Argonaut: Agentic Scenario Mining for Autonomous Driving

Zhijie Qiao Xiangzhong Ye Kunda Yang Jiawei Wang Henry X. Liu
University of Michigan

Abstract

Scenario mining retrieves events of interest from autonomous vehicle fleet logs via natural-language descriptions, supporting downstream tasks such as safety validation, world model training, and synthetic data generation. This task spans diverse spatial, temporal, and visual conditions that demand both scale and accuracy. To this end, we present Argonaut, an agentic scenario mining workflow that pairs scalable rule-based geometric matching with fine-grained vision-language understanding. Argonaut outperforms both official baselines across all metrics on the validation and test sets of the CVPR 2026 Argoverse 2 Scenario Mining Competition.

1. Introduction

Autonomous vehicle fleets continuously generate large volumes of driving data, yet the most valuable events, such as safety-critical and long-tail scenarios, are inherently rare. Scenario mining addresses this by retrieving targeted events from fleet logs via natural-language descriptions, enabling downstream tasks such as safety validation, world model training, and synthetic data generation at scale. Manual annotation cannot keep pace with fleet-sized data volumes, motivating LLM- and VLM-based methods for automated retrieval, though precisely and reliably localizing scenes that match a given description remains an open challenge.

To advance research in this direction, the CVPR 2026 Argoverse 2 Scenario Mining Challenge [6] pairs natural-language descriptions with driving logs for spatial and temporal localization. Its official baseline in 2025, RefAV [1], employs LLM coding agents to write scenario-matching functions from these descriptions, operating over Le3DE2E [5] perception tracker outputs and map data. While this avoids expensive manual annotation, unconstrained code generation yields inconsistent implementations across similar descriptions and is difficult to audit or verify. Furthermore, functions operating over tracker outputs have no access to visual-only cues (weather, road conditions, traffic-signal state, or actor appearance) that reside exclusively in the camera streams.

To address these limitations, we propose an agentic scenario mining method that systematically combines the strengths of scalable rule-based geometric matching and fine-grained VLM-based visual reasoning. Four major components are as follows:

- **Scenario Routing.** A thinking agent interprets each scenario description, identifying the referred and related actor types and routing it to either rule-based geometric matching or VLM-based visual reasoning.
- **Atomic Decomposition.** A coding agent develops a shared library of rule-based atomic functions and iteratively refines them on the validation split.
- **Bird's-eye-view (BEV) Verification.** A VLM-based verification agent reviews the rule-stage output through rendered BEV videos to mitigate borderline cases, tracker noise, and approximate geometric matches.
- **Camera Analysis.** A VLM-based analysis agent examines raw camera streams to process scenario descriptions that require visual cues the rule stage cannot capture.

2. Methodology

Scenario Routing. We use a GPT-5.5-based [3] thinking agent to identify the referred and related actor types in each scenario description and assign it to one of two paths. When its conditions can be evaluated from tracked geometry, the description is routed to the rule stage and composed from atomic functions, as detailed below. When it depends on explicit visual cues, it is routed to the camera stage for visual inference by a VLM.

Atomic Decomposition. Following the official baselines, we use Argoverse 2 static map annotations and Le3DE2E object tracks aligned with ego poses and evaluation timestamps. To make the rule-based approach consistent and reliable over these inputs, we structure the agentic coding process around a compact set of atomic functions capable of representing or approximating a majority of scenario descriptions in the dataset. Accordingly, each scenario description is expressed as one or more compositions of the shared atomic functions. Because related scenario descriptions share underlying functions, they fail in comparable,

Split	Method	HOTA-Temporal ↑	HOTA-Track ↑	Timestamp BA ↑	Log BA ↑
Validation	RefProg	26.27	36.18	68.07	70.46
	SM-Agent	23.25	31.15	66.95	67.66
	Argonaut	**33.57**	**48.15**	**73.75**	**77.99**
Test	RefProg	26.27	36.18	68.07	70.46
	SM-Agent	23.25	31.15	66.95	67.66
	Argonaut	**37.04**	**55.11**	**75.50**	**80.75**

Table 1. Official evaluation metrics on the 2026 Argoverse 2 Scenario Mining benchmark for the validation and test splits.

predictable ways, so one correction can fix a whole family of scenarios at once.

This set comprises 33 atomic functions in four groups: map functions (e.g., *at crosswalk*, *on intersection*), single-agent dynamics (e.g., *acceleration*, *turning*, *lane change*), multi-agent spatial relations (e.g., *on left*, *in front*, *same lane*), and multi-agent interactions (e.g., *approaching*, *overtaking*). For example, the scenario *"accelerating vehicle changing lanes to the right"* is expressed as the composition of *acceleration* and *right lane change*. Under this design, a GPT-5.5-based coding agent implements the atomic functions, then iteratively refines them on the validation split using their failure modes against the ground truth.

BEV Verification. The rule stage has two failure modes: geometric approximations can produce false positives when a match holds numerically but not semantically, and tracker noise can corrupt the numerical computations the rules rely on. Both are typically resolvable from a bird's-eye view, which enables holistic reasoning over scene layout, actor trajectories, and multi-agent interactions that geometric rules alone cannot capture. We therefore introduce a BEV verification stage built on Gemini-3.5-Flash [2] that re-examines each rule-stage positive through a short rendered video. Each clip overlays the involved actors on the map with their immediate surroundings; given the clip and the scenario description, the model returns a binary verdict with a brief justification.

Figure 1 shows a representative example: the rule stage accepts *"active vehicle at stop sign"* because the vehicle passes close to the sign, but from the BEV clip, the VLM model sees that it travels along the main road, which is not under regulation of the sign, and rejects the positive.

Camera Analysis. Certain scenarios rely on explicit visual cues that fall beyond the scope of the rule stage, such as weather and road conditions, traffic-signal state and signage, and the visual state of actors (e.g., *"ego vehicle reflected in the windows of a nearby building"*). For these, we feed the raw camera streams directly to a VLM, splitting each into five-second sliding windows with a one-second stride, sampled at 5 Hz. Since this dense video processing is computationally expensive, we opt for Qwen3.5-27B [4], a capable open-source video-language model that offers competitive performance at substantially lower cost than commercial alternatives.

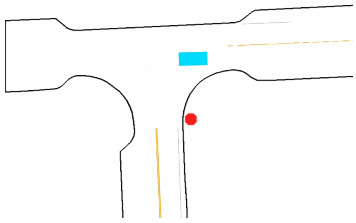

Figure 1. BEV verification frame for *