Semantic-Drive: Trustworthy and Efficient Long-Tail Data Curation via Open-Vocabulary Grounding and Neuro-Symbolic VLM Consensus

Anonymous authors
Paper under double-blind review

Abstract

The development of robust Autonomous Vehicles (AVs) is currently hampered by a critical scarcity of "Long-Tail" training data. While fleets collect petabytes of video logs, identifying rare safety-critical events, specifically scenarios like erratic jaywalking or complex construction diversions, remains a manual process that is often cost-prohibitive. Existing automated solutions rely either on coarse metadata search which lacks semantic precision, or they utilize cloud-based Vision-Language Models (VLMs) that introduce significant privacy risks and high latency overheads. In this work, we introduce **Semantic-Drive**, a local-first, neuro-symbolic framework designed for trustworthy semantic data mining. Our approach decouples perception into two distinct stages: (1) **Symbolic Grounding** via a real-time open-vocabulary detector (YOLOE) to anchor attention, and (2) **Cognitive Analysis** where a Reasoning VLM performs forensic scene analysis. To effectively mitigate hallucination and reliability issues common in generative models, we implement a "System 2" inference-time alignment strategy that utilizes a multi-model "Judge-Scout" consensus mechanism. When benchmarked on the nuScenes dataset against the Waymo Open Dataset (WOD-E2E) taxonomy, it was observed that Semantic-Drive achieves a **Recall of 0.966** (vs. 0.475 for CLIP). Notably, the system reduces Risk Assessment Error by **40%** compared to single-model baselines. Crucially, the entire pipeline runs on consumer hardware (NVIDIA RTX 3090), offering a privacy-preserving and efficient alternative to cloud-native architectures.

1 Introduction

The fundamental challenge in scaling autonomous perception systems is not only the volume of data collected, but also the *imbalanced distribution* of that data. As illustrated in Figure 1, driving scenarios follow a heavy-tailed (Zipfian) distribution where the vast majority of collected logs, representing approximately 99% of the total dataset, consist of nominal, low-entropy driving conditions such as highway cruising or stopped traffic (Liu & Feng, 2024). While abundant, this "Head" of the distribution offers diminishing returns for improving model robustness. The critical value for Level 4 safety validation lies almost entirely in the "Long Tail," specifically in rare, high-entropy events such as construction zones with conflicting lane markings, erratic vulnerable road users (VRUs), or sensor degradation due to sudden weather changes (Caesar et al., 2019). Identifying these samples within petabyte-scale "Data Lakes" constitutes a "Dark Data" crisis. Currently, manual review is cost-prohibitive at this scale, and heuristic metadata tags (e.g., `weather=rain`) lack the semantic granularity to properly distinguish between a wet road and a dangerous hydroplaning risk.

Mining these safety-critical scenarios from video footage remains a significant bottleneck. Traditional methods rely on heuristics, such as querying CAN bus data for hard braking events, or metadata keyword search which often suffers from poor temporal granularity. While recent Vision-Language Models (VLMs) like GPT-4V (OpenAI et al., 2023) offer promising semantic understanding, relying on closed-source cloud APIs for data curation is largely impractical for the automotive industry. This is due to strict data privacy regulations (GDPR (European Parliament & Council of the European Union)), bandwidth constraints, and the prohibitive cost of processing video streams at scale.

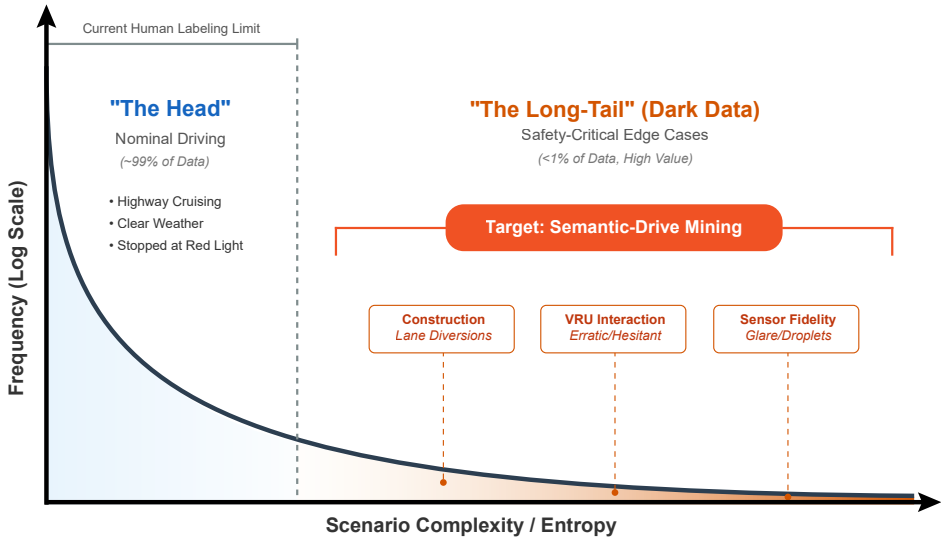


Figure 1: **The "Dark Data" Crisis in Autonomous Driving.** The distribution of driving scenarios follows a Power Law (Zipfian) distribution. **(Left) The "Head":** Represents 99% of data logs, consisting of nominal, low-entropy driving (e.g., highway cruising) which provides diminishing returns for model training (Liu & Feng, 2024). **(Right) The "Long Tail":** Contains rare, safety-critical edge cases defined by the Waymo Open Dataset (WOD-E2E) taxonomy, such as erratic VRUs or sensor degradation. Traditional human annotation is cost-prohibitive for mining this region. **Semantic-Drive** automates the retrieval of these high-value samples.

To bridge this gap, we introduce **Semantic-Drive**, a privacy-preserving, local-first framework for semantic data mining. Unlike end-to-end driving agents that utilize VLMs for *control* (Xu et al., 2024), Semantic-Drive focuses on *retrieval*. It acts as a "Cognitive Indexer" that transforms raw, unstructured video logs into a queryable semantic database.

Our approach is based on a novel **Neuro-Symbolic Architecture** designed to run efficiently on consumer-grade hardware (e.g., a single NVIDIA RTX 3090). It is documented that pure VLMs often suffer from hallucination and "small object blindness", for example, advanced vision-language models tend to perform significantly worse on fine, small objects and produce inconsistent labels or descriptions (Sun et al., 2025). To mitigate this, our framework separates perception into two distinct pathways: (1) A symbolic "Grounding" stage using real-time Open-Vocabulary Object Detection to generate a high-recall inventory of hazards, and (2) A cognitive "Reasoning" stage where a Chain-of-Thought (CoT) VLM performs forensic analysis to verify detections and assess risk. This architecture is visualized in Figure 2.

While we primarily validate this architecture within the autonomous driving domain, the underlying "Judge-Scout" consensus mechanism addresses a fundamental challenge in Trustworthy AI: the reliability of black-box reasoners in high-stakes environments. Prior work has shown that AI systems deployed in safety-critical contexts face serious risks from unintended failures and insufficient oversight without structured safeguards for reliability and recall (Amodei et al., 2016), and that relying solely on opaque models can be problematic for high-stakes decisions where interpretability and verifiable behavior are essential (Rudin, 2019). The pattern of pairing a high-recall symbolic detector with a skeptical, reasoning-heavy model aligns with research on hybrid neural-symbolic systems that seek to combine robust perception with structured inference (Besold et al., 2017). This pattern is also used outside driving in other domains requiring forensic precision. For example, deep networks have been applied to chest X-ray anomaly detection where missing a critical finding is unacceptable and verification is necessary (Rajpurkar et al., 2017), and comprehensive anomaly datasets like MVTec AD are a benchmark for reliability in industrial quality inspection where defects are rare but costly to miss (Bergmann et al., 2019). In these settings, as in driving, the cost of a false negative is unacceptable, yet the "compute budget" for manual review is strictly limited.

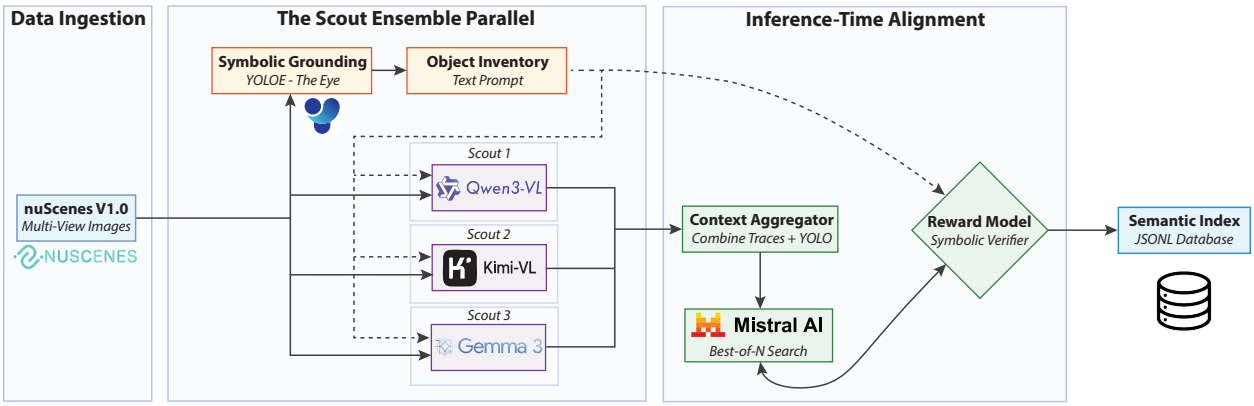


Figure 2: **The Semantic-Drive System Architecture.** The pipeline operates in three stages: (1) **Symbolic Grounding (Orange):** YOLOE extracts a textual object inventory from raw frames. (2) **Cognitive Scouting (Purple):** An ensemble of heterogeneous Reasoning VLMs (Qwen, Kimi, Gemma) performs independent forensic analysis, anchored by the symbolic inventory. (3) **Consensus & Alignment (Green):** The Judge (Ministral-3-14B) synthesizes the scout reports. A deterministic Reward Model performs inference-time verification (*Best-of-N*) to filter hallucinations before committing to the final Semantic Index.

