github.com/David-FR/MIRAGE_CLEAR

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

Talk Roadmap

Five sections, one core contribution



We attribute perception failures to specific modules with regulatory-grounded reasoning.

AV Perception Systems

A modular stack of specialized subsystems



TSR

Traffic Sign Recognition

Reads regulatory signs; triggers braking, stopping, etc.



ALC

Automated Lane Centering

Detects lane boundaries for lateral control.



OD

Object Detection

Localizes obstacles, vehicles, and pedestrians.

Modularity = engineering benefit + *the source of the attribution problem.*

The Regulatory Landscape

Compliance now requires subsystem observability

NHTSA AV STEP

U.S. transparency proposal

EU AI Act 2024/1689

High-risk AI mandate

What regulators now demand

- Transparency in high-risk AI systems
- Subsystem observability
- Component-level traceability
- Identification of failed module + rationale

Generic high-level alerts ("the system failed") are no longer sufficient.
Regulators want the failed component and the failure rationale.

We Can Detect Failures — But We Can't Trace Them

Disengagement causes → clearly identified



4%

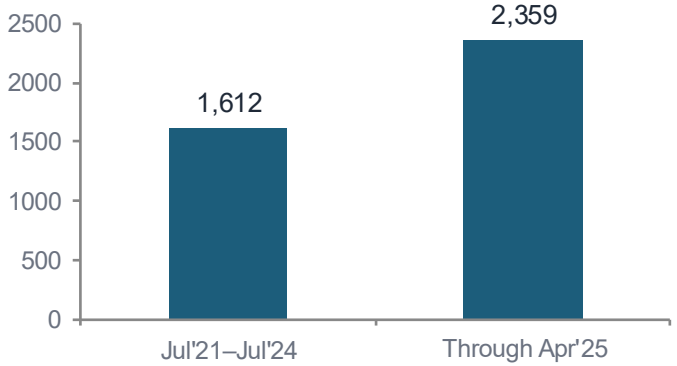
of disengagements have a clearly identified cause.

17% of disengagements are caused by perception failures.

What is a disengagement?

Moment when the autonomous system transfers control back to the human driver, typically because it cannot safely proceed on its own.

Level-2 ADAS crashes (NHTSA)



Crashes are rising. Attribution capability isn't.

The Core Problem: The Attribution Gap

Existing methods detect THAT, not WHERE

Method	Anomaly Detection	Component Attribution	Explainability	Regulatory Compliance
Statistical Methods	✓	✗	✗	✗
Deep Learning AD	✓	✗	✗	✗
Rule-based Systems	✓	◐	✓	◐
Vision Language Models	✓	✗	✓	✗
Counterfactual Analysis	✓	✓	✗	✗
★ **CLEAR (ours)**	✓	✓	✓	✓

Legend: ✓ Full - ◐ Partial - ✗ None

Key insight

Symptoms rarely localize to a single module.

Fog example

Fog simultaneously degrades:

- TSR contrast sensitivity
- ALC edge detection
- OD depth estimation

Why This Is Hard

One scene → multiple module symptoms

SCENARIO



A faded STOP sign at a foggy intersection

Single environmental condition.
Multiple module-level symptoms.

