SAVANT: Semantic Analysis with Vision-Augmented Anomaly deTection

Roberto Brusnicki, David Pop, Yuan Gao, Mattia Piccinini, Johannes Betz

Abstract— **Autonomous driving systems remain critically vulnerable to the long-tail of rare, out-of-distribution scenarios with semantic anomalies. While Vision Language Models (VLMs) offer promising reasoning capabilities, naive prompting approaches yield unreliable performance and depend on expensive proprietary models, limiting practical deployment. We introduce SAVANT (Semantic Analysis with Vision-Augmented Anomaly deTection), a structured reasoning framework that achieves high accuracy and recall in detecting anomalous driving scenarios from input images through layered scene analysis and a two-phase pipeline: structured scene description extraction followed by multi-modal evaluation. Our approach transforms VLM reasoning from ad-hoc prompting to systematic analysis across four semantic layers: Street, Infrastructure, Movable Objects, and Environment. SAVANT achieves 89.6% recall and 88.0% accuracy on real-world driving scenarios, significantly outperforming unstructured baselines. More importantly, we demonstrate that our structured framework enables a fine-tuned 7B parameter open-source model (Qwen2.5VL) to achieve 90.8% recall and 93.8% accuracy—surpassing all models evaluated while enabling local deployment at near-zero cost. By automatically labeling over 9,640 real-world images with high accuracy, SAVANT addresses the critical data scarcity problem in anomaly detection and provides a practical path toward reliable, accessible semantic monitoring for autonomous systems.**

I. INTRODUCTION

The widespread and safe deployment of autonomous vehicles (AVs) depends on their ability to respond well to the "long-tail" of rare, low-probability occurrences that are impossible to exhaustively collect and exhibit in training datasets [22]. Modern autonomous driving systems (ADS) work successfully in the usual ordinary casesyet remain brittle in unexpected cases, undermining public trust. Real world failures, such as mistaking a full moon for a traffic light or a billboard stop sign for a real one, highlight the gravity of the problem [16]. Figure 2 displays common semantic mismatches, where objects are misaligned with their context. These instances reflect critical failures of contextualization that threaten the safe adoption of autonomous vehicles.

Vision Language Models (VLMs) are pretrained on large image-text datasets, enabling contextual reasoning with broad world knowledge [1], [2]. With this rich semantics, VLMs can interpret complex scenes, reasoning about object relationships, spatial configuration, and context beyond current perception systems. Just as foundation models have emerged in large language models (LLMs), LLMs have already been explored for autonomous driving [3]. Recent

R. Brusnicki, D. Pop, Y. Gao, M. Piccinini, and J. Betz are with the Professorship of Autonomous Vehicle Systems, TUM School of Engineering and Design, Technical University of Munich, 85748 Garching, Germany; Munich Institute of Robotics and Machine Intelligence (MIRMI)
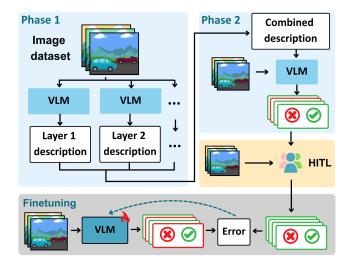
Fig. 1. Overview of the SAVANT framework. Driving images are processed in two phases: (1) Scene Description Extraction, where a VLM generates textual descriptions of scene layers (street, infrastructure, movable objects, environmental); (2) Scene Evaluation, where the original image and aggregated descriptions are jointly analyzed for anomaly classification. The resulting classifications undergo human verification and correction through Human-in-the-Loop (HITL) curation to create a high-quality training dataset, which is subsequently used to fine-tune an enhanced VLM capable of single-shot anomaly detection while maintaining compatibility with the original multi-phase framework.

developments in VLMs have provided promising ability in visual perception and natural language reasoning, suggesting potential to detect subtle contextual inconsistencies in driving anomalies.

However, a big limitation arises in using VLMs for safety-critical anomaly detection: unstructured prompting is unreliable, often requiring costly proprietary models. Simple queries like "Is this scene anomalous?" lack robustness, interpretability, and consistency. Our findings show that prompt-based strategies miss critical anomalies and yield false positives, while large proprietary models add prohibitive cost and latency for real-time use.

A. Related Work

1) Foundation Models for Driving Intelligence: Recent works have used foundation models in the context of autonomous driving in two ways that are complementary: as an end-to-end driving agent and as an advanced scene-understanding system. The first trend removes modularity, mapping sensor data directly to vehicle controls. LMDrive, DriveGPT4, and more recent models like EMMA and ImagiDrive exemplify VLM use, including complex driving characteristics with a closed loop methodology [8], [9], [10],

| (a) Moon as traffic light | (b) Traffic lights on truck |
| (c) Stop sign on billboard | (d) Police cars crossing diagonally |

Fig. 2. Examples of semantic anomalies in autonomous driving scenarios. These real-world cases demonstrate contextually unusual situations where individually recognizable objects combine to create potentially unsafe or confusing scenarios for autonomous systems.

[11]. Hybrid approaches, leverage LLM reasoning while integrating classical control (e.g., MPC) to guide vehicle behavior [12]. Other architectures, such as DriveVLM, LMAD, and DualAD, propose dual systems coupling VLM reasoning with classical perception stacks, highlighting the need for both deliberative and reactive control [13], [14], [15]. While powerful, these end-to-end models are black boxes, limiting safety verification and interpretability, and thus unsuitable for safety-critical monitoring [16], [17].