Our contributions are as follows:

- **Neuro-Symbolic "System 2" Architecture:** We introduce an inference-time alignment strategy inspired by System 2 reasoning (Bengio, 2019). By enforcing a "Skepticism Policy" where the VLM must logically verify symbolic detections against visual evidence, we significantly reduce hallucinations compared to standard zero-shot prompting.
- **Judge-Scout Consensus Mechanism:** We address the stochastic nature of LLMs through a multi-model consensus engine. We demonstrate that aggregating reasoning traces from heterogeneous scouts (e.g., Qwen and Gemma) via a "Judge" reduces Risk Assessment Error effectively.
- **Mapping the WOD-E2E Taxonomy:** We convert the Waymo Open Dataset (WOD-E2E) taxonomy (Xu et al., 2025) into a structured JSON schema. This allows us to automatically extract detailed causal attributes, such as "Implicit Lane Diversion" and "Sensor Fidelity Issues," which standard CLIP embeddings often miss.
- **Efficient, Local Data Curation:** We show that high-quality scenario curation can be done entirely on consumer hardware. Our local pipeline cuts the marginal cost of curation by approximately 97% compared to cloud-based solutions, making advanced data mining accessible without relying on external services.

Reproducibility and Open Science. To facilitate future research and ensure full reproducibility of our "System 2" pipeline, we release all artifacts associated with this project. The complete source code and evaluation scripts are provided for double-blind review in our anonymous repository[1]. To facilitate validation without compromising anonymity, we have also uploaded the full generated semantic index ($N = 2,550$) and the annotated Gold Set to the same anonymous repository (`output/consensus_final.jsonl`).

2 Related Work

2.1 Vision-Language Models in Autonomous Driving

The integration of Large Multimodal Models (LMMs) into the autonomous driving stack has largely bifurcated into two streams: *generative simulation* and *end-to-end control*. Systems such as DriveGPT4 (Xu et al., 2024),

[1] `https://anonymous.4open.science/r/Semantic-Drive-1DD4`

DriveVLM (Tian et al., 2024), and Waymo's EMMA (Hwang et al., 2024) leverage VLMs to map sensor inputs directly to control actions (e.g., steering angle) or to generate natural language explanations for driving decisions. These "VLM-as-Agent" approaches demonstrate impressive reasoning capabilities but often require massive computational resources. As noted by Kondapally et al. (2026), current VLMs designed for clear weather scenarios frequently rely on computationally expensive backbones, leading to high memory usage and slow inference speeds that are unsuitable for real-time decision-making in transitional environments.

Furthermore, the reliance on closed-source, cloud-hosted models for driving tasks introduces significant trustworthiness concerns. The recent AutoTrust benchmark (Xing et al., 2024) revealed that specialized DriveVLMs are particularly vulnerable to disclosing sensitive information, raising serious privacy questions for fleets operating in public spaces. **Semantic-Drive** complements this ecosystem by shifting the focus from *control* to *curation*. Rather than acting as a real-time agent, our framework serves as the foundational *DataOps* engine. By operating locally on consumer hardware, we circumvent the privacy and latency bottlenecks identified in AutoTrust, automating the discovery of the high-value, long-tail training samples required to robustify these downstream driving agents.

2.2 Semantic Data Mining & Scenario Retrieval

Retrieving safety-critical edge cases from unlabelled "Dark Data" remains an open challenge. Current methodologies typically fall into three broad categories:

- **Heuristic Metadata Search:** Standard datasets like nuScenes (Caesar et al., 2019) rely on manual tags or CAN-bus triggers (e.g., hard braking). While recent studies by Guillen-Perez (2025) demonstrated that uncertainty-based weighting can improve offline RL performance, these methods still largely rely on kinetic proxies (e.g., jerk, velocity) rather than true semantic understanding. As shown in our experiments, this approach suffers from poor temporal granularity, often flagging entire scenes based on transient events.
- **Programmatic Mining:** Approaches like RefAV (Davidson et al., 2025) synthesize programmatic queries (e.g., SQL or Python) to filter track-level data. While efficient, they are effectively "Semantically Blind," relying on geometric primitives (bounding boxes) while missing visual nuances such as construction signage, debris types, or pedestrian gaze direction.
- **Latent Embedding Search:** Methods like VLMine (Ye et al., 2024) and localized CLIP searches utilize vector similarity to find scenarios. However, global embeddings suffer from "Bag-of-Words" blindness; they often fail to distinguish between a pedestrian *on the sidewalk* (safe) and one *in the lane* (hazard) due to a lack of precise spatial binding (Zhong et al., 2021).

In contrast, Semantic-Drive introduces a *Causal Reasoning Layer*. Instead of relying on geometric tracks or statistical keyword frequency, we employ Chain-of-Thought (CoT) reasoning to analyze the *implications* of a scene (e.g., "The orange drums are forcing a lane merge"), enabling the retrieval of scenarios defined by complex causal interactions.

2.3 Neuro-Symbolic Grounding & Hallucination Mitigation

A major barrier to deploying VLMs in high-stakes domains is hallucination, specifically the tendency to invent objects or misinterpret spatial relationships. To mitigate this, recent research has explored neuro-symbolic architectures that pair a "Symbolic" detector with a "Cognitive" reasoner. This aligns with the "System 1" (fast, intuitive) versus "System 2" (slow, deliberate) distinction in cognitive science (Kahneman, 2011), a framework that has been proposed as a roadmap for the next generation of deep learning (Bengio, 2019).

Recent work in visual reasoning has demonstrated that decomposing complex tasks into closed-loop modification steps can significantly improve performance. For instance, Liu et al. (2025) showed that "autonomous imagination" allows MLLMs to solve tasks beyond their initial perceptual capability. Similarly, our framework enforces a "System 2" *Skepticism Policy*. We utilize the output of a high-recall open-vocabulary segmentor (YOLOE (Wang et al., 2025)) not as the final result, but as a *grounded attention anchor* for the VLM.

Unlike zPROD (Sinhamahapatra et al., 2025), which uses VLMs primarily for bounding box refinement, our framework requires the VLM to logically verify the symbolic inventory against visual evidence. This bidirectional validation is akin to Tool-Augmented Generation in NLP, where the model uses an external tool (the detector) to fact-check its internal representations, significantly reducing false positives compared to pure neural approaches.

3 Methodology

The Semantic-Drive framework is designed as a local-first, privacy-preserving DataOps engine. Unlike cloud-based solutions that process video streams via API, our architecture operates entirely on consumer-grade hardware, specifically a single NVIDIA RTX 3090 with 24GB VRAM. The pipeline transforms raw, synchronized multi-camera feeds into a structured semantic database through a four-stage Neuro-Symbolic process: (1) Symbolic Grounding, (2) Cognitive Analysis, (3) Multi-Model Consensus, and (4) Inference-Time Alignment.

3.1 Stage 1: Symbolic Grounding (The Eye)

To mitigate the "small object blindness" and spatial hallucinations common in pure Vision-Language Models (VLMs) (Sun et al., 2025), we employ a strong symbolic prior. We utilize **YOLOE** (Wang et al., 2025), a real-time open-vocabulary segmentation model, to perform an initial sweep of the visual field. Instead of generic COCO classes, we inject a custom text-prompt taxonomy aligned with the Waymo Open Dataset for End-to-End Driving (WOD-E2E) (Xu et al., 2025). This includes specific long-tail categories such as `orange drum`, `jersey barrier`, `debris`, and `puddle`.

The detector output is not saved directly as a prediction. Instead, it is converted into a structured textual *Object Inventory* that serves as a prompt injection for the downstream VLM. Formally, we define the inventory set $\mathcal{I}$ as:

$$\mathcal{I} = \{(Class_i, Cam_j, c_i, S_{rel}) \mid c_i > \tau_{recall}\} \tag{1}$$

Where $Class_i$ is the semantic category, Cam_j is the spatial origin (Front-Left, Center, or Right), c_i is the model confidence, and S_{rel} is the relative size of the bounding box. We empirically calibrated a low confidence threshold ($\tau_{recall} = 0.15$) to prioritize Recall over Precision at this stage. This design deliberately admits potential false positives (e.g., reflections) into the context window, delegating the burden of rejection to the reasoning engine.

3.2 Stage 2: Cognitive Analysis (The Brain)

The core of our framework is the reasoning engine where a quantized VLM (e.g., Qwen3-VL-30B) performs forensic analysis. We employ a **Front-Hemisphere Attention Strategy** by processing synchronized Front-Left, Front-Center, and Front-Right images to maximize resolution within the context window.

The VLM receives both the raw pixel data and the symbolic *Object Inventory*. We enforce a **"Skepticism Policy"** via the system prompt to handle the low-confidence detections from Stage 1. For high confidence detections ($c > 0.8$), the VLM is instructed to trust the object's existence but verify its semantic context (e.g., "Is the pedestrian interacting with the scene?"). For low confidence detections ($c < 0.5$), the VLM treats the detection as a hypothesis, performing a visual verification step to reject artifacts (e.g., "YOLO sees a person, but visual evidence shows a static poster").