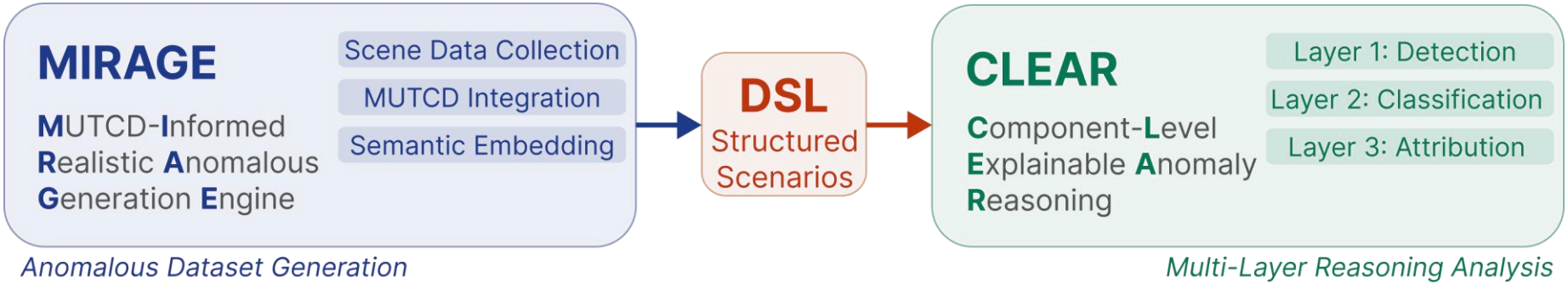
→ **TSR**
Can't read the faded sign clearly.

→ **ALC**
Can't see lane edges through fog.

→ **OD**
Can't estimate depth to other vehicles.

Framework Overview

Two-part system: dataset + reasoning



MIRAGE

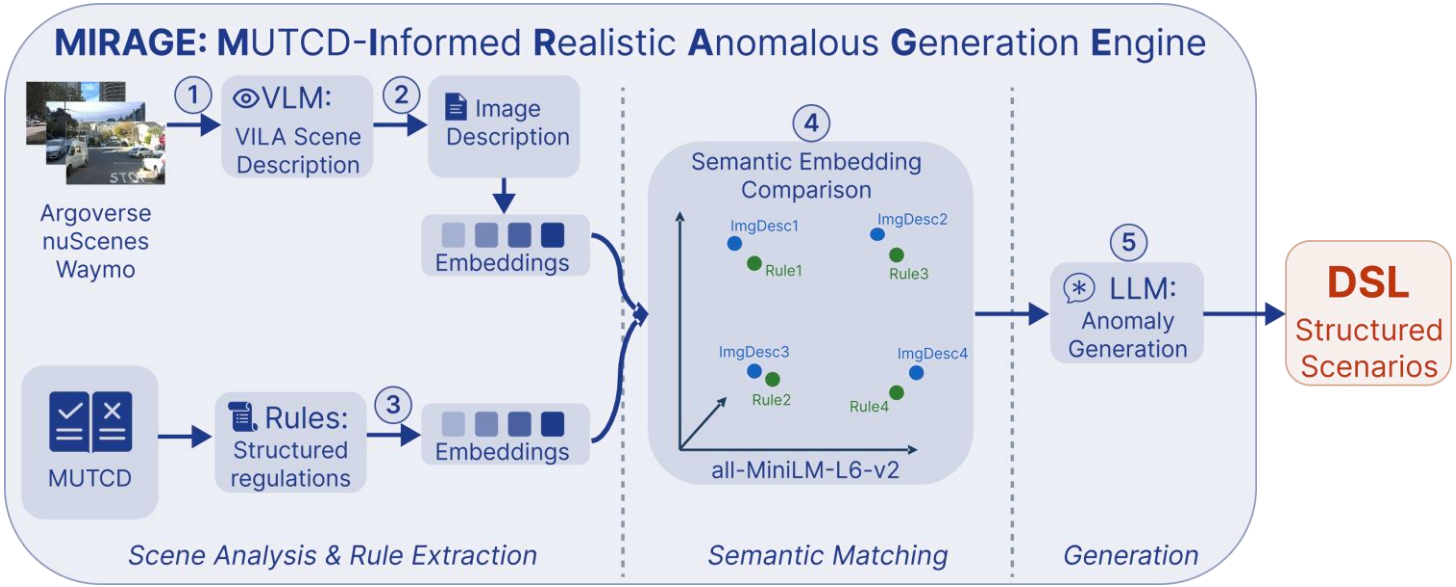
Generates the policy-grounded anomalous dataset.

CLEAR

Reasons over scenes to attribute failures.