A different approach has utilized VLMs for inclusive, offline scene interpretation. DriveLM reformulates driving as a visual question answering task enabling multistep reasoning, such as analyzing object interactions and their future states [18]. Benchmarks like CODA-LM, NuRisk, and recent pedestrian behavior analyses push the limits of VLM scene understanding [20], [19], [21]. End-to-end systems, though rich in output, are unsuitable for real-time safety monitoring, while VQA systems remain too detached from control tasks. Our work fills this gap by repurposing VLM reasoning for online safety validation.

2) Detection of Semantic Anomaly and Out-of-Distribution Inputs: A foundational component of safety for an autonomous system is to monitor its input and react to situations deviating from training distributions. Beyond ML, this includes covariate shifts (e.g., noise, heavy rain) and contextual plausibility of novel objects and scene [22]. Our focus is on contextual shifts, requiring rich world understanding beyond statistical pattern matching.

Previous works used object detectors to extract a bag of objects, then applied text-only LLMs to assess plausibility from object co-occurrence [6], [7]. These methods were limited by reliance on text (ignoring rich visual data) and by validation mainly in simplified simulations.

Other works took data-centric approaches, using VLMs to mine corner cases or generative models to create challenging scenario [23], [24], [25], [26], [27]. These aim to improve training robustness with more diverse data, while our method provides a runtime safety net for anomalies persisting after training.

Prior research suffers from poor interpretability, text-only reasoning, or reliance on simulation. In contrast, we use multimodal VLMs directly on real-world images, bridging modal and simulation gaps. Our layered method, to our knowledge the first for automated semantic anomaly classification on real driving data, provides direct visual evidence for more reliable and trustworthy autonomous systems.

B. Critical Summary

The emergence of foundation models is transforming autonomous driving research, moving from modular pipelines to end-to-end system [1], [2], [3], [4]. This transition advances reasoning and scene understanding, but raises the challenge of reliability in rare long-tail events. We address this challenge by using a Vision-Language Model (VLM) to monitor semantic failures—scenes where object arrangements are contextually unsafe [6], [7], [5].

Current approaches suffer from several critical limitations: unstructured prompting lacks robustness, interpretability, and consistency; end-to-end models are black boxes unsuitable for safety-critical monitoring; previous methods rely on text-only reasoning while ignoring rich visual data; and existing solutions are validated mainly in simplified simulations rather than real-world scenarios. These limitations collectively prevent reliable deployment of VLM-based anomaly detection in safety-critical autonomous driving applications.

C. Contributions

To address the previous limitations, focusing in robust anomaly detection and practical deployment for real-world operation, the contributions of this paper are the following:

1) We propose **SAVANT**, a structured reasoning framework that shifts anomaly detection from unstructured prompting to principled analysis, achieving 89.6% recall and 88.0% accuracy through layered scene analysis and two-phase evaluation.

2) We provide a **systematic evaluation** analysis of 33 state-of-the-art VLMs across performance, cost, and efficiency dimensions, establishing practical deployment guidelines for semantic anomaly detection.

3) We demonstrate an **accessible deployment solution** through fine-tuned 7B open-source models reaching 90.8% recall and 93.8% accuracy, surpassing all other models while enabling cost-free local deployment without API dependencies.

4) We release **extensive research resources**[1] including complete framework implementation, optimized prompts for all evaluated models, fine-tuned models, web interface for efficient label correction, and extended CODA dataset with 9,640 annotated real-world driving images to facilitate further research in semantic anomaly detection.

---

[1]All resources will be made publicly available in an updated version of this preprint.

II. THE SAVANT FRAMEWORK

This section presents SAVANT, our structured reasoning framework for semantic anomaly detection in autonomous driving. We address the limitations of naive VLM prompting through a two-phase approach that decomposes complex driving scenes into analyzable semantic components.

A. Layered Anomaly Formulation

We formalize semantic anomaly detection as a binary classification task that transcends traditional object-level analysis. Given an input image I from an autonomous vehicle's forward-facing camera, our goal is to determine whether the scene contains contextually inappropriate arrangements of objects that could compromise system safety.

Unlike conventional out-of-distribution detection that identifies unknown objects, semantic anomalies involve familiar elements in contextually invalid configurations. We define a semantic anomaly as a scenario where individually recognizable objects appear in locations, states, or relationships that violate common-sense expectations about the driving environment. For example, traffic lights being transported on a truck represent a semantic anomaly: both "traffic light" and "truck" are common driving objects, but their combination creates a potentially dangerous context.

Our approach decomposes the complex task of scene-level anomaly detection into four semantic layers, each capturing distinct aspects of traffic scenes that contribute to anomalous situations:

Layer 1: Street - Road topology, geometry, surface conditions, and lane markings that define the driving surface and its structural integrity.

Layer 2: Infrastructure - Traffic control devices, signs, signals, and barriers that regulate traffic flow and provide guidance.

Layer 3: Movable Objects - Vehicles, pedestrians, and other dynamic entities that navigate through the environment.

Layer 4: Environment - Weather, lighting, and visibility conditions that affect scene perception and safety.

This hierarchical decomposition enables systematic analysis of complex driving scenarios while maintaining interpretability and allowing for targeted anomaly identification within specific semantic domains.

B. Two-Phase Anomaly Detection Pipeline

SAVANT employs a structured two-phase pipeline that transforms unstructured VLM analysis into systematic reasoning. Figure 1 illustrates this process, which addresses the core limitation of naive prompting approaches.