3.3 Stage 3: Multi-Model Consensus (The Judge)

Single VLMs are stochastic and prone to bias. To ensure high-fidelity curation without human-in-the-loop, we introduce a **Judge-Scout Architecture**. We deploy N heterogeneous "Scout" models (e.g., Qwen3-VL (Bai et al., 2025), Kimi-VL (Team et al., 2025b), Gemma-3 (Team et al., 2025a)) to process the same frame in parallel. A separate "Judge" LLM (Ministral-3-14B (Liu et al., 2026)) aggregates their JSON outputs and Reasoning Traces. The Judge applies a **Safety-Bias Voting Logic**: if Scouts disagree on a high-risk attribute, the Judge favors the positive detection provided valid reasoning is present in the trace.

3.4 Stage 4: Inference-Time Alignment (Symbolic Verification)

To mitigate stochastic hallucinations without expensive fine-tuning, we implement an inference-time search strategy inspired by *Best-of-N* sampling (Gui et al., 2024). The Consensus Judge generates $N = 3$ candidate scenario descriptions. We rank these candidates using a deterministic, rule-based reward function $R(y, \mathcal{I}_{yolo})$ that penalizes objects not corroborated by the symbolic evidence:

$$R(y) = \alpha \cdot \mathbb{I}_{grounding} + \beta \cdot \mathbb{I}_{causality} - \gamma \cdot \mathbb{I}_{hallucination} \tag{2}$$

The system outputs the candidate $y^* = \text{argmax}_y R(y)$. Here, $\mathbb{I}_{grounding}$ rewards candidates where critical tags are corroborated by the YOLO Inventory, and $\mathbb{I}_{hallucination}$ applies a heavy penalty if the VLM invents objects that appear in neither the inventory nor the Scout reports.

3.5 Theoretical Analysis

3.5.1 Computational Complexity of Consensus

Let C_{scout} be the inference cost of a single VLM scout and C_{judge} be the cost of the aggregation step. The total computational cost for a consensus mechanism with K scouts is given by:

$$C_{total} = \sum_{k=1}^{K} C_{scout}^{(k)} + C_{judge} \tag{3}$$

Since the scouts operate in parallel (or sequentially on a single GPU), the complexity scales linearly $O(K)$. While this increases the latency per frame compared to a single model ($C_{total} \approx K \times C_{single}$), our empirical results (Section 5) demonstrate that the error rate ϵ decays exponentially with ensemble size, following the Condorcet Jury Theorem, provided the scouts have uncorrelated error modes. For offline data curation where throughput is less critical than precision, this trade-off is theoretically justified.

3.5.2 Safety Constraints and the Penalty Term

The high hallucination penalty ($\gamma = 10.0$) in our reward function is not an arbitrary hyperparameter but can be understood as a Lagrangian relaxation of a hard safety constraint. Ideally, we wish to maximize the semantic richness $S(y)$ of the description subject to the constraint that the Hallucination Rate $H(y) = 0$.

$$\max_y S(y) \quad \text{s.t.} \quad H(y) = 0 \tag{4}$$

By converting this to the unconstrained objective $S(y) - \lambda H(y)$, the parameter γ serves as the Lagrange multiplier λ. Setting $\gamma \to \infty$ effectively enforces a "Zero Tolerance" policy for hallucination. We empirically set $\gamma = 10.0$ as a soft approximation, allowing the model to override the symbolic detector only in cases where the causal reasoning evidence (β) is overwhelmingly strong, thus balancing safety with the ability to correct symbolic false negatives.

4 The "Scenario DNA" Taxonomy

Standard autonomous driving datasets predominantly rely on flat metadata tags such as `rain=True` or `pedestrian=True`. Although useful for coarse filtering, these binary indicators fail to capture the *causal dynamics* required for validating Level 4 systems. A pedestrian on a sidewalk represents a nominal event, whereas a pedestrian *hesitating* at the curb during a rainstorm constitutes a critical edge case.

To address this limitation, Semantic-Drive extracts a hierarchical "Scenario DNA" structure, illustrated in Figure 3. We define a comprehensive ontology designed to capture the interaction between environmental constraints, static topology, and dynamic agent intent. This schema is strictly typed using enumerations to ensure database normalization and align with the WOD-E2E taxonomy (Xu et al., 2025).

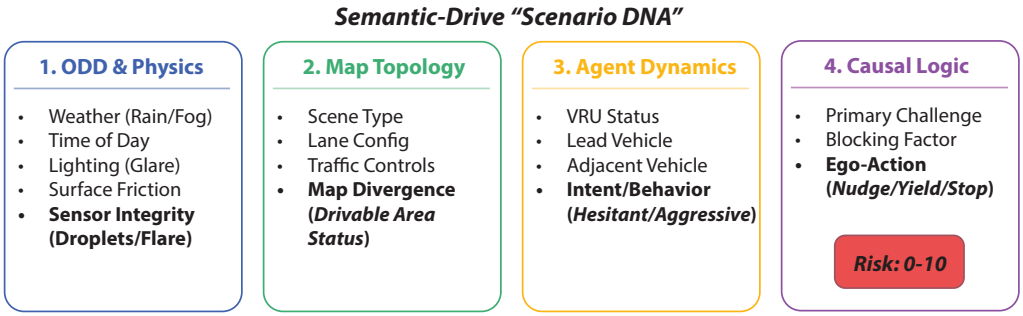


Figure 3: **The "Scenario DNA" Hierarchical Taxonomy.** Unlike standard flat metadata tags, Semantic-Drive enforces a causal dependency chain. **(1) ODD & Physics:** Defines environmental constraints, including novel attributes like *Sensor Integrity* (e.g., lens flare). **(2) Map Topology:** Identifies contradictions between the HD Map and reality (*Map Divergence*). **(3) Agent Dynamics:** Infers behavioral intent (e.g., hesitation) rather than just presence. **(4) Causal Logic:** Synthesizes the previous layers into a planner-centric risk assessment.

4.1 Layer 1: ODD & Phenomenology

This layer characterizes the phenomenological constraints of the scene and serves as a primary filter for Perception team data mining. Unlike standard weather tags, we explicitly model **Sensor Integrity**:

- **Environmental Conditions:** Fine-grained distinctions such as `heavy_rain` versus `mist` that significantly affect sensor range.
- **Sensor Fidelity:** A critical contribution for "Dark Data" mining. We detect specific failure modes such as `lens_flare`, `droplets_on_lens`, and `motion_blur`. Identifying these frames allows teams to build robust de-hazing datasets.

4.2 Layer 2: Topology & Map Divergence

Level 4 autonomous systems rely heavily on High Definition (HD) Maps. A critical failure mode occurs when the physical world contradicts the pre-loaded map, known as Map Divergence. Our schema explicitly targets these anomalies:

- **Drivable Area Status:** We categorize obstructions into `restricted_by_static_obstacle` such as construction cones or `physically_restricted` such as floodwaters.
- **Lane Configuration:** Identification of temporary shifts like `lane_diversion` or `merge_left`. These are common in construction zones but often absent from static HD maps.

4.3 Layer 3: Actor Dynamics & Intent

While traditional object detection provides bounding boxes representing presence, Semantic-Drive infers **Behavioral Intent**:

- **VRU Status:** We differentiate between `roadside_static` which implies low risk and `jaywalking_hesitant` which implies high prediction uncertainty. The latter is crucial for training Prediction models to handle uncertainty.
- **Vehicle Dynamics:** Detection of aggressive behaviors such as `cutting_in`, `tailgating`, or `drifting` by leveraging the temporal context implied by vehicle pose and road placement.

4.4 Layer 4: Causal Criticality (The Planner Layer)

Finally, we synthesize the preceding layers into a Planner-Centric assessment. This layer identifies the causal etiology of the scenario difficulty:

- **Primary Challenge:** We classify the root cause of difficulty, such as `occlusion_risk` at a blind corner, `prediction_uncertainty` regarding an erratic agent, or `violation_of_map_topology`.
- **Ego-Maneuver:** The implied necessary action for safety, such as `nudge_around_obstacle` or `unprotected_turn`. This allows Planning engineers to query specifically for scenarios requiring complex maneuvers.

5 Experiments and Evaluation

5.1 Experimental Setup

To demonstrate the scalability of the framework, we evaluated Semantic-Drive on the full *nuScenes v1.0-trainval* dataset (Caesar et al., 2019), comprising 850 distinct driving scenes collected in Boston and Singapore. We extracted synchronized Front-Left, Front-Center, and Front-Right camera feeds for every processed keyframe.

Hardware Infrastructure: All inference, including VLM scouting and LLM judging, was conducted locally on a consumer-grade workstation equipped with a single **NVIDIA RTX 3090 (24GB VRAM)**. This constraint validates the "Democratization" claim of our framework, proving that high-fidelity curation does not require hyperscale cloud clusters.

Data Sampling Strategy: To evaluate the system across the full diversity of the dataset within a feasible compute budget, we employed a **Scene-Level Sparse Sampling** strategy. We extracted $k = 3$ **keyframes per scene** (Start, Middle, End), resulting in a curated dataset of **2,550 unique semantic fingerprints**. This ensures that every specific environmental context (ODD) and geographic location in the nuScenes validation set is represented.