MIRAGE: Dataset Generation Pipeline

From real driving scenes to annotated anomalies



48,022
real-world scenes

3
datasets (nuScenes/Waymo/Argo2)

5,847
annotated anomalies

100%
U.S. scenes (MUTCD aligned)

Grounding in MUTCD

Verifiable ground truth, not subjective labels

Why MUTCD?

U.S. national standard for traffic signs, signals, and road markings.

Rule categories

- **Standard:** Legally mandatory
- **Guidance:** Recommended practice
- **Option:** Conditional flexibility
- **Support:** Non-binding context

Four violation types injected

Direct

Missing required sign; improper placement.

Subtle

Faded markings; partially occluded signs.

Contextual

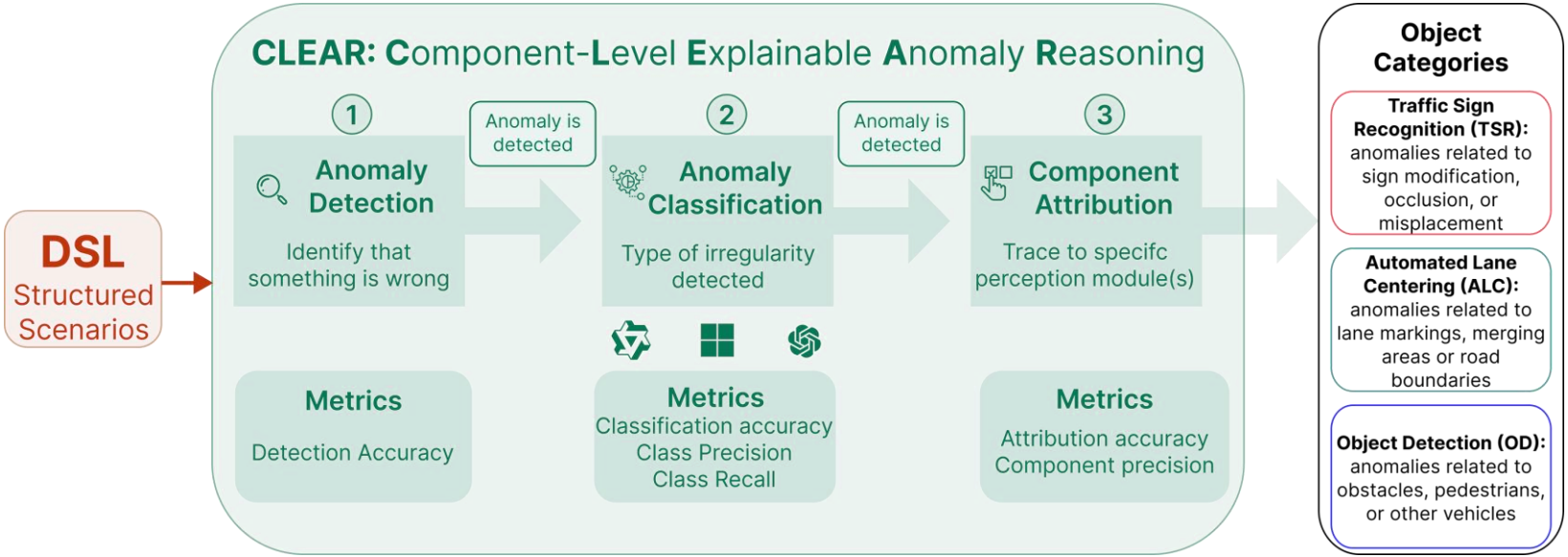
School-zone timing without warning.

Environmental

Glare, fog, deep shadows, heavy rain.

CLEAR: Hierarchical Reasoning Architecture

Three layers of structured reasoning



1 **LAYER 1 · Detection**
Anomaly present?

2 **LAYER 2 · Classification**
Which violation type?

3 **LAYER 3 · Attribution**
Which module(s)?

Key Architectural Innovations

Why this isn't just prompt engineering

Confidence-gated propagation



Higher confidence required at each successive layer.
Reasoning halts on low confidence → prevents error cascades.