Phase 1: Structured Scene Description Extraction. Rather than directly querying for anomaly detection, we first guide the VLM to systematically describe the scene according to our four-layer decomposition. For each semantic layer $l \in \{1, 2, 3, 4\}$, we employ carefully designed prompt templates P_l that direct the model's attention to specific aspects of the scene:

$$D_l = \text{VLM}(I, P_l) \tag{1}$$

where D_l represents the textual description extracted for layer l. This structured extraction forces the model to examine each semantic aspect systematically, reducing the likelihood of overlooking critical details that might indicate anomalies.

The complete scene description aggregates information across all layers:

$$D_{scene} = \text{Aggregate}(D_1, D_2, D_3, D_4) \tag{2}$$

This phase serves multiple purposes: it ensures comprehensive scene coverage, provides interpretable intermediate representations, and creates rich textual context for the subsequent evaluation phase.

Phase 2: Multi-Modal Scene Evaluation. The second phase leverages both visual and textual information for robust anomaly classification. The VLM receives the original image I and the structured scene description D_{scene} as joint inputs:

$$\text{Classification} = \text{VLM}(I, D_{scene}, P_{eval}) \tag{3}$$

where P_{eval} is the evaluation prompt that instructs the model to analyze the scene for semantic inconsistencies using both visual evidence and the extracted textual descriptions. This multi-modal approach combines the richness of visual perception with the structured reasoning provided by the textual analysis.

The evaluation follows a systematic process: layer-wise anomaly assessment, cross-layer interaction analysis, and final binary classification with supporting rationale. This structured approach improves both accuracy and interpretability compared to direct prompting methods.

C. Fine-tuning Integration Strategy

While our two-phase pipeline achieves high accuracy, the requirement for multiple VLM queries limits real-time deployment feasibility. To address this challenge, we leverage SAVANT's high-quality outputs as an automated data annotation engine.

We apply our structured framework to automatically label large-scale datasets, generating high-quality training data that captures the nuanced reasoning process embedded in SAVANT's two-phase approach. Using this data, we fine-tune compact VLMs to internalize the structured reasoning process:

$$f_{fine-tuned}(I) = \text{VLM}_{fine-tuned}(I, P_{direct}) \tag{4}$$

This strategy enables us to distill SAVANT's multi-phase reasoning into an efficient single-shot model suitable for real-time deployment. The fine-tuned model maintains the benefits of structured analysis while achieving the computational efficiency required for practical autonomous system integration. Crucially, this approach enables smaller, open-source models to achieve performance levels that rival or exceed larger proprietary alternatives, providing a practical and accessible path toward widespread deployment.

III. EXPERIMENTAL SETUP

This section describes our experimental design for evaluating SAVANT's structured reasoning approach across multiple dimensions: performance against baselines, scalability across VLM architectures, and practical deployment.

A. Datasets

We build three incremental datasets, as described below.

CODALM_small (Model Selection). We begin evaluation with CODALM_small, a dataset that we curate to contain 100 real-world driving images (50 anomalous, 50 normal) derived from the CODA corner case dataset [30]. Each image receives manual annotation with detailed textual scene descriptions and anomaly evaluations, ensuring a high-quality ground truth. This dataset enables our initial scanning across 30+ VLM candidates: the small size of the dataset allows us to identify the best models of each family without excessive resource consumption, while providing sufficient data for measuring average response times and API costs.

CODALM_medium (Comparative Evaluation). To validate our approach at scale, we created CODALM_medium by combining automated framework evaluation with human expert validation. Starting with the full CODA dataset (9,640 images), we used Gemini-2.0-Flash-Exp (the top-performing model in our evaluation at that time) to generate scene descriptions and anomaly evaluations. From these, 5,078 annotations were reviewed and corrected by two human evaluators, resulting in a high-quality balanced dataset with validated ground truth. For comparative experiments, 1,020 examples were selected to form a balanced subset, while the remaining served as the fine-tuning dataset. Figure 3 shows the anomaly distribution across CODALM_medium: 60.9% contain anomalies spanning four semantic layers. Movable Objects anomalies are most frequent (81.7%), followed by Street Layer (44.2%), Infrastructure (39.7%), and Environmental (18.3%). The dataset also captures varying complexity levels, from single-layer anomalies (27.4%) to quad-layer anomalies (2.6%), providing broad coverage for training robust detection models.
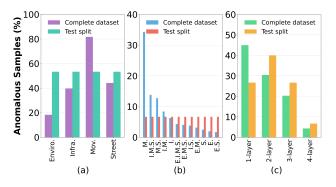

Fig. 3.   Layer-wise anomaly distribution comparing CODALM_medium dataset (5,078 samples) and its test split (1020 samples):(a) Individual layer frequency across the four semantic layers, (b) Anomaly layer combination frequency, (c) Multi-layer anomaly frequency distribution.

CODALM_large (Framework Application). Finally, we release the complete 9,640-image CODALM_large dataset, fully labeled using SAVANT with the best-performing VLM. This dataset represents the largest semantically-annotated dataset for anomaly detection in autonomous driving, and demonstrates our framework's capability as a scalable data annotation engine.

B. Models Evaluated

Our evaluation encompasses both proprietary and open-source VLMs to provide baseline comparisons and assess the accessibility of different approaches.

Proprietary Models. We evaluate the best available models from major commercial VLM families accessed via API, including representatives from Google Gemini, OpenAI GPT, Anthropic Claude, Mistral, and Qwen-VL families. These models represent the current pinnacle of VLM capabilities but require ongoing costs and external dependencies.

Open-source Models. We include leading open-source alternatives from key model families, including Qwen2.5-VL variants, Mistral models, Pixtral, Gemma, and LLaVA architectures spanning different parameter scales from 7B to 72B. This comparison shows the broad applicability of SAVANT across diverse model architectures.