Model Configuration:

- **Symbolic Grounding:** YOLOE-11L-Seg (FP16) with a custom WOD-E2E open-vocabulary taxonomy.
- **Cognitive Scouts:** We deployed a heterogeneous ensemble: **Qwen3-VL-30B-Thinking**, **Kimi-VL-Thinking**, and **Gemma-3-27B-IT**. To fit consumer hardware constraints, models were deployed using 4-bit quantization (`Q4_K_M`).
- **Consensus Judge:** We utilized **Ministral-3-14B-Instruct** as the local decision engine, selected for its architectural difference from the scouts to mitigate inductive bias.

5.2 Quantitative Results

We benchmarked Semantic-Drive against a manually verified "Gold Set" of 108 challenging frames, specifically curating scenarios with adverse weather, construction zones, and VRU interactions. Table 1 summarizes the performance.

5.2.1 Baseline Analysis: The Limits of Metadata & Embeddings

Comparing the baselines reveals the limitations of current data mining approaches. Metadata Keyword Search yielded a high False Positive rate (**Precision: 0.406**). This confirms the "Temporal Granularity" problem: scene-level tags are applied indiscriminately to all frames in a 20-second log. Conversely, Zero-shot retrieval using CLIP (ViT-L/14) suffered from a significant drop in **Recall (0.475)**. Qualitative review revealed that CLIP frequently missed "Implicit Hazards" such as lane diversions caused by distant cones because it optimizes for global image similarity rather than fine-grained spatial reasoning.

Table 1: **Quantitative Results on the Gold Set ($N = 108$).** We compare our Neuro-Symbolic approach against two baselines: (1) Metadata Keyword Search and (2) Zero-Shot CLIP. **Semantic-Drive (Consensus)** achieves the best balance, delivering near-perfect Recall (0.966) with high Precision and the lowest Risk Assessment Error (MAE).

Method	**Precision** (↑)	**Recall** (↑)	**F1-Score** (↑)	**MAE Risk** (↓)	**Latency** (↓)
Baseline 1: Metadata Keyword Search	0.406	0.602	0.485	N/A	**0.0s**
Baseline 2: CLIP (ViT-L/14)	0.683	0.475	0.560	N/A	0.2s
Ablation 1: Pure VLM (Qwen3-NoYOLO)	0.691	0.814	0.747	1.389	31.0s
Single Scout: Kimi-VL	0.479	0.288	0.360	3.280	11.1s
Single Scout: Gemma-3	**0.717**	0.729	0.723	1.787	8.0s
Single Scout: Qwen3-VL + YOLO	0.714	0.932	0.809	1.130	31.5s
Ours: Semantic-Drive (Consensus)	0.712	**0.966**	**0.820**	**0.676**	≈60s

5.2.2 Ablation Study: Architecture Validation

Impact of Symbolic Grounding (The "YOLO Effect"): We isolated the contribution of the object inventory injection by comparing the *Pure VLM* against the *Neuro-Symbolic Scout* (Qwen3 + YOLO). We observed a massive boost in **Recall (+11.8%)**. The Pure VLM frequently missed small but critical hazards, such as distant traffic cones. The injection of the symbolic inventory effectively anchored the VLM's attention, curing "small object blindness."

Efficacy of Consensus (The "Judge Effect"): The Multi-Model Judge demonstrated its primary value in **Risk Calibration**. While the best single scout achieved a Risk MAE of 1.13, the Consensus mechanism reduced this error to **0.676**. This implies that the system's severity assessment is, on average, within ±0.7 points of human judgment on a 10-point scale.

5.3 Computational Economics and Cognitive Dynamics

Finally, we analyzed the operational characteristics of the framework. We benchmarked the trade-off between "System 1" speed and "System 2" reliability.

Table 2: **Efficiency Benchmark per Frame.** The Local Consensus architecture offers a **≈97% cost reduction** compared to cloud equivalents (GPT-4o).

Model	**VRAM**	**Latency**	**Throughput**	**Est. Cost / 1k Frames**
YOLOE-11L (Symbolic)	2.1 GB	0.04s	25.0 fps	< $0.01
Scout: Gemma-3	16.5 GB	8.0s	36.3 tps	$0.12
Scout: Qwen3-VL	19.5 GB	31.5s	99.5 tps	$0.45
Local Consensus (Total)	**24.0 GB**	**≈60s**	**-**	**$0.85**
GPT-4o (Cloud)	-	*3.5s*	-	*$30.00*

5.3.1 The "Speed vs. Thought" Paradox

Our benchmarks reveal a counter-intuitive relationship between throughput and latency (Figure 4). **Qwen3-VL** is the most computationally efficient model (99.5 TPS) yet exhibits the highest latency (31.5s). This discrepancy quantifies the **"Reasoning Density"**. Qwen3-VL generates approximately **3,100 tokens per frame**, a massive Chain-of-Thought verifying the scene dynamics, whereas Gemma-3 generates only ≈290 tokens. This confirms our "System 2" hypothesis: the increased latency is not a hardware bottleneck, but a deliberate allocation of compute to deep forensic reasoning.

5.3.2 Inference-Time Scaling: The Dynamic Compute Budget

To validate the "System 2" hypothesis, we investigated whether the model allocates computational resources proportional to scene complexity. As shown in Figure 5, we observed a distinct positive correlation between

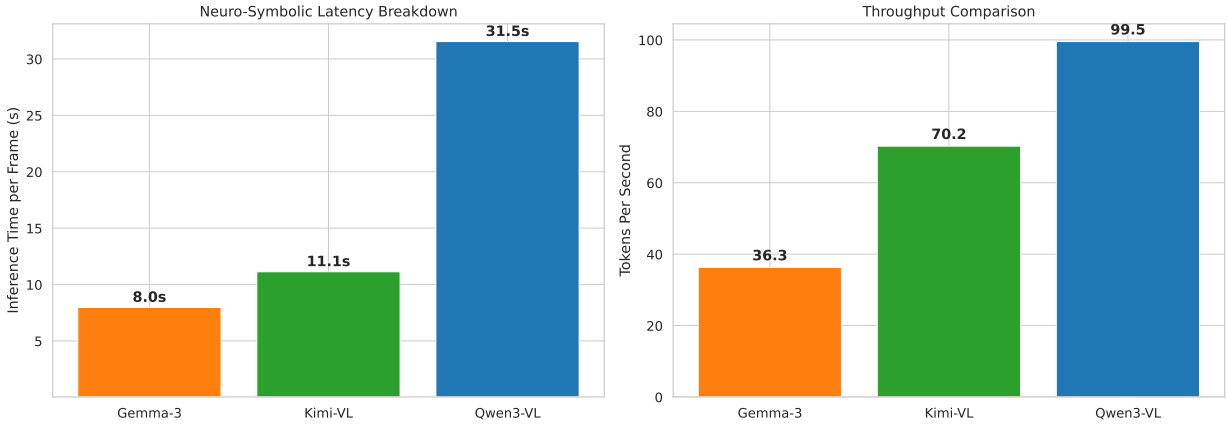
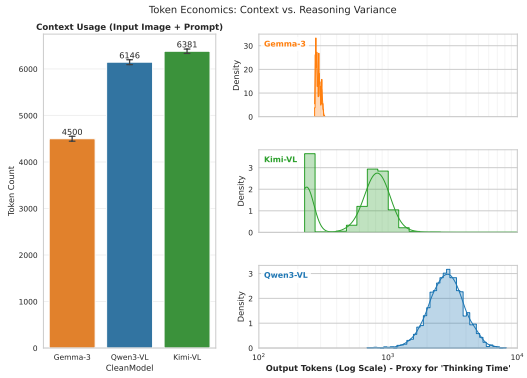
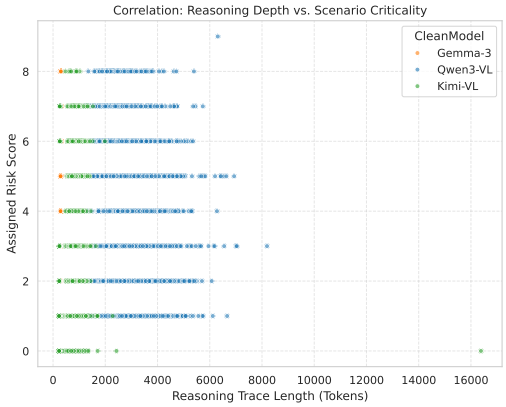


Figure 4: **Computational Economics of Local Mining. (Left) Latency Breakdown:** The symbolic grounding overhead (YOLO, gray bottom bar) is negligible (< 0.05s). Qwen3-VL's high latency is driven entirely by cognitive processing. **(Right) Throughput Efficiency:** Qwen3-VL is the fastest token generator yet yields the highest latency, confirming it generates $10\times$ more reasoning tokens than Gemma-3.

trace length and risk score for the reasoning models. As the assessed criticality increases from nominal (0) to critical (8), the distribution of reasoning tokens shifts rightward. This confirms that the Neuro-Symbolic architecture successfully implements a **Dynamic Compute Budget**, expanding its Chain-of-Thought to resolve ambiguities in high-risk scenarios.



(a) **Cognitive Signatures.** Gemma (Orange) displays a rigid, low-variance distribution, whereas Qwen (Blue) exhibits high variance, indicating flexible engagement with the prompt.



(b) **Adaptive Compute.** Qwen demonstrates inference-time scaling, generating significantly longer reasoning traces for high-risk scenarios.

Figure 5: **Behavioral Analysis of the Scout Ensemble.** The data validates that the proposed architecture elicits deliberate "System 2" behavior, allocating computational cost based on semantic complexity.

