffic context.

Environmental violations (EV), showed the lowest per-module attribution accuracy (32–40% across individual modules; 21.7% overall for this violation type; see Figure 7 and Table VII), and the overall Layer-3 accuracy is 41.7%. We analyze this further in Table VI-F through a Top-2 evaluation and confidence distribution analysis. Environmental anomalies create diffuse, multi-modal effects that violate the single-point-of-failure assumption underlying component attribution. Fog simultaneously degrades TSR's contrast sensitivity, ALC's edge detection, and OD's depth estimation, making "primary" attribution difficult. Unlike direct violations with explicit MUTCD specifications (e.g., "Stop signs shall be octagonal"), environmental conditions lack precise regulatory definitions. MUTCD provides only high-level guidance like "maintain visibility," without quantitative thresholds for fog density or glare intensity. This gap between environmental physics and regulatory language prevents effective alignment. Moreover, environmental factors such as fog, glare and shadows, simultaneously affect all perception modules, making single-module attribution inherently ambiguous.

TABLE VI: Layer-2 Classification: Normalized Confusion Matrix.

Actual \ Predicted	DV	SV	CV	EV
DV	1.0 (823)	0.29 (245)	0.10 (89)	0.03 (32)
SV	0.22 (156)	1.0 (687)	0.34 (234)	0.06 (45)
CV	0.09 (45)	0.26 (123)	1.0 (456)	0.19 (89)
EV	0.09 (23)	0.19 (45)	0.28 (67)	1.0 (234)

F. Layer 3 Results: Component Attribution

In Layer 3, **CLEAR** attributes each detected anomaly to one or more perception modules. Figure 7 illustrates attribution accuracy across modules and violation types. Attribution accuracy correlates strongly with how clearly a violation maps to a single module. Each module achieves its best accuracy on their specific violation type: Traffic Sign Recognition (TSR) reaches 84% on direct violations, where sign-related rules map unambiguously to sign recognition; Automated Lane Centering (ALC) reaches 75% on subtle violations, where lane degradation falls within its primary scope; and Object Detection (OD) hits 60% on contextual violations, where dynamic environmental elements align with object detection.

Accuracy typically declines when violations cross module boundaries. For instance, a faded stop sign involves both TSR for sign readability and ALC for intersection context, creating ambiguity that reduces TSR's accuracy to 58% in these cases. Environmental violations represent the most difficult category because they affect more than one module simultaneously.

Direct violations (DV) show the strongest attribution signals, with 84% accuracy for TSR, 70% for ALC, and 55% for OD. These violations are typically unambiguous and well-mapped to regulatory rules, making them easier to attribute. Subtle violations (SV) show strong attribution to ALC (75%) due to sensitivity to lane degradation. TSR performs moderately (58%), while OD yields a lower attribution rate of 37%. Contextual violations (CV) are best attributed to OD (60%), consistent with its role in handling dynamic objects. TSR and ALC underperform (42% and 35%), reflecting difficulty in mapping contextual rules to fixed infrastructure. Environmental violations (EV) remain the hardest to localize, with accuracy ranging from 32% (TSR) to 40% (ALC). These often impact multiple subsystems simultaneously.

Ranked Attribution and Confidence Analysis. The strict Top-1 evaluation penalizes scenarios where multiple modules are affected. To better characterize attribution behavior, we extended the Layer 3 JSON schema to produce a ranked list of affected modules with confidence scores rather than a single primary module. This enables two complementary analyses: a Top-2 accuracy metric, where a prediction is correct if the ground-truth module appears among the two highest-ranked outputs, and a per-module confidence distribution summarized by normalized Shannon entropy (H). Entropy of 1.0 indicates perfectly uniform confidence across all three modules, while values near 0.0 indicate concentration on a single module.

TABLE VII: Top-2 Attribution Accuracy. Comparison of strict single-module (Top-1) and relaxed (Top-2) attribution accuracy across violation types. Recovery indicates the percentage of Top-1 errors corrected by allowing a second-ranked module.