C. Evaluation Methods

We evaluate multiple configurations to establish baseline comparisons, determine the best setup to maximize the performance, and validate improvements:

Baseline Methods:

- `image_baseline`: Direct VLM prompting with images only, representing the naive approach of asking: "Is this traffic scene anomalous? Yes or No." This baseline captures unstructured, single-shot VLM performance without our layered reasoning framework.
- `text_baseline`: Uses only unstructured scene descriptions for evaluation, without visual grounding or layered analysis.
- `baseline`: Combines unstructured scene description with image-based evaluation, but without our four-layer description decomposition.

Structured Methods:

- `image`: Direct image analysis using structured four-layer decomposition without textual scene descriptions.
- `text`: Uses structured scene descriptions from Phase 1 with layered evaluation, but without visual grounding in Phase 2.
- `full`: Complete SAVANT pipeline combining structured scene descriptions with multi-modal layered evaluation.
- `*_opt`: Optimized versions using DSPy [28] MIPROv2 prompt optimization for enhanced performance.

Table I summarizes the key properties of each evaluation method, including input modalities, structural components, and computational requirements.

These configurations enable systematic analysis of our framework's core innovations: structured reasoning, multi-modal evaluation, layered scene decomposition, and prompt optimization effects.

TABLE I

EVALUATION CONFIGURATIONS AND THEIR PROPERTIES.

Method	Input	L	O	#Q
`image_baseline`	image	✗	✗	1
`text_baseline`	scene description	✗	✗	2
`baseline`	image + scene description	✗	✗	2
`image`	image	✓	✗	1
`text`	scene description	✓	✗	5
`text_opt`	scene description	✓	✓	5
`full`	image + scene description	✓	✗	5
`full_opt`	image + scene description	✓	✓	5

L = Layered analysis, O = Optimization, #Q = Number of queries

D. Evaluation Metrics and Implementation

Performance Metrics. We report precision, recall, F1-score, and accuracy across all experiments. For safety-critical anomaly detection, recall is paramount as missing true anomalies (false negatives) poses greater risk than false alarms.

Image Resolution. All models process images at 360p resolution, representing an optimal trade-off between performance and computational efficiency based on our comprehensive resolution analysis presented in Section IV.

Efficiency Metrics. We measure inference time (average seconds per query) and cost analysis (average number of input/output tokens per query) to assess practical deployment feasibility.

Implementation Details. Experiments utilize DSPy for prompt optimization, applying MIPROv2 optimization with few-shot examples where applicable. Fine-tuning employs LoRA (Low-Rank Adaptation) for parameter-efficient training. Local model evaluation uses standard GPU hardware to ensure reproducible results.

IV. RESULTS AND ANALYSIS

We demonstrate SAVANT's effectiveness through experimental evidence showing that structured reasoning and DSPy optimization improve VLM anomaly detection performance and enable accessible deployment through fine-tuning.

A. Baseline Performance: Image-Only Evaluations

Table II presents our baseline evaluation across 32 state-of-the-art VLMs using the `image_baseline` method—direct image-only prompting without SAVANT's structured reasoning framework. Models receive only visual input with a simple prompt asking "Is this traffic scene anomalous? Yes or No." Response times T represent the average inference time per query, while token counts reflect the total tokens consumed (input + output) per evaluation. This baseline assessment establishes the performance floor for unstructured approaches and identifies the top-performing models of each family for subsequent structured reasoning evaluation.

This evaluation reveals clear performance hierarchies: top proprietary models (Gemini 2.5 Pro: 85%, GPT-5: 83%) outperform open-source alternatives (Qwen2.5-VL 72B: 75%,

TABLE II

COMPARISON OF VLMS FOR IMAGE-ONLY ANOMALY DETECTION AT 360P RESOLUTION. BEST SCORES WITHIN EACH GROUP ARE SHOWN IN **BOLD**, SECOND-BEST SCORES ARE <u>UNDERLINED</u>.

Proprietary Models	Accu.	Prec.	Rec.	F1	T(s)	Tokens
Gemini 2.5 Pro	**0.85**	0.94	0.70	<u>0.80</u>	24.9	938
GPT-5	<u>0.83</u>	0.88	<u>0.77</u>	**0.82**	28.7	51851
Gemini 2.5 Flash	0.81	0.91	0.69	0.78	10.6	947
Gemini 1.5 Flash	0.81	0.90	0.70	0.79	2.9	958
Mistral Medium 3.1	0.80	0.80	**0.80**	0.80	7.9	1259
Mistral Medium 3	0.81	0.94	0.66	0.78	6.4	1158
Qwen-VL Max	0.79	0.87	0.68	0.76	7.7	2193
GPT-4o	0.78	0.94	0.60	0.73	7.4	1032
Gemini 2.5 Pro Prev	0.77	0.87	0.64	0.74	23.6	**898**
Claude 3.5 Sonnet	0.74	0.79	0.66	0.72	7.3	1109
Gemini 2.0 Flash Exp	0.73	0.80	0.62	0.70	**2.6**	2441
Gemini 2.0 Thinking	0.75	0.90	0.56	0.69	8.1	<u>937</u>
Gemini 2.5 Flash Prev	0.75	0.90	0.56	0.69	11.2	956
GPT-4 Turbo	0.73	**0.96**	0.48	0.64	7.6	1055
Claude 3.5 Haiku	0.72	0.89	0.50	0.64	7.0	1144
Claude Sonnet 4	0.70	0.83	0.50	0.63	9.8	1145
Claude Opus 4.1	0.68	<u>0.95</u>	0.38	0.54	19.4	1238
Claude 3.7 Sonnet	0.65	0.80	0.40	0.53	10.8	1276
GPT-4.1 Mini	0.66	0.94	0.34	0.50	5.4	1075
GPT-4o Mini	0.65	0.90	0.34	0.49	4.4	14787
Claude Opus 4	0.62	0.83	0.31	0.45	13.1	1224
Qwen-VL Plus	0.58	0.90	0.18	0.30	<u>2.8</u>	2155
GPT-4.1 Nano	0.56	0.80	0.16	0.27	4.3	1262
Open Models	**Accu.**	**Prec.**	**Rec.**	**F1**	**T(s)**	**Tokens**
Qwen2.5-VL 72B	**0.75**	0.82	0.64	<u>0.72</u>	9.4	<u>961</u>
Mistral Small 3.1	<u>0.74</u>	0.71	**0.82**	**0.76**	11.9	1057
Pixtral Large 2411	0.67	0.64	<u>0.78</u>	0.70	5.8	1817
Mistral Small 3.2	0.70	0.78	0.56	0.65	5.5	996
Qwen2.5-VL 32B	0.62	0.65	0.53	0.59	19.9	2380
Gemma3 12B	0.44	0.46	0.70	0.56	9.2	1254
Pixtral 12B	0.64	**1.00**	0.28	0.44	5.4	1637
LLaVA 1.5 7B	0.58	0.79	0.22	0.34	6.7	1029
Qwen2.5-VL 7B	0.55	0.55	0.48	0.52	<u>4.2</u>	**686**
Qwen2.5-VL 3B	0.50	0.50	0.29	0.37	**3.6**	1489