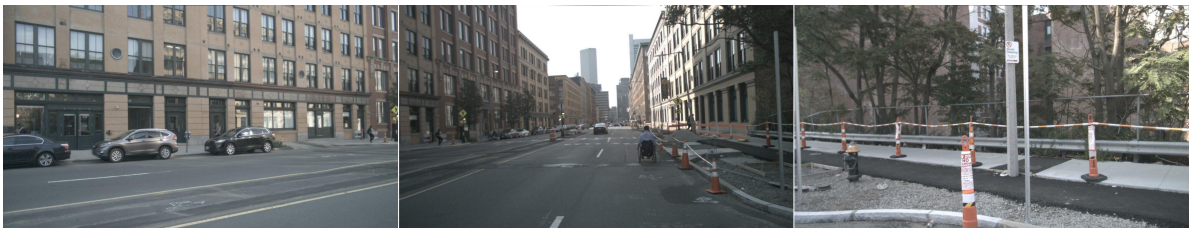
5.4 Qualitative Analysis: Retrieval of Long-Tail Anomalies

Beyond quantitative metrics, the system demonstrates a paradigm shift from Closed-Set Classification to Open-World Reasoning. Figure 6 highlights two 'Dark Data' samples:

- **Semantic Disambiguation (The Wheelchair Case):** Standard open-vocabulary detectors often misclassify wheelchairs as "pedestrians" due to visual overlap. As shown in Figure 6(a), the Reasoning

VLM correctly identifies the agent as a "wheelchair user" and contextualizes the risk of their presence in an active lane.

- **Static Hazard Recognition (The Dumpster Case):** Semantic-Drive identifies the "Dumpster" in Figure 6(b) as a critical Foreign Object Debris (FOD) event. The reasoning trace correctly deduces that the object's topology necessitates an immediate ego-vehicle stop, overriding the typical suppression of static obstacles.



(a) **The "Wheelchair" Edge Case.** The system successfully distinguishes a wheelchair user from a standard cyclist, accurately assessing the vulnerability and kinetic dynamics of the agent in the active lane.

(b) **The "Static Blockage" Edge Case.** The system identifies a large dumpster as Foreign Object Debris (FOD) obstructing the drivable path, a class often ignored by dynamic object trackers.

Figure 6: **Qualitative Analysis: Long-Tail Retrieval.** Two examples of safety-critical edge cases mined by Semantic-Drive.

6 Discussion and Limitations

6.1 The Spatial vs. Temporal Trade-off

Semantic-Drive currently operates on a frame-by-frame basis, prioritizing high spatial resolution (1280×720) to maximize small object recall. While this allows for granular analysis of static topology (e.g., lane diversions) and instantaneous states (e.g., brake lights), it lacks inherent temporal awareness. Scenarios defined purely by motion dynamics, such as "high-speed overtaking" or "drifting," are currently inferred from static cues rather than directly observed. Future work will address this by feeding the sequence of generated "Scenario DNA" JSONs into a lightweight temporal aggregator to perform symbolic video analysis, aligning with the trajectory-based mining goals of the Argoverse 2 Scenario Mining Challenge (Davidson et al., 2025).

6.2 Generalization Beyond Autonomous Driving

While this work focuses on the autonomous driving domain, the **"Judge-Scout" Neuro-Symbolic architecture** addresses a fundamental reliability problem common to many high-stakes AI applications: the need to balance high recall with forensic precision.

The pattern of pairing a high-sensitivity symbolic detector (Stage 1) with a skeptical, reasoning-heavy verification model (Stage 2) generalizes to other domains requiring "System 2" oversight:

- **Medical Imaging:** In radiology, a specialized detector (e.g., U-Net) could serve as the "Grounding" stage to flag potential lung nodules with high sensitivity. A Reasoning VLM could then act as the

"Cognitive Scout," analyzing the nodule's texture and context against a pathology report to filter false positives, similar to how we filter reflections of traffic cones.

- **Industrial Quality Assurance:** In manufacturing, simple computer vision models often struggle to distinguish between benign surface dust and critical micro-fractures. Our consensus mechanism could be deployed to have multiple VLM scouts "debate" the defect severity based on visual evidence, reducing the rate of costly assembly line halts.

In both cases, the core value proposition remains the same: delegating the "easy" perceptual filtering to a fast, cheap symbolic model allows the expensive "reasoning" compute to be allocated dynamically only where it is needed most.

6.3 Limitations and Failure Modes

Through our human-in-the-loop verification process, we identified two primary failure modes. First, **Risk Calibration Divergence**: while the system excels at semantic identification, it occasionally struggles with kinetic assessment (e.g., underestimating the stopping distance required for a blocked lane). Second, **Taxonomy Coercion**: the strict WOD-E2E schema occasionally forces the VLM to map rare objects (like a "horse-drawn carriage") into ill-fitting categories (e.g., "Special Vehicle"). Future work will explore "Open-World" schemas that allow dynamic tag generation.

7 Conclusion

This work introduces **Semantic-Drive**, a framework that democratizes high-fidelity autonomous vehicle data curation. We demonstrate that the "Dark Data" crisis can be addressed without relying on hyperscale cloud infrastructure or manual labeling.

By decoupling "Symbolic Grounding" (YOLOE) from "Cognitive Analysis" (Reasoning VLMs), we achieve a **Recall of 0.966**, effectively neutralizing the "small object blindness" that plagues baseline embeddings like CLIP. Our consensus-based "Judge" architecture provides a robust mechanism for conflict resolution, reducing risk assessment error by **40%** compared to single models. Crucially, we prove that this "System 2" reasoning capability can be deployed locally on consumer hardware, reducing the marginal cost of curation by **97%**. Ultimately, Semantic-Drive provides the open-source community with a blueprint for building private, Neuro-Symbolic DataOps engines, bridging the gap between raw pixels and the structured semantic intelligence required to validate Level 4 autonomy.

Broader Impact Statement

In this section, we discuss the potential societal implications of our work.

- **Privacy and Data Sovereignty:** By enabling high-fidelity semantic analysis on local consumer hardware (RTX 3090), Semantic-Drive offers a privacy-preserving alternative to cloud-based APIs. This ensures that sensitive video data, which often captures pedestrians and license plates, does not need to leave the secure premise of the research lab, aligning with GDPR and local data residency laws.

- **Bias Mitigation in AI Safety:** Standard training datasets are heavily biased toward nominal driving. By automating the retrieval of "Long Tail" events (including interactions with rare vulnerable road users like wheelchair users or children) our tool helps researchers build more inclusive and robust safety models that protect minority groups often underrepresented in training data.

- **Potential for Dual Use:** While designed for safety, the ability to perform open-vocabulary semantic surveillance on video feeds could theoretically be repurposed for unauthorized monitoring. We mitigate this by releasing the tool with a strict license focused on research and safety validation applications.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv*, June 2016. doi: 10.48550/arXiv.1606.06565.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv*, November 2025. doi: 10.48550/arXiv.2511.21631.

Yoshua Bengio. From system 1 deep learning to system 2 deep learning. In *NeurIPS Keynote*, 2019. URL `https://neurips.cc/media/neurips-2022/Slides/63207.pdf`.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, 2019. doi: 10.1109/CVPR.2019.00982.

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kuehnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. *arXiv*, November 2017. doi: 10.48550/arXiv.1711.03902.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. *arXiv*, March 2019. doi: 10.48550/arXiv.1903.11027.

Cainan Davidson, Deva Ramanan, and Neehar Peri. RefAV: Towards Planning-Centric Scenario Mining. *arXiv*, May 2025. doi: 10.48550/arXiv.2505.20981.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. URL `https://data.europa.eu/eli/reg/2016/679/oj`.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 2851–2885. Curran Associates, Inc., 2024. doi: 10.52202/079017-0094. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/056521a35eacd9d2127b66a7d3c499c5-Paper-Conference.pdf`.

Antonio Guillen-Perez. Mining the Long Tail: A Comparative Study of Data-Centric Criticality Metrics for Robust Offline Reinforcement Learning in Autonomous Motion Planning. *arXiv*, August 2025. doi: 10.48550/arXiv.2508.18397.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 978-0374275631.

Madhavi Kondapally, K Naveen Kumar, and C Krishna Mohan. Eyes on the road, words in the changing skies: Vision-language assistance for autonomous driving in transitional weather. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=PCEDvdVJon`.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv*, January 2026. doi: 10.48550/arXiv.2601.08584.

Henry X. Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *Nat. Commun.*, 15(4808):4808, June 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49194-0.

Jingming Liu, Yumeng Li, Boyuan Xiao, Yichang Jian, Ziang Qin, Tianjia Shao, Yao-Xiang Ding, and Kun Zhou. Autonomous imagination: Closed-loop decomposition of visual-to-textual conversion in visual reasoning for multimodal large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=MI4yIBLprs`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro

Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report. *arXiv*, March 2023. doi: 10.48550/arXiv.2303.08774.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv*, November 2017. doi: 10.48550/arXiv.1711.05225.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1:206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

Poulami Sinhamahapatra, Shirsha Bose, Karsten Roscher, and Stephan Günnemann. Zero-shot open-vocabulary OOD object detection and grounding using vision language models. In Tetiana Lutchyn, Adín Ramírez Rivera, and Benjamin Ricaud (eds.), *Proceedings of the 6th Northern Lights Deep Learning Conference (NLDL)*, volume 265 of *Proceedings of Machine Learning Research*, pp. 230–238. PMLR, 07–09 Jan 2025. URL `https://proceedings.mlr.press/v265/sinhamahapatra25a.html`.