Schema-constrained outputs

Outputs restricted to fixed taxonomies and JSON schemas → minimizes free-form hallucinations in reasoning descriptions.

Architecture, not prompts

Performance varies only ±1.5% across prompt variations.

→ The LLM is the reasoning engine

Together: hierarchical decomposition + confidence gating + schema constraints = reliable, interpretable forensics.

Threat Model

Realistic adversaries, observable failures

Attacker goals

Stealth
Imperceptible to humans

Targeted misperception
Single module fooled

Cross-module evasion
Other modules unaffected

Regulatory ambiguity
MUTCD-plausible designs

Constraints / assumptions

- **Physical access only —** stickers; no internal access to models
- **MUTCD-plausible —** looks like valid traffic control
- **Black-box knowledge —** knows architecture, not parameters
- **Sensor visibility —** within standard FoV / resolution
- **Immutable reasoning —** can't tamper with LLM or rule DB

Experimental Setup

Anomaly distribution and model configuration

Violation Type	TSR	ALC	OD	Total
Direct	**980**	320	410	**1,710**
Subtle	450	**720**	380	**1,550**
Contextual	380	290	**540**	**1,210**
Environmental	460	450	467	**1,377**
Total	**2,270**	**1,780**	**1,797**	**5,847**

Shaded cells = dominant module per violation type.
Environmental violations affect all three modules nearly equally. This previews the multi-module attribution story.

Models evaluated

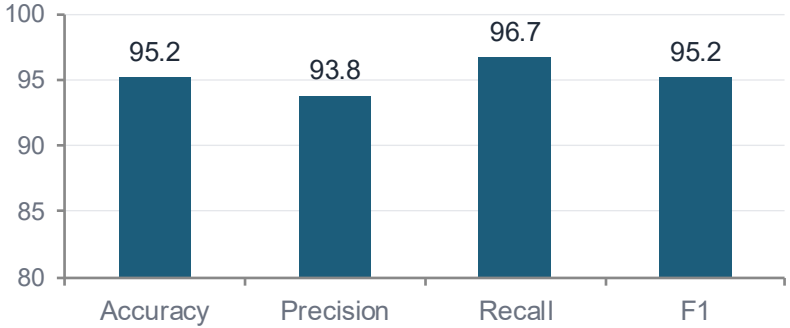
- **GPT-4o-mini** *(primary)*
- **Phi-4** *(structured reasoning)*
- **Phi-4-mini** *(fastest)*

11,694	5,847	17,541
L1 scenes	L2/L3 scenes	API calls

Layer 1 & Layer 2 Results

Detection is solid; classification reveals boundary cases

LAYER 1 — Detection



95.2%
detection accuracy

Avg. confidence: 0.89
Processing: 2.3s/scene

LAYER 2 — Classification

True ↘ Pred.	Direct	Subtle	Contextual	Environmental
Direct	**69.2%**	13.5%	10.5%	6.8%
Subtle	**15.0%**	**60.0%**	**18.0%**	7.0%
Contextual	10.5%	**18.0%**	**58.0%**	13.5%
Environmental	6.5%	12.0%	**19.5%**	**62.0%**