Mistral Small 3.1: 74%), with inference times varying from 2.6 to 24.9 seconds.

Figure 4 shows performance scores (F1, accuracy, precision, recall) for the best-performing models of each family evaluated on a balanced split of one thousand examples across resolutions of 180p, 240p, 360p, 540p, and 720p. The analysis reveals significant performance improvements up to 360p for most models, with only marginal gains from 360p to 540p or 720p. Considering the substantial increase in token consumption costs for higher resolutions (2.25x for 540p, 4x for 720p), 360p represents the optimal balance between performance and efficiency.

B. Structured Reasoning Outperforms Naive Prompting

Our first key finding demonstrates that SAVANT's layered analysis approach significantly outperforms unstructured baseline methods across all evaluated VLMs. Table III presents our core performance comparison using Gemini-2.0-Flash-Exp, selected over the higher-performing Gemini-2.5-Pro due to cost considerations[1]

---

[1]Our extensive evaluation would incur prohibitive costs, while Gemini-2.0-Flash-Exp was available at no cost during our evaluation period.

TABLE III

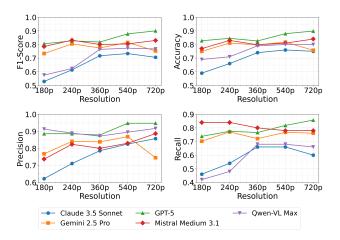| Method	Gemini-2.0-FE				Qwen2.5-VL-72B				Qwen2.5-VL-32B				Qwen2.5-VL-7B				Qwen2.5-VL-3B			
	Acc.	Rec.	Prec	F1	Acc.	Rec.	Prec	F1	Acc.	Rec.	Prec	F1	Acc.	Rec.	Prec	F1	Acc.	Rec.	Prec	F1
image_baseline	0.73	0.62	0.79	0.70	0.75	0.64	0.82	0.72	0.62	0.53	0.65	0.59	0.55	0.48	0.55	0.52	0.50	0.29	0.50	0.37
text_baseline	0.65	0.40	0.80	0.53	0.69	0.48	**0.85**	0.61	0.61	0.43	**0.67**	0.53	0.59	0.39	**0.65**	0.49	0.50	0.35	0.50	0.41
baseline	0.77	0.66	0.85	0.74	0.73	0.56	0.83	0.67	0.65	0.72	0.63	0.67	0.60	0.46	0.64	0.53	0.56	**0.47**	0.57	0.52
image	0.78	**0.94**	0.71	0.81	0.69	**0.86**	0.64	0.74	0.64	0.62	0.65	0.63	0.45	0.36	0.43	0.39	0.46	0.28	0.44	0.34
text	0.77	0.92	0.71	0.80	0.79	0.84	0.76	0.80	0.60	0.48	0.62	0.54	0.43	0.31	0.40	0.35	0.51	0.22	0.52	0.31
text_opt	0.80	0.86	0.77	0.81	**0.82**	0.82	0.82	**0.82**	0.64	0.71	0.62	0.66	**0.61**	0.53	0.64	0.58	0.51	0.45	0.51	0.48
full	0.85	0.90	0.82	0.86	0.80	0.78	0.81	0.80	0.64	0.74	0.62	0.67	0.59	0.53	0.60	0.56	0.45	0.30	0.43	0.35
full_opt	**0.88**	0.90	**0.87**	**0.88**	0.82	0.84	0.81	0.82	**0.66**	**0.75**	0.63	**0.69**	0.60	**0.65**	0.64	**0.62**	0.59	0.45	0.63	0.52

Fig. 4. Resolution comparison analysis across different image resolutions for the best-performing models of each family. Performance shows significant improvements up to 360p with diminishing returns (or worse performance) at higher resolutions.

When paired with Gemini-2.0-Flash-Exp, SAVANT's full framework achieves 90% recall and 85% accuracy, representing a 36% relative improvement in recall over the two-phase baseline (absolute improvement of 24%). Most notably, the DSPy-optimized version reaches 88% accuracy while maintaining 90% recall, demonstrating that structured reasoning enables reliable anomaly detection even in safety-critical scenarios where missing true positives carries significant risk.