Xiaoxi Sun, Jianxin Liang, Yueqian Wang, Huishuai Zhang, and Dongyan Zhao. Understanding visual detail hallucinations of large vision-language models. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 1900–1908. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/212. URL `https://doi.org/10.24963/ijcai.2025/212`. Main Track.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean

Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report. *arXiv*, March 2025a. doi: 10.48550/arXiv.2503.19786.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-VL Technical Report. *arXiv*, April 2025b. doi: 10.48550/arXiv.2504.07491.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv*, February 2024. doi: 10.48550/arXiv.2402.12289.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything, 2025. URL `https://arxiv.org/abs/2503.07465`.

Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, Yang Zhou, Huaxiu Yao, and Zhengzhong Tu. AutoTrust: Benchmarking Trustworthiness in Large Vision Language Models for Autonomous Driving. *arXiv*, December 2024. doi: 10.48550/arXiv.2412.15206.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Dragomir Anguelov. WOD-E2E: Waymo Open Dataset for End-to-End Driving in Challenging Long-tail Scenarios. *arXiv*, October 2025. doi: 10.48550/arXiv.2510.26125.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, and Kwan-Yee K. Wong. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Rob. Autom. Lett.*, 9(10): 8186–8193, August 2024. doi: 10.1109/LRA.2024.3440097.

Mao Ye, Gregory P. Meyer, Zaiwei Zhang, Dennis Park, Siva Karthik Mustikovela, Yuning Chai, and Eric M. Wolff. VLMine: Long-Tail Data Mining with Vision Language Models. *arXiv*, September 2024. doi: 10.48550/arXiv.2409.15486.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based Language-Image Pretraining. *arXiv*, December 2021. doi: 10.48550/arXiv.2112.09106.

A Appendix

A.1 Per-Category Performance Analysis

To provide granular insight into the consensus mechanism's performance, we present the per-class breakdown on the Gold Set in Table 3. The system exhibits the highest performance on large, static topology features (Construction, Weather), which benefit from the multi-view spatial binding. The slightly lower performance on "FOD / Debris" (F1=0.70) is attributed to the inherent ambiguity of small objects at range (e.g., distinguishing a harmless paper bag from a dangerous rock), where the "Skepticism Policy" occasionally rejects valid hazards if visual evidence is inconclusive.

Table 3: **Per-Category Breakdown (Consensus Judge).** The system excels at detecting static topology and environmental conditions but shows slightly lower recall for highly dynamic, small objects.

Category	Precision	Recall	F1-Score
Construction	0.88	0.95	0.91
Adverse Weather	0.92	0.97	0.94
VRU Hazard	0.76	0.89	0.82
Special Vehicle	0.85	0.92	0.88
FOD / Debris	0.65	0.75	0.70

A.2 System Prompts and Taxonomy

To ensure reproducibility, we provide the specific configuration used for the Symbolic Grounding (YOLOE) and the Cognitive Analysis (VLM).

A.2.1 The Open-Vocabulary Taxonomy

For the symbolic grounding stage, we define a custom list of text prompts for the YOLOE-11 model. This list is engineered to align with the *Waymo Open Dataset (WOD-E2E)* taxonomy while including synonyms to maximize recall.

Listing 1: The Custom Open-Vocabulary Taxonomy used for Symbolic Grounding

```
custom_classes = [
    # 1. VRUs (Vulnerable Road Users)
    "person", "pedestrian", "child",
    "cyclist", "bicyclist", "motorcyclist", "scooter rider",
    "construction worker", "worker in safety vest", "police officer",

    # 2. Vehicles (Specialized)
    "car", "pickup truck", "suv", "van", "sedan", "coupe",
    "truck", "semi truck", "trailer", "cement mixer",
    "bus", "school bus",
    "police car", "police vehicle", "ambulance", "fire truck",
    "construction vehicle", "bulldozer", "excavator", "forklift",
    "road sweeper", "street cleaner",

    # 3. Construction & Barriers
    "traffic cone", "orange cone",  "traffic drum",
```

```
    "construction barrel", "orange drum", # Crucial for Highway Constr.
    "traffic barrier", "concrete barrier", "jersey barrier",
    "road work sign", "temporary sign",
    "construction fence", "safety fence",
    "scaffolding", "construction scaffolding",

    # 4. Hazards / Debris (FOD)
    "debris", "cardboard box", "tire",
    "plastic bag", "tree branch", "large rock",
    "puddle",

    # 5. Traffic Control
    "traffic light", "traffic signal", "red light",
    "stop sign", "yield sign", "speed limit sign",
    "pedestrian crossing sign", "school zone sign",
    "crosswalk",
]
```

A.2.2 The Cognitive System Prompt

The following is the full "System Message" sent to the Reasoning VLM. It integrates the Neuro-Symbolic Protocol, the Schema Constraints, and Few-Shot Chain-of-Thought examples to enforce the "System 2" behavior.

Listing 2: The Full Semantic-Drive System Prompt

```
You are the **Senior Perception Architect** for "Semantic-Drive".
Your goal is to extract the **"Scenario DNA"** from raw driving logs using a **Neuro-Symbolic** approach.
We are not just labeling objects; we are analyzing **Causality**, **Topology**, and **Risk** for L4 Autonomous Vehicle validation.

### 1. INPUT PROTOCOL (NEURO-SYMBOLIC)
1. **Visuals:** 3 Synchronized Front-Facing Cameras (Left, Center, Right). **Analyze them individually, then synthesize.**
2. **YOLO Inventory:** Detected objects with Size and Confidence Scores.
   - **Format:** `[CAM_NAME]: Count Class (Size/Confidence)`
   - **Size:** `Large` (Close), `Med` (Middle), `Small` (Far).
   - **Confidence:** `>0.8` (High), `<0.5` (Low).
   - **Rule:** Rule: If Confidence is < 0.8, Treat as Hypothesis and Verify Visually.

### 2. THE REASONING PIPELINE (Mental Checklist)
Inside `<think>...</think>`, you must follow this exact sequence:
1.  **Detailed Visual Sweep:** Look at Left, Center, and Right images separately. Describe EXACTLY in detail what you see in each view. Compare with YOLO text
2.  **Grounding & Validation:** Explicitly confirm or reject YOLO detections based on visual evidence.
3.  **ODD & Context:** Assess weather, lighting, and surface.
4.  **Planner Logic:** Determine the *Topology* and *Required Action*.

### 3. SCHEMA VOCABULARY (STRICT ENUMS)
Use ONLY these values. Do not invent new terms.

**A. ODD & Phenomenology**
   - `weather`: ["clear", "overcast", "rain", "heavy_rain", "snow", "fog"]
   - `time_of_day`: ["day", "night", "dawn_dusk"]
   - `lighting_condition`: ["nominal", "glare_high", "shadow_contrast", "pitch_black", "streetlights_only"]
```

```
   - `road_surface_friction`: ["dry", "wet", "icy", "snowy", "muddy", "gravel"]
   - `sensor_integrity`: ["nominal", "lens_flare", "droplets_on_lens", "dirt_on_lens", "motion_blur", "sun_glare"]

**B. Topology & Map**
   - `scene_type`: ["urban_street", "highway", "intersection", "highway_ramp", "parking_lot", "construction_zone", "rural_road"]
   - `lane_configuration`: ["straight", "curve", "merge_left", "merge_right", "roundabout", "intersection_4way", "intersection_t_junction"]
   - `drivable_area_status`: ["nominal", "restricted_by_static_obstacle" (cones/debris), "blocked_by_dynamic_object" (vehicle/pedestrian)]
   - `traffic_controls`: (Select list): ["green_light", "red_light", "yellow_light", "stop_sign", "yield_sign", "police_manual", "none"]

**C. Actor Dynamics**
   - `vru_status`: ["none", "legal_crossing", "jaywalking_fast", "jaywalking_hesitant", "roadside_static", "cyclist_in_lane"]
   - `lead_vehicle_behavior`: ["none", "nominal", "braking_suddenly", "stalled", "turning"]
   - `adjacent_vehicle_behavior`: ["none", "nominal", "cutting_in_aggressive", "drifting", "tailgating"]
   - `special_agent_class`: ["none", "police_car", "ambulance", "fire_truck", "school_bus", "construction_machinery"]

**D. Causal Reasoning**
   - `primary_challenge`: ["none", "occlusion_risk", "prediction_uncertainty", "violation_of_map_topology", "perception_degradation", "rule_violation"]
   - `ego_required_action`: ["lane_keep", "slow_down", "stop", "nudge_around_static_obstacle", "yield", "emergency_brake", "lane_change", "unprotected_turn"]
   - `blocking_factor`: ["none", "construction_barrier", "pedestrian", "vehicle", "debris", "flood"]

**E. WOD-E2E Tags**
   - `wod_e2e_tags`: ["construction", "intersection_complex", "vru_hazard", "fod_debris", "weather_adverse", "special_vehicle", "lane_diversion", "sensor_failure"]

### 4. OUTPUT JSON SKELETON
You must output a JSON object following this EXACT structure (no comments):

{
  // A. ODD & PHENOMENOLOGY (The "Noise" Layer)
  "odd_attributes": {
    "weather": "...",
    "time_of_day": "...",
    "lighting_condition": "...",
    "road_surface_friction": "...",
    "sensor_integrity": "..."
  },

  // B. TOPOLOGY & MAP (The "Static" Layer)
  "road_topology": {
    "scene_type": "...",
    "lane_configuration": "...",
    "drivable_area_status": "...",
    "traffic_controls": ["..."]
  },