Violation Type	Top-1 (%)	Top-2 (%)	Recovery (%)
Direct	69.7	83.2	45.2
Subtle	56.7	78.4	50.1
Contextual	52.2	78.8	55.8
Environmental	21.7	75.0	68.1
Overall	41.7	74.6	56.4

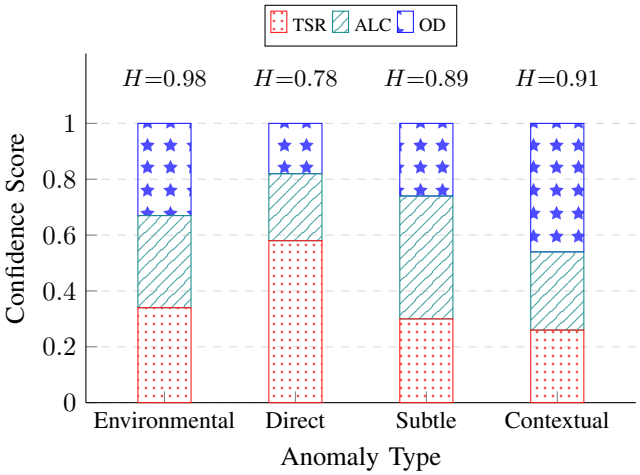

Fig. 8: Per-module confidence distribution across violation types. Each bar shows the mean confidence CLEAR assigns to TSR, ALC, and OD. Normalized entropy (H) is annotated to the right: values near 1.0 indicate uniform confidence across all modules (multi-module impact), while lower values indicate concentration on a single module.

Table VII shows Top-2 results. Overall accuracy rises from 41.7% to 74.6%. The recovery column measures what fraction of Top-1 errors are corrected by allowing a second candidate. For environmental violations, 68.1% of previously incorrect cases become correct under Top-2, meaning CLEAR typically ranked the correct module second. Since only three modules exist, random Top-2 selection would achieve 66.7%. To show that CLEAR's performance is not simply a consequence of this, we analyze the confidence scores across modules.

While Figure 7 reports *attribution accuracy* (whether the Top-1 prediction matched ground truth), Figure 8 shows the mean confidence CLEAR assigns to TSR, ALC, and OD for each violation type, with normalized entropy annotated. For direct violations, confidence concentrates on one module (H=0.78) and recovery is low (45.2%). For environmental violations, confidence is near-uniform (H=0.98) and recovery is high (68.1%). This shows that if the model were assigning confidence arbitrarily, entropy and recovery would be uniform across violation types. Instead, they co-vary systematically: high entropy predicts high recovery, and low entropy predicts low recovery. This confirms that CLEAR distributes confidence based on genuine multi-module impact, not random selection. The near-uniform confidence for environmental violations signals system-wide degradation rather than a single-module fault, which is the correct characterization of how fog, glare, or heavy rain degrade AV perception.

Reasoning Quality Analysis. We conduct a manual evaluation of 50 randomly sampled outputs per layer (150 total across all three layers) [40]. Each sample is scored on three dimensions:

completeness, the rationale addresses every reasoning element required by the layer prompt, such as hazard identification, rule invocation, and module responsibility; consistency, the rationale is internally coherent and matches the final prediction or attribution; and regulatory citation accuracy, MUTCD references are correct, including part or section numbers and applicability to the scene. Completeness and consistency of reasoning decrease from Layer 1 through Layer 3, see Table VIII, reflecting increasing complexity in justifying module-level attribution as scenarios grow more complex and introduce cross-module dependencies. Our results show that while `Phi-4-mini` offers the fastest inference, `GPT-4o-mini` provides better accuracy-efficiency tradeoffs for component-level attribution. Regulatory citation accuracy remains strong overall, with the highest performance observed in Layer 2.