The bigger improvement in text-only scenarios (40% to 92% recall) highlights the importance of structured scene descriptions. Our layered approach captures critical semantic information that unstructured descriptions miss, proving that systematic decomposition enhances VLM reasoning capabilities.

However, our extensive evaluation across the Qwen2.5-VL family reveals a persistent performance gap between proprietary and open-source models. Despite extensive experimentation with multiple model sizes (from 3B up to 72B parameters) and evaluation methods, the best-performing open model (Qwen2.5-VL-72B with 82% F1) falls short of the proprietary baseline (Gemini-2.0-Flash-Exp with 88% F1). This gap motivated our investigation into fine-tuning approaches to achieve competitive performance with locally deployable, cost-effective solutions.

C. The Critical Importance of Multi-Modal Evaluation

Our second key finding establishes that combining visual and textual information in SAVANT's Phase 2 evaluation provides superior performance compared to text-only reasoning. This validates our design choice of multi-modal evaluation rather than purely textual analysis after the description extraction in Phase 1.

In Table III, our layered approach shows substantial improvements in image-only evaluation for high-capacity models, where recall increases from 62% to 94% for Gemini-2.0-Flash-Exp when VLMs are guided to analyze specific semantic layers before classification. This demonstrates that explicit reasoning guidance enhances VLM performance on complex visual scenes for models with sufficient representational capacity, though the effect varies across model architectures.

Our multi-modal approach leverages both the richness of visual perception and the structured reasoning provided by textual analysis, consistently outperforming single-modality alternatives across all evaluated configurations.

D. Model Scale and Optimization Effectiveness

Table III reveals critical insights about the relationship between model scale and optimization effectiveness. DSPy optimization shows varying benefits across model sizes: while the 72B model benefits from optimization (text: 80% → text_opt: 82% F1), smaller models show mixed results. The 7B model demonstrates substantial optimization gains (text: 35% → text_opt: 58% F1), while the 3B model shows moderate improvement (text: 31% → text_opt: 48% F1), indicating that optimization effectiveness depends on both model capacity and the specific optimization target.

This scaling relationship has profound implications for deployment strategies. Larger models (72B, 32B) can leverage sophisticated reasoning frameworks effectively, while smaller models (7B, 3B) require alternative approaches such as fine-tuning to achieve competitive performance. The consistent performance degradation from 72B (82% F1) → 32B (69% F1) → 7B (62% F1) → 3B (52% F1) establishes clear trade-offs between model accessibility and task performance.

Furthermore, structured reasoning methods (image, text, full) generally outperform their baseline counterparts across model architectures, demonstrating that explicit decomposi-

tion of the reasoning process benefits VLMs. However, the magnitude and consistency of improvement varies significantly with model capacity and architecture, reinforcing the importance of model selection for deployment scenarios with varying computational constraints.

E. Fine-Tuned Open Model Outperforms Proprietary Ones

Our most significant contribution addresses the practical deployment challenge through fine-tuning. Table IV presents our headline result: SAVANT enables a fine-tuned 7B parameter open-source model to achieve performance that rivals proprietary alternatives while enabling local deployment.

TABLE IV
PERFORMANCE COMPARISON OF FINE-TUNED MODELS VERSUS BASELINES AND TOP PROPRIETARY MODELS.

Model	Acc.	Rec.	Prec.	F1	Q
Qwen2.5-VL-7B (image_baseline)	0.546	0.484	0.553	0.516	1
Qwen2.5-VL-7B (Fine-tuned)	**0.938**	**0.908**	**0.967**	**0.936**	1
Qwen2.5-VL-7B (Pipeline FT)	0.837	0.818	0.851	0.834	2
Gemini-2.0-FE (full_opt)	0.880	0.896	0.860	0.878	5
Gemini 2.5 Pro (image_baseline)	0.850	0.700	0.940	0.800	1
GPT-4o (image_baseline)	0.780	0.600	0.940	0.730	1
Claude 3.5 Sonnet (image_baseline)	0.740	0.660	0.790	0.720	1

Q = Number of queries

Our fine-tuning approach yields two complementary solutions. The single-shot model performs anomaly detection with only one query, directly classifying images without intermediate steps. The pipeline model preserves SAVANT's two-phase structure (scene description extraction followed by multi-modal evaluation) using fine-tuned components. Evaluating these approaches, the single-shot model achieves 93.8% accuracy and 90.8% recall, while the pipeline model reaches 83.7% accuracy and 81.8% recall. Both dramatically improve recall from the 48.4% baseline—1.9× and 1.7× improvements respectively—while maintaining computational efficiency.

Notably, the single-shot model outperforms the pipeline variant despite the latter's use of structured reasoning that proved beneficial in our framework evaluation. This counterintuitive result likely stems from the increased training complexity of the pipeline approach, which must learn both scene description generation and multimodal anomaly detection simultaneously. Both models were trained with identical hyperparameters and epochs, but the pipeline model's additional complexity may require extended training to reach its full potential, suggesting that further optimization could yield even higher performance.

The expanded comparison in Table IV demonstrates that our single-shot fine-tuned model (93.8% accuracy) surpasses all proprietary baselines including Gemini 2.5 Pro (85% accuracy), GPT-4o (78% accuracy), and Claude 3.5 Sonnet (74% accuracy). Our pipeline fine-tuned model (83.7% accuracy) also outperforms GPT-4o and Claude 3.5 Sonnet while requiring only two queries compared to the five queries needed by Gemini-2.0-FE's full optimized approach.