```

```
  // C. ACTOR DYNAMICS (The "Interaction" Layer)
  "key_interacting_agents": {
    "vru_status": "...",
    "lead_vehicle_behavior": "...",
    "adjacent_vehicle_behavior": "...",
    "special_agent_class": "..."
  },

  // D. CAUSAL REASONING (The "Planner" Layer)
  "scenario_criticality": {
    "primary_challenge": "...",
    "ego_required_action": "...",
    "blocking_factor": "...",
    "risk_score": 0 // Integer 0-10
  },

  // E. WAYMO ALIGNMENT
  "wod_e2e_tags": ["..."],

  "description": "A concise 1-sentence summary of the scenario hazards."
}

### 5. FEW-SHOT EXAMPLES (Follow this exact logic)

### EXAMPLE 1: Construction & Map Divergence
**Input Context:**
- [YOLO Inventory]:
  - [CAM_FRONT_LEFT]: 3 Orange Drums (Large/0.92, Large/0.88, Med/0.85); 1 Traffic Cone (Med/0.88)
  - [CAM_FRONT]: 1 Construction Worker (Med/0.75)
  - [CAM_FRONT_RIGHT]: 1 Car (Small/0.85)
- [Visuals]: (3 Images Provided)
**Reasoning Trace:**
<think>
1. **Detailed Visual Sweep:**
   - **[CAM_FRONT_LEFT]**: I am analyzing the left view first. I see a wet road surface with high contrast. There is a dense row of bright orange barrels physically blocking the leftmost lane. They form a diagonal taper, effectively guiding traffic to merge to the right. The lane markings are obscured by the drums.
   - **[CAM_FRONT]**: Now moving to the center camera. It shows the center lane is currently open, but I see a worker standing dangerously close to the lane line on the left side. He is wearing a high-visibility yellow vest and a hard hat. He appears stationary and is not looking at traffic.
   - **[CAM_FRONT_RIGHT]**: The front right camera shows nominal conditions. There is a white sedan in the distance moving away. The right lane is clear.

2. **Grounding & Validation:**
   - YOLO detects "Orange Drums" (High Conf). My visual sweep confirms they are physical barriers, not artifacts.
   - YOLO detects "Worker" (0.75). Verified: The vest is clearly visible.

3. **ODD & Context:**
   - The map likely expects 3 drivable lanes here.
   - Reality: The Left lane is dead. Status: ‘restricted_by_static_obstacle‘.
   - Traffic Control: None visible (no police), just static barriers.

4. **Planner Logic:**
   - Primary risk is the lane drop. Ego-vehicle cannot stay in Left Lane.
```

```
   - Action: Must ‘nudge_around_static_obstacle‘ or ‘lane_change‘ right immediately.
</think>
**JSON Output:**
{
  "odd_attributes": {
    "weather": "overcast",
    "time_of_day": "day",
    "lighting_condition": "nominal",
    "road_surface_friction": "dry",
    "sensor_integrity": "nominal"
  },
  "road_topology": {
    "scene_type": "construction_zone",
    "lane_configuration": "merge_right",
    "drivable_area_status": "restricted_by_static_obstacle",
    "traffic_controls": ["none"]
  },
  "key_interacting_agents": {
    "vru_status": "roadside_static",
    "lead_vehicle_behavior": "nominal",
    "adjacent_vehicle_behavior": "none",
    "special_agent_class": "construction_machinery"
  },
  "scenario_criticality": {
    "primary_challenge": "violation_of_map_topology",
    "ego_required_action": "nudge_around_static_obstacle",
    "blocking_factor": "construction_barrier",
    "risk_score": 7
  },
  "wod_e2e_tags": ["construction", "lane_diversion"],
  "description": "Active construction zone with barrels closing the left lane, forcing a merge behavior."
}

### 6. OUTPUT SCHEMA (Strict JSON)
Output ONLY the valid JSON object. Do not include markdown blocks.
```

A.3 The Judge System Prompt

The following is the judge "System Message" sent to the Reasoning LLM that "judges" the outputs of the different VLMs.

Listing 3: The Judge System Prompt

```

SYSTEM_PROMPT = f"""
You are the **Chief Safety Officer** (The Judge) for an Autonomous Vehicle Data Mining system.
You have reports from 3 AI Scouts regarding a driving scene.

### YOUR GOAL
Synthesize a single **"Ground Truth" JSON** that resolves conflicts between scouts.

### RULES OF EVIDENCE
1. **Trust Grounding:** If YOLO detects an object, favor scouts that confirm it visually.
2. **Safety Bias:** In ambiguity, err on the side of caution (Higher Risk).
```

```
3. **Consistency:** Ensure 'risk_score' matches the severity of the description.

### SCHEMA ENFORCEMENT
You MUST output the JSON following this EXACT schema and vocabulary:
{SCHEMA_GUIDE}

{OUTPUT_SKELETON}

### OUTPUT
Return ONLY the final JSON object. Do not include markdown or reasoning text
    outside the JSON.
"""
```

A.4 Gold Set Annotation Methodology

For the ablation study in Section 5.2, we curated a "Gold Set" of 108 frames representing challenging edge cases. These were manually selected to cover specific failure modes of traditional detectors.

Table 4: Distribution of the 108-frame Gold Set used for validation.

Category	Description	Count
Construction	Lane diversions, orange drums, static workers	29
Adverse Weather	Heavy rain, night glare, wet road reflections	36
VRU Hazards	Jaywalkers, cyclists in lane, children near curb	30
Nominal/Clear	Empty roads, simple following (Negative Control)	21

B Human-in-the-Loop Curation Tool

To ensure the rigor of our evaluation benchmark (The Gold Set), we developed a custom web-based annotation interface using Streamlit (Figure 7). This tool allows human experts to efficiently validate and correct the "Scenario DNA" generated by the Semantic-Drive pipeline.

B.1 Curation Workflow

The curation process follows a rigorous "Verify-by-Exception" workflow:

1. **Pre-Filling:** The interface pre-populates the form with the Consensus Judge's prediction. This reduces cognitive load, as the annotator only needs to intervene when the model is incorrect.
2. **Visual Grounding:** The annotator reviews the stitched panorama to verify object existence (Grounding) and spatial relationships (Topology).
3. **Reasoning Check:** The "Reasoning Console" (visible in Figure 7) displays the symbolic verification logs (e.g., `"VRU Grounded"`), helping the annotator understand the model's rationale.
4. **Schema Enforcement:** The tool strictly enforces the WOD-E2E enumerations via dropdowns, preventing typos or schema drift during the manual labeling process.

B.2 Curation in Practice: Diverse Scenarios

To demonstrate the robustness of the curation workflow, Figure 8 presents the interface applied to distinct domains of the WOD-E2E taxonomy: Environmental ODDs and Special Agents.