VII. Discussion

Our evaluation yields several insights into component-level anomaly attribution in AV perception systems:

Hierarchical Complexity. As shown in Figure 9, the performance under strict single-module evaluation declines across layers (95.2% $\rightarrow$ 62.3% $\rightarrow$ 41.7%), reflecting the increasing difficulty of transitioning from binary detection to attribution. However, our Top-2 Analysis (Table VII) shows that much of this decline stems from multi-module ambiguity rather than reasoning failure, with attribution rising to 74.6% when allowing a second-ranked module. The remaining gap reflects the genuine complexity of isolating a single affected module when violations cross component boundaries.

Impact of Anomaly Type. Attribution accuracy is highly sensitive to the type of anomaly. Direct violations yield the highest accuracy (84%) for TSR, whereas environmental violations are the most challenging (32%), due to their diffuse and multi-modal impact. Our Top-2 evaluation and confidence analysis (Table VII, Figure 8) quantify this observation. As shown in Section VI-F, the correlation between confidence entropy and Top-2 recovery confirms that apparent misattributions reflect genuine multi-module impact rather than reasoning failure.

Reasoning Transparency. Explicit reasoning chains offer interpretability benefits beyond raw performance. For example, misattributions such as assigning lane-marking anomalies to OD instead of ALC, often stem from ambiguous boundary definitions between components. These explanations surface latent failure modes, guiding debugging and system design.

Beyond core performance metrics, CLEAR provides several practical advantages over traditional detection frameworks:

Regulatory Alignment. With 78.6% accuracy in citing relevant MUTCD regulations, the system demonstrates promising alignment with traffic safety standards. This suggests potential for regulatory-auditable AI outputs.

Diagnostic Utility. Even when misattribution occurs, per-module confidence scores reveal which components share exposure to the anomaly. For example, a TSR failure caused by an occluded sign may be misattributed to OD. This happens because both modules are affected by the same upstream signal (poor visibility of the sign), and the confidence distribution reflects this shared exposure.

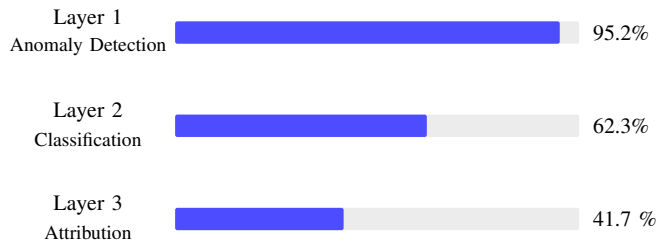

Fig. 9: Cascading performance across CLEAR's layers under strict single-module evaluation.

A. Limitations

Deployment Feasibility. The average processing time of 3.1 seconds per scene positions CLEAR for offline forensic workflows rather than real-time intervention. This focus aligns with its intended deployment in post-incident analysis where investigators review recorded data, validation pipelines for auditing perception behavior, and safety audits where regulators verify compliance. In these contexts, thoroughness and interpretability are more critical than immediate latency.

Attribution Granularity. Our default evaluation uses single-label attribution, which forces environmental anomalies into a single module despite their multi-modal impact. As shown in Table VI-F, the correlation between confidence entropy and Top-2 recovery (Figure 8) demonstrates that CLEAR ranks modules based on genuine evidence of impact rather than arbitrary selection. Nevertheless, the limitation remains: modern OD architectures are trained end-to-end, creating entangled representations where sub-module boundaries become semantically blurred, making isolated attribution difficult.

Structured Output Reliability. While reasoning and attribution outputs, are typically correct, deviations from the specified format complicate automated evaluation and downstream integration. This structured output reliability issue affects approximately 15% of responses across all reasoning layers, manifesting as incomplete JSON objects, missing required fields, or extraneous text outside the schema boundaries. To mitigate this, CLEAR employs a JSON schema validation step after each layer's output. Malformed responses are automatically re-queried up to three times with the same prompt. This retry mechanism resolves the majority of formatting issues: of the 15% requiring at least one retry, fewer than 2% fail all three attempts. Persistently malformed outputs are flagged for human review and excluded from automated evaluation.