This comparison shows that our fine-tuned models substantially outperform their baseline version and achieve competitive results that match or exceed proprietary alternatives. The performance transformation from baseline (54.6% accuracy, 48.4% recall) to fine-tuned variants (84-94% accuracy, 82-91% recall) demonstrates a significant improvement in accessibility for practical anomaly detection deployment.
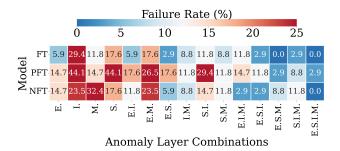

Fig. 5. Failure rates (%) across semantic layer combinations for three Qwen2.5-VL-7B variants: FT (Fine-Tuned), PFT (Pipeline Fine-Tuned), and NFT (Non-Fine-Tuned). Layer abbreviations: S (Street), I (Infrastructure), M (Movable Objects), E (Environmental). Multi-layer combinations (e.g., E.S.I.M) represent anomalies spanning multiple semantic contexts.

F. Layer-Specific Error Analysis

To understand how fine-tuning affects error patterns across SAVANT's semantic decomposition, we analyze layer-wise failure rates for the three Qwen2.5-VL-7B variants from Table IV. Figure 5 compares the fine-tuned single-shot (FT), pipeline fine-tuned (PFT), and non-fine-tuned baseline (NFT) models across all anomaly layer combinations.

The FT model exhibits the lowest failure rates, remaining below 10% for most anomalies, particularly across Street (S) and Environment (E) layer combinations. This demonstrates that fine-tuning effectively internalizes SAVANT's structured reasoning. In contrast, the NFT baseline shows high error rates exceeding 30% for most single-layer anomalies—especially Infrastructure (I) and Movable Objects (M)—confirming that without fine-tuning, the 7B model struggles with semantic layer associations.

The PFT variant achieves moderate gains over baseline but exhibits elevated failure rates ($\approx$ 40%) for Infrastructure-related anomalies (I, I.M, S.I.), supporting our observation that two-phase reasoning complexity may hinder training convergence.

Across all variants, Environmental (E) anomalies remain most difficult due to their subtle nature (e.g., lighting, fog). Interestingly, failure rates decrease for multi-layer scenarios (e.g., E.S.I.M), where cross-layer cues improve detection through redundant semantic evidence.

This analysis reinforces two key findings: (1) single-shot fine-tuning enables robust, context-aware detection across all semantic layers, and (2) Environmental conditions remain the primary challenge—requiring future work on visual-semantic reasoning under adverse conditions.

V. CONCLUSION

In this paper, we addressed the critical challenge of semantic anomaly detection in autonomous driving, where data scarcity and unreliable VLM performance have hindered practical deployment. We introduced SAVANT, a structured reasoning framework that transforms ad-hoc VLM prompting into systematic analysis across four semantic layers: Street, Infrastructure, Movable Objects, and Environment. Through comprehensive evaluation of 33 state-of-the-art VLMs, we demonstrated that structured reasoning significantly outperforms unstructured baselines, with SAVANT achieving 89.6% recall and 88.0% accuracy.

Our framework addresses multiple critical gaps simultaneously: it enables researchers to efficiently evaluate large numbers of models, provides systematic prompt optimization through DSPy integration, and facilitates efficient human curation of model-generated annotations. This multi-faceted approach produces high-quality labeled data that enables fine-tuning open-source models to achieve state-of-the-art performance at dramatically reduced costs. Most significantly, our fine-tuned 7B Qwen2.5-VL model achieves 90.8% recall and 93.8% accuracy, surpassing all evaluated proprietary models while enabling cost-free local deployment.

By automatically annotating 9,640 real-world driving images and demonstrating scalable data curation workflows, SAVANT provides a practical solution to data scarcity in semantic anomaly detection. This work establishes a foundation for accessible, reliable safety monitoring that can accelerate autonomous driving research and deployment. For future work, we plan to extend SAVANT to temporal analysis through video input and validate the framework through real-world on-vehicle integration.

REFERENCES

[1] Z. Yang, R. Li, X. Wen, H. Zhang, B. Zheng, R. Zheng, C. Wen, J. Xu, M. Yang, and K. Jia, "LLM4Drive: A Survey of Large Language Models for Autonomous Driving," in 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024.

[2] Y. Zhou, Q. Yuan, K. Chen, Z. Tian, and H. Li, "Vision-Language Models for Autonomous Driving: A Survey," arXiv preprint arXiv:2402.03756, 2024.

[3] C. Jin, Z. Zhou, J. Li, X. Zhu, Z. Zhou, J. Lu, L. Wang, Y. Qiao, Y. Wang, and J. Yan, "Large Language Models for Autonomous Driving (LLM4AD): Concept, Benchmark, Experiments, and Challenges," arXiv preprint arXiv:2410.15281, 2024.

[4] Y. Gao, M. Piccinini, Y. Zhang, D. Wang, K. Moller, R. Brusnicki, B. Zarrouki, A. Gambi, J. F. Totz, K. Storms, S. Peters, A. Stocco, B. Alrifaee, M. Pavone, and J. Betz, "Foundation Models in Autonomous Driving: A Survey on Scenario Generation and Scenario Analysis," arXiv preprint arXiv:2506.11526, 2025.

[5] D. Pop, "LENS-AD: A Foundation Model-based Safety Monitor for Semantic Anomaly Detection in Autonomous Driving," M.S. thesis, Technical University of Munich, 2025.