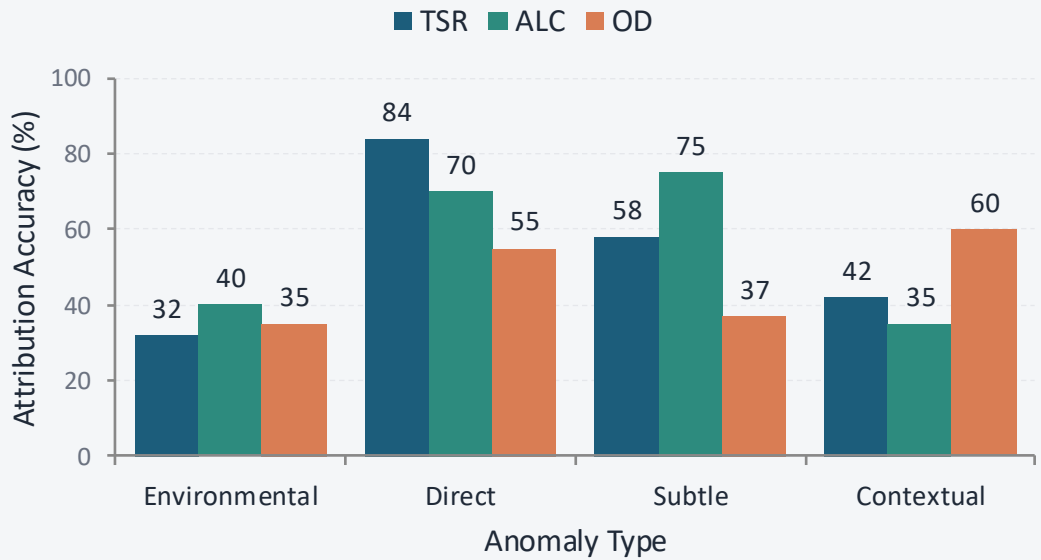
Row-normalized confusion matrix for Layer 2 violation classification.
Rows = ground-truth class, columns = predicted class; diagonal cells are correct predictions.

Intensity scale: ≥ 50% (correct) · ≥ 15% · 5–15% · < 5%

Main confusion: Subtle ↔ Contextual (18.0 / 18.0), and Environmental → Contextual (19.5).
Errors concentrate on category boundaries — foreshadows multi-module attribution.

Layer 3 Results: Component Attribution

The contribution no prior method offers



Attribution accuracy across anomaly types and perception modules.
Direct violations yield the highest performance for TSR (84%); environmental anomalies affect multiple modules, lowering per-module accuracy.

84%

attribution accuracy on direct TSR violations

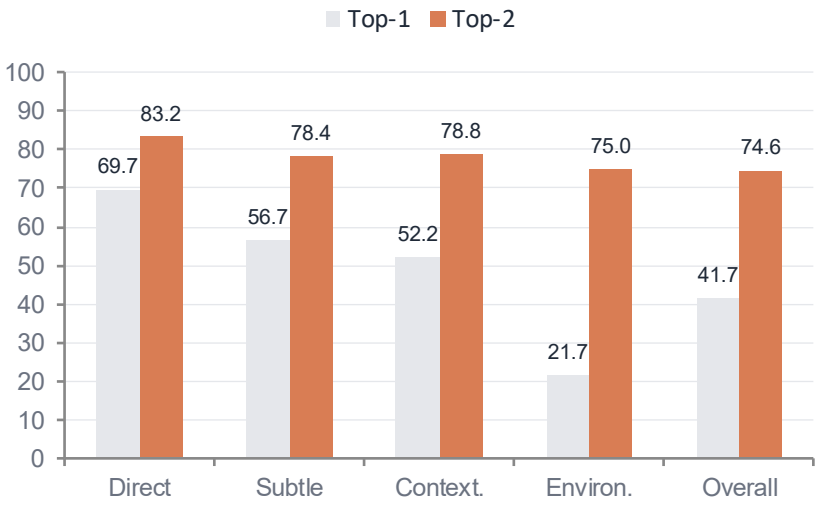
Best-per-module accuracy

- **TSR 84%** on direct violations
- **ALC 75%** on subtle violations
- **OD 60%** on contextual violations