LLM Limitations. Our approach inherits limitations of current LLMs, including hallucinations, overconfidence in uncertain scenarios, and difficulty with counterfactual reasoning.

Regulatory Coverage. Our regulatory grounding is currently limited to MUTCD standards, restricting applicability to U.S. traffic scenarios. As the EU AI Act and other international regulations increasingly demand component-level accountability for autonomous vehicles, this limitation becomes significant

TABLE VIII: Reasoning quality across three layers based on manual evaluation of 50 randomly sampled scenarios.

Quality Dimension	Layer 1	Layer 2	Layer 3
Completeness	94%	82%	71%
Consistency	91%	76%	68%
MUTCD Citation Accuracy	N/A	88%	73%
Explanation Clarity	High	Medium	Medium

for global deployment. However, this constraint is practical rather than fundamental. The **MIRAGE** pipeline is jurisdiction agnostic and can be instantiated with the rule corpus of the target geography, meaning we could substitute the MUTCD with the corresponding national or municipal traffic-control standards, enabling portability beyond the US.

B. Future Work

Current reasoning chains are technically accurate but may prove challenging for non-technical stakeholders such as regulators, legal professionals, and insurance adjusters. Future research could address five key areas.

First, subsequent efforts could focus on developing multi-level explanations tailored to audience expertise. These might range from high-level executive summaries for regulators to detailed technical breakdowns for engineering teams.

Second, there is potential for finer-grained attribution. Our Top-2 evaluation and per-module confidence analysis represent a first step toward multi-label attribution. Future iterations could formalize this into a full probability distribution over affected modules with calibrated thresholds, rather than the ranked list currently produced. This would require updating dataset annotations, modifying the DSL schema to support weighted outputs, systematic comparison against direct-prompting and retrieval-augmented baselines to quantify the contribution of each pipeline component, and adapting the reasoning pipeline to handle multi-label ground truth.

Third, the development of visualization tools could enhance interpretability. These interfaces would allow stakeholders to explore attribution reasoning by navigating from high-level component identification down to specific evidence chains.

Fourth, establishing standardized explanation formats remains a priority. These formats must align with emerging international frameworks for regulatory compliance while maintaining the necessary technical rigor.

Fifth, several pathways exist to reduce the computational footprint of CLEAR and minimize latency:

1) **Knowledge Distillation:** The current pipeline relies on `GPT-4o-mini` for reasoning. However, **CLEAR's** structured outputs provide high-quality labels for training smaller, specialized models.
2) **Caching and Retrieval:** Future systems could cache scene-rule similarity computations. By using approximate nearest neighbor search, similar scenes can retrieve precomputed attributions. This reduces redundant LLM calls for recurring scenario patterns.
3) **Speculative Execution:** Layers 2 and 3 could begin processing optimistically while Layer 1 computes confidence scores. If the confidence falls below the threshold, the system discards the results. This approach reduces effective latency for high-confidence cases.
4) **Batched Inference:** The current 3.1s latency reflects sequential API calls. Implementing batched LLM inference and parallel layer execution could reduce per-scene latency to under one second.

VIII. Conclusion

We presented **CLEAR**, a novel framework for component-level explainable anomaly reasoning in autonomous vehicle perception systems that addresses critical gaps in current AV safety analysis. Our key contributions include: (1) **MIRAGE**, the first dataset generation methodology creating semantically rich anomalous scenarios with ground-truth attribution by integrating 48,022 real-world scenes with MUTCD regulatory knowledge; (2) a three-layer LLM-based reasoning pipeline achieving 95.2% anomaly detection accuracy, 84% attribution accuracy for direct violations, and 74.6% overall accuracy under Top-2 evaluation; (3) interpretable reasoning chains satisfying regulatory transparency requirements from the EU AI Act and NHTSA guidelines; and (4) validation that structured LLMs can provide reliable component-level diagnostics for safety-critical systems. **CLEAR** addresses increasing regulatory demands for transparency and accountability as autonomous vehicles approach widespread deployment. The framework enables targeted remediation by identifying which specific perception module (TSR, ALC, or OD) failed and why, rather than generic system-wide alerts. While environmental anomalies remain challenging due to multi-modal impacts, our approach provides a foundation for systematic safety analysis and regulatory compliance. **CLEAR** represents the first integrated system unifying detection, attribution, explainability, and regulatory compliance for AV perception diagnostics. Beyond autonomous vehicles, our methodology for combining regulatory knowledge with LLM-based reasoning has applications in safety-critical AI systems including medical devices and aviation. As AI systems become increasingly integrated into safety-critical applications, **CLEAR** establishes a new paradigm for explainable AI that builds public trust through systematic understanding of failure modes and targeted corrective action.