[6] M. Elhafsi, A. Brem, and P. C. Gembarski, "Semantic Anomaly Detection for Autonomous Driving," in 2023 IEEE International Conference on Omni-layer Intelligent Systems (COINS), 2023, pp. 1–8.

[7] A. Sinha and A. Choudhury, "Real-time Semantic Anomaly Detection and Novelty Identification in Autonomous Driving," arXiv preprint arXiv:2404.05312, 2024.

[8] H. Shao, Y. Li, L. Li, S. Liu, H. Chen, X. Qi, K. Liu, C. Li, Y. Ge, A. Anandkumar, and others, "LMDrive: Closed-Loop End-to-End Driving with Large Language Models," arXiv preprint arXiv:2312.07488, 2023.

[9] Z. Xu, Y. Han, Z. Zhang, Z. Wang, S. Ge, H. Xu, and L. Li, "DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model," arXiv preprint arXiv:2310.01412, 2023.

[10] E. Erlich, G. Sharir, A. Noy, I. Schwartz, Y. Friedman, Y. Chai, and D. He, "EMMA: End-to-End Multimodal Model for Autonomous Driving," arXiv preprint arXiv:2410.23262, 2024.

[11] Y. Wu, J. Zhang, Z. Lin, Z. Zhou, J. Yan, and Y. Qiao, "ImagiDrive: A Unified Imagination-and-Planning Framework for Autonomous Driving," arXiv preprint arXiv:2508.11428, 2025.

[12] W. Sha, Y. Chen, B. Li, Q. Cui, Z. Chen, B. Li, and D. Zhao, "LanguageMPC: Large Language Models as Decision Makers for Autonomous Driving," arXiv preprint arXiv:2310.04301, 2023.

[13] Z. Tian, K. Chen, Y. Zhou, and H. Li, "DriveVLM: The Finding of VLM's Strong ZERO-SHOT Planning Capabilities in Autonomous Driving," arXiv preprint arXiv:2403.03928, 2024.

[14] X. Chen, Z. Liu, Z. Zhang, L. Zhang, Z. Wu, and T. Zhang, "LMAD: Integrated End-to-End Vision-Language Model for Explainable Autonomous Driving," arXiv preprint arXiv:2508.12404, 2025.

[15] D. Wang, M. Kaufeld, and J. Betz, "DualAD: Dual-Layer Planning for Reasoning in Autonomous Driving," arXiv preprint arXiv:2409.18053, 2024.

[16] Microsoft, "Failure modes in machine learning," Microsoft Learn, 2023. [Online]. Available: https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

[17] Y. Li, A. Liu, and J. Yang, "A Comprehensive Survey on Physical Risk Control for Foundation Model-enabled Robotics," arXiv preprint arXiv:2505.12583, 2025.

[18] C. Sima, K. Renz, K. Chitta, L. Chen, and A. Geiger, "DriveLM: Driving with graph visual question answering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22091–22102.

[19] Y. Gao, M. Piccinini, R. Brusnicki, Y. Zhang, and J. Betz, "NuRisk: A Visual Question Answering Dataset for Agent-Level Risk Assessment in Autonomous Driving," arXiv preprint arXiv:2509.25944, 2025.

[20] J. Chen, C. Singh, D. Chen, A. Vijayaraghavan, and S. Manivasagam, "Automated Driving Systems Data (CODA-LM): A Labeled Video Dataset for Training and Benchmarking LVLMs in Autonomous Driving," arXiv preprint arXiv:2402.10375, 2024.

[21] Y. Shu, Z. Zhou, Z. Liu, and J. Wang, "Application of Vision-Language Model to Pedestrians Behavior and Scene Understanding in Autonomous Driving," arXiv preprint arXiv:2501.06680, 2025.

[22] J. Yang, J. Zhou, Y. Liu, J. Chen, and Y.-G. Wang, "Generalized Out-of-Distribution Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, 2024.

[23] Z. Hu, L. Zhang, R. Yang, Z. Li, X. Li, and L. Li, "VLM-C4L: A VLM-based framework for continuous corner case learning in autonomous driving," arXiv preprint arXiv:2502.04321, 2025.

[24] A. Hu, G. Stan, T. Pavlov, M. Cvitkovic, S. Pang, A. Rusu, F. Viola, P. Munk, O. Vinyals, T. Lillicrap, and others, "GAIA-1: A Generative World Model for Autonomous Driving," in Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[25] X. Wang, Z. Xie, A. Zhu, G. Yu, W. Li, W.-X. Chu, G. Chen, L. Wang, H. Li, and H. Yu, "DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving," arXiv preprint arXiv:2309.09777, 2023.

[26] D. Wang, Z. Sun, Z. Li, C. Wang, Y. Peng, H. Ye, B. Zarrouki, W. Li, M. Piccinini, L. Xie, and J. Betz, "Enhancing Physical Consistency in Lightweight World Models," arXiv preprint arXiv:2509.12437, 2025.

[27] A. Su, L. Yang, C. Li, and J. Wang, "Generating Multimodal Driving Scenes via Next-Scene Prediction," arXiv preprint arXiv:2503.14945, 2025.

[28] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, and C. Potts, "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," in The Twelfth International Conference on Learning Representations, 2024.

[29] O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, and M. Zaharia, "Demonstrate-Search-Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP," arXiv preprint arXiv:2212.14024, 2022.

[30] K. Li, K. Chen, H. Wang, L. Hong, C. Ye, J. Han, Y. Chen, W. Zhang, C. Xu, D.-Y. Yeung, X. Liang, Z. Li, and H. Xu, "CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving," arXiv preprint arXiv:2203.07724, 2022.