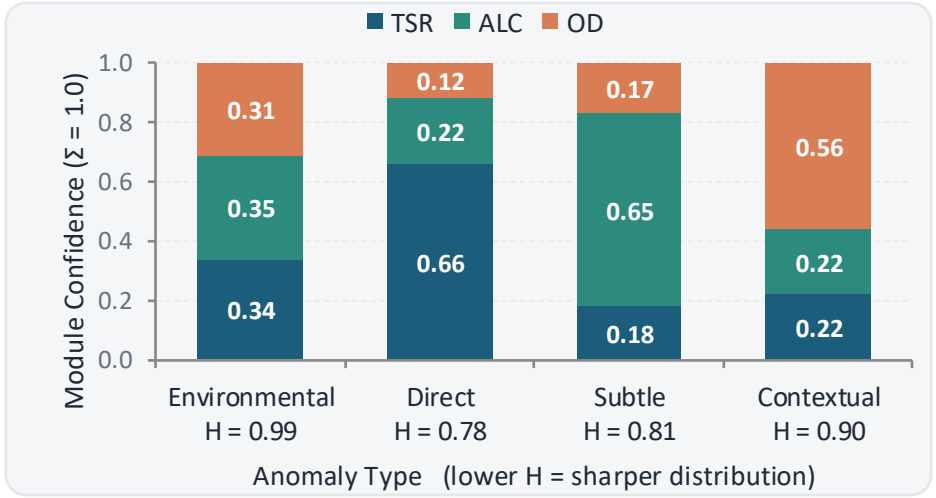
Reframing the 41.7%: Top-2 + Confidence Analysis

Misattributions reflect physics, not reasoning failure

Top-1 vs Top-2 attribution accuracy



Top-2 raises overall accuracy to **74.6%** and **75.0%** for environmental.



Key insight

High entropy (H = 1) → uniform across modules → high recovery.
Low entropy (H = 0) → all confidence on one module → low recovery.

Direct → confidence concentrates on TSR (low H = sharp).
Environmental → confidence ≈ uniform (high H ≈ 1) → correct characterization of multi-module impact.

CLEAR distributes confidence based on genuine multi-module impact, not random guessing.

Reasoning Quality: Auditable by Design

Citations, completeness, consistency

CLEAR Layer	Completeness	Consistency	MUTCD Citation
Layer 1 Detection	94.0%	91.5%	82.4%
Layer 2 Classification	86.5%	84.0%	77.8%
Layer 3 Attribution	71.0%	73.2%	75.6%
Average	**83.8%**	**82.9%**	**78.6%**

Manual evaluation of CLEAR's reasoning quality across the three layers.
Completeness = coverage of relevant aspects; Consistency = internal logical coherence;
MUTCD Citation = accuracy of cited regulations.

78.6%

MUTCD citation accuracy

Completeness and consistency drop as reasoning complexity grows (L1 → L3)

MUTCD citation stays high → *outputs remain regulator-auditable across the entire pipeline.*

Layer 1 output · waymo_sf_0234

```
{
  "scene_id": "waymo_sf_0234",
  "anomaly_detected": true,
  "confidence": 0.92,
  "reasoning": "STOP sign placed at
   ground level — violates MUTCD
   2B.04 (7-ft mounting height).",
  "anomaly_indicators": [
    "improper sign placement",
    "height violation"
  ]
}
```

Structured · cited · auditable.

Key Takeaways

Three things to remember

1

Attribution is the unsolved gap.

Existing methods detect THAT something failed, not WHERE. CLEAR is the first to reach component-level resolution.

2

Regulatory grounding makes LLM reasoning auditable.

MUTCD-aligned justifications turn opaque LLM outputs into traceable, regulator-readable evidence.

3

Confidence behavior reveals genuine multi-module physics.

Apparent misattributions reflect the reality of fog, glare, and rain; **not reasoning failure.**

Limitations · Future Work · Questions

Limitations

- **Offline only** — ~3.1s/scene — forensic, not real-time
- **MUTCD only** — U.S.-grounded; pipeline is portable
- **Single-label default** — Multi-module attribution still open
- **Schema fragility** — ~15% need retry; <2% persist as malformed

Future work

- Knowledge distillation → smaller specialist models
- Caching + ANN retrieval for recurring scenes
- Batched inference → < 1s/scene latency
- Jurisdiction portability (EU, others)
- Multi-label attribution with calibrated probs

Questions?

Happy to discuss

Code & data

github.com/David-FR/MIRAGE_CLEAR

Acknowledgments

BMW · U.S.ARMY · DEVCOM GROUND VEHICLE SYSTEMS CENTER