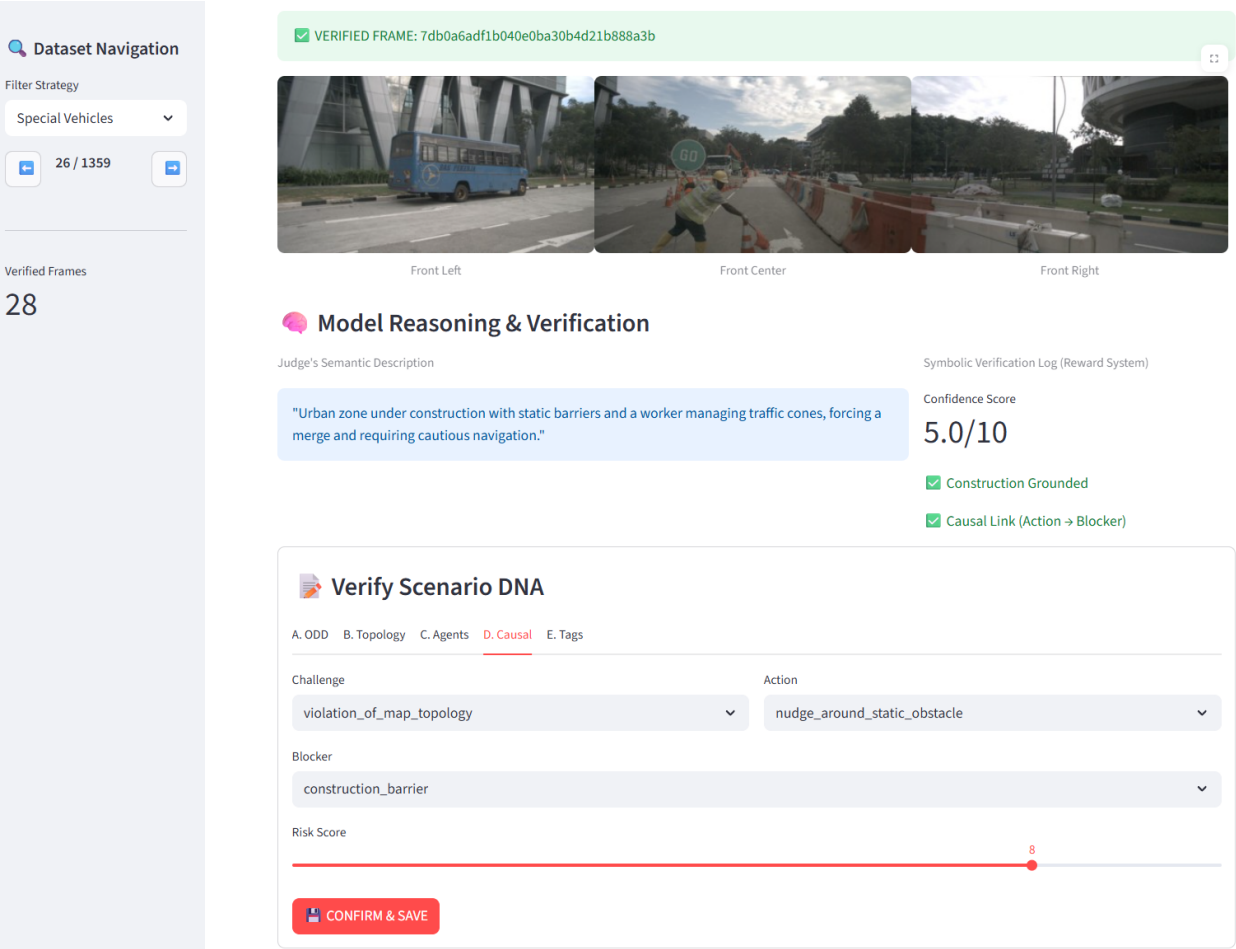


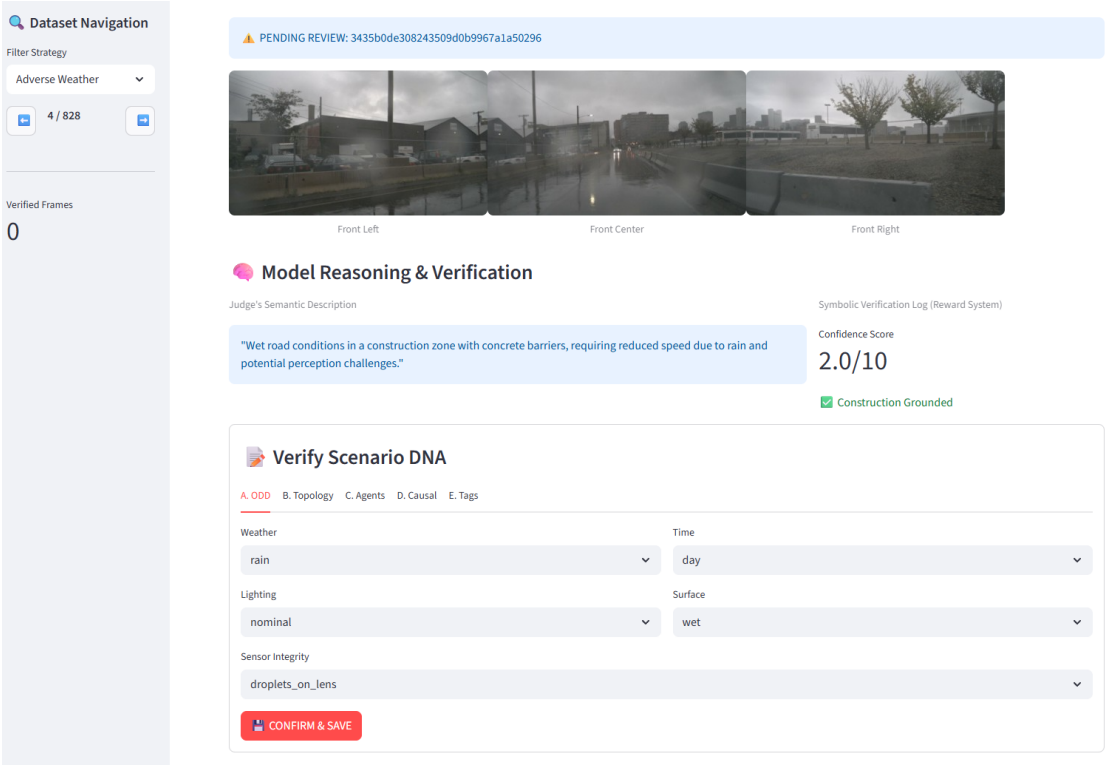
Figure 7: **The Semantic-Drive Curator Interface.** The tool facilitates "Verify-by-Exception." **Top:** The Consensus Judge provides a semantic description and a symbolic verification log (e.g., "VRU Grounded"). **Bottom:** The annotator corrects the structured "Scenario DNA" tags. In this example, the system correctly identifies a construction zone with a worker managing the traffic flow.

B.3 Detailed Breakdown of Gold Set Performance

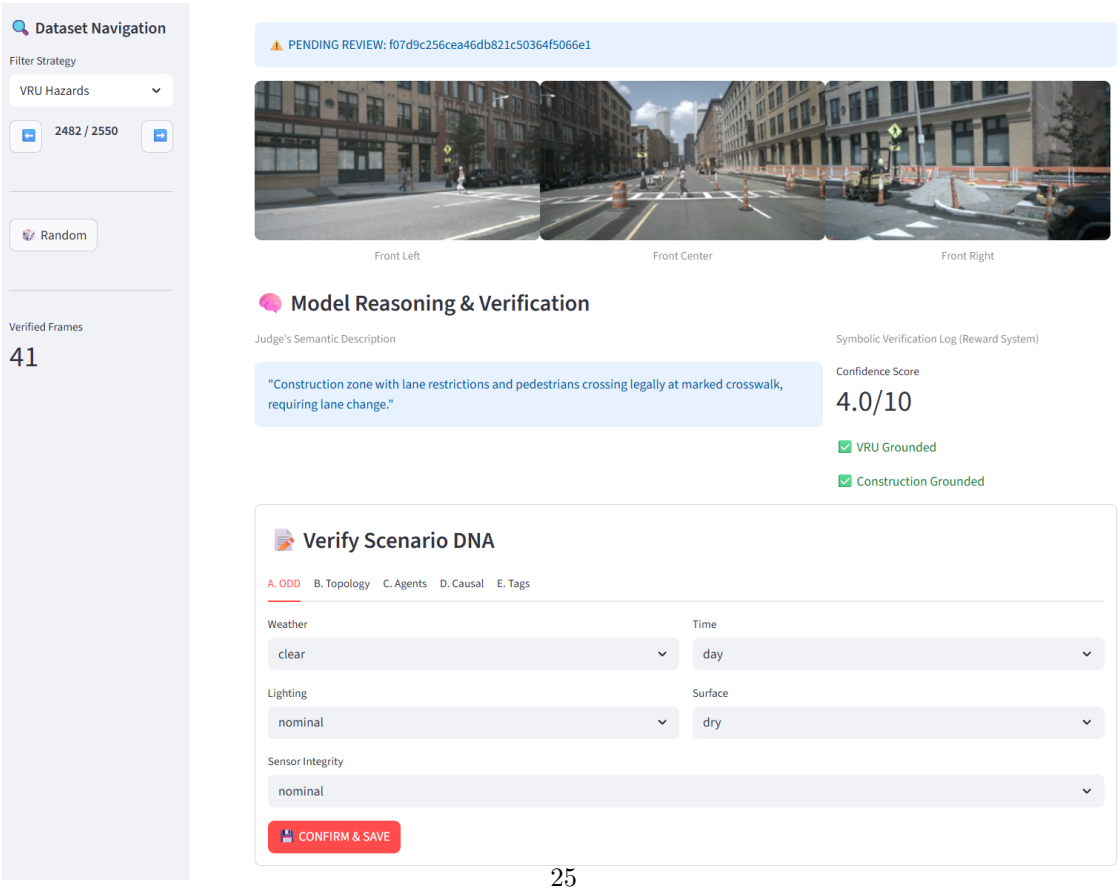
Table 5 details the specific performance of the Neuro-Symbolic architecture on the 108 curated test cases. "Correction" indicates instances where the VLM successfully rejected a False Positive from the symbolic detector or identified a semantic attribute (e.g., "Hesitation") missed by the detector.

Table 5: **Qualitative Failure Analysis on Gold Set.** Real-world examples illustrating the Neuro-Symbolic interaction. **Recovered:** The VLM correctly identified a rare class (Wheelchair) that standard detectors misclassified. **Success:** Correct consensus on complex static obstacles. **Corrected:** The VLM successfully contextualized a high-confidence YOLO detection (Pedestrians) as non-hazardous (Sidewalk), preventing a false alarm.

Token ID	**Primary Hazard**	**Symbolic (YOLO)**	**Cognitive (VLM)**	**Outcome**
`8104e0...`	VRU (Wheelchair)	"Cyclist" (Low Conf)	Id'ed Wheelchair User	**Recovered**
`dc73ce...`	FOD (Dumpster)	"Truck/Object"	Static Dumpster Blocking	**Success**
`7db0a6...`	Special Vehicle	"Bus"	School Bus (Stop Logic)	**Success**
`990723...`	False Pos. Risk	"3 Persons (0.80)"	Context: Safe on Sidewalk	**Corrected**
Aggregated	*Nominal Frames*	*Various Artifacts*	*Contextually Filtered*	**98% Acc**



(a) **Environmental ODD Verification.** The system identifies a "Rain" scenario with sensor degradation ("droplets_on_lens"). The interface allows the annotator to confirm these phenomenological attributes, which are critical for training de-hazing or robust perception models. Note the low risk score (2.0) despite the weather, as the road topology is open.



(b) **Construction zone identification with VRUs.** A "Green State" example where the human annotator has verified the scene. The system correctly flagged the construction zone and the pedestrians crossing legally at marked crosswalks.