Acknowledgment

This work was supported in part by a grant from The BMW Group, and in part by Clemson University's Virtual Prototyping of Autonomy Enabled Ground Systems (VIPR-GS), under Cooperative Agreement W56HZV-21-2-0001 with the US Army DEVCOM Ground Vehicle Systems Center (GVSC). DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited. OPSEC #10607.

REFERENCES

[1] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz *et al.*, "Self-driving cars: A survey," *Expert systems with applications*, vol. 165, p. 113816, 2021.

[2] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," pp. 2722–2730, 2015.

[3] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, vol. 37, p. 100270, 2020.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[5] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, and B. Li, "Adversarial objects against lidar-based autonomous driving systems," 2019. [Online]. Available: https://arxiv.org/abs/1907.05418

[6] National Highway Traffic Safety Administration, "Level 2 adas crash totals (20 jul 2021–15 jul 2024)," Standing General Order dataset, 2024, accessed 29 May 2025.

[7] ——, "Level 2 adas crash totals through 15 apr 2025," Standing General Order dataset, 2025, accessed 29 May 2025.

[8] U. S. DOT Highly Automated Systems Safety Center of Excellence, "Understanding safety challenges of vehicles equipped with ads," Technical report, 2024, accessed 29 May 2025.

[9] European Commission, "Artificial Intelligence Act (EU Regulation 2024/1689)," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, 2024, accessed: 2025-05-27.

[10] U.S. National Highway Traffic Safety Administration, "Automated Vehicles Transparency and Engagement for Safe Testing (AV TEST) Initiative," 2023. [Online]. Available: https://www.nhtsa.gov/press-releases/nhtsa-proposes-national-program-vehicles-automated-driving-systems

[11] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, Mar. 2021. [Online]. Available: https://doi.org/10.1145/3439950

[12] S. Nazat, L. Li, and M. Abdallah, "Xai-ads: An explainable artificial intelligence framework for enhancing anomaly detection in autonomous driving systems," *IEEE Access*, vol. 12, pp. 48 583–48 607, 2024.

[13] S. Feng, X. Yan, H. Sun, Y. Feng, and H. X. Liu, "Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment," *Nature communications*, vol. 12, no. 1, p. 748, 2021.

[14] S. Fourati, W. Jaafar, N. Baccar, and S. Alfattani, "Xlm for autonomous driving systems: A comprehensive review," 2024. [Online]. Available: https://arxiv.org/abs/2409.10484

[15] Z. Wan, Y. Huai, Y. Chen, J. Garcia, and Q. A. Chen, "Towards automated driving violation cause analysis in scenario-based testing for autonomous driving systems," 2024. [Online]. Available: https://arxiv.org/abs/2401.10443

[16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020. [Online]. Available: https://arxiv.org/abs/1903.11027

[17] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[18] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," 2023. [Online]. Available: https://arxiv.org/abs/2301.00493

[19] Federal Highway Administration, "Manual on uniform traffic control devices (mutcd)," https://mutcd.fhwa.dot.gov/, 2025, accessed: 2025-04-15.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=_VjQlMeSB_J

[21] P. MohajerAnsari, A. Salarpour, D. Fernandez, C. Kokenoz, B. Li, and M. D. Pesé, "Attention-aware temporal adversarial shadows on traffic sign sequences," in *The 5th Workshop of Adversarial Machine Learning on Computer Vision: Foundation Models + X*, 2025.

[22] D. Fernandez, P. MohajerAnsari, A. Salarpour, L. Cheng, A. Razi, and M. D. Pesé, "Comparative analysis of patch attack on vlm-based autonomous driving architectures," 2026. [Online]. Available: https://arxiv.org/abs/2603.08897

[23] P. Jing, Q. Tang, Y. Du, L. Xue, X. Luo, T. Wang, S. Nie, and S. Wu, "Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 3237–3254. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/jing

[24] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:207947087

[25] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. V. Poor, "Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey," *ArXiv*, vol. abs/2303.06302, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257496029

[26] N. Habib, Y. Cho, A. Buragohain, and A. Rausch, "Towards exploring adversarial learning for anomaly detection in complex driving scenes," Springer, pp. 35–55, 2023.

[27] J. R. V. Solaas, E. Mariconti, and N. Tuptuk, "Systematic literature review: Anomaly detection in connected and autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 1, pp. 43–58, 2025.

[28] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. MA, Y. Li, L. XU, D. Shang, Z. Zhu, S. Sun, Y. BAI, X. Cai, M. Dou, S. Hu, B. Shi, and Y. Qiao, "On the road with GPT-4v(ision): Explorations of utilizing visual-language model as autonomous driving agent," 2024. [Online]. Available: https://openreview.net/forum?id=2UBexKm8TE

[29] D. Fernandez, P. MohajerAnsari, A. Salarpour, and M. D. Pesé, "Avoiding the crash: A vision-language model evaluation of critical traffic scenarios," SAE Technical Paper, Tech. Rep., 2025.

[30] D. Fernandez, P. MohajerAnsari, C. Kokenoz, A. Salarpour, B. Li, and M. D. Pesé, "Wip: From detection to explanation: Using llms for adversarial scenario analysis in vehicles," in *Proceedings of the 3rd USENIX Symposium on Vehicle Security and Privacy (VehicleSec '25)*. USENIX Association, 2025. [Online]. Available: https://www.usenix.org/conference/vehiclesec25/presentation/fernandez

[31] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *CoRR*, vol. abs/2112.11561, 2021. [Online]. Available: https://arxiv.org/abs/2112.11561

[32] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, p. 362–386, Nov. 2019. [Online]. Available: http://dx.doi.org/10.1002/rob.21918

[33] S. Bhattacharjee, M. J. Islam, and S. Abedzadeh, "Robust anomaly based attack detection in smart grids under data poisoning attacks," in *Proceedings of the 8th ACM on Cyber-Physical System Security Workshop*, ser. CPSS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3–14. [Online]. Available: https://doi.org/10.1145/3494107.3522778

[34] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[35] X. Huang, E. M. Wolff, P. Vernaza, T. Phan-Minh, H. Chen, D. S. Hayden, M. Edmonds, B. Pierce, X. Chen, P. E. Jacob, X. Chen, C. Tairbekov, P. Agarwal, T. Gao, Y. Chai, and S. Srinivasa, "DriveGPT: Scaling autoregressive behavior models for driving," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff,

and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 25 908–25 921. [Online]. Available: https://proceedings.mlr.press/v267/huang25ak.html
[36] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
[37] M. Grootendorst, "Keybert: Minimal keyword extraction with bert," https://github.com/MaartenGr/KeyBERT, 2020, accessed: 2025-05-30.
[38] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han, "Vila: On pre-training for visual language models," 2024. [Online]. Available: https://arxiv.org/abs/2312.07533
[39] P. Bauerfeind, A. Salarpour, D. Fernandez, P. MohajerAnsari, J. Reschke, and M. D. Pesé, "David vs. goliath: A comparative study of different-sized llms for code generation in the domain of automotive scenario generation," 2025. [Online]. Available: https://arxiv.org/abs/2510.14115
[40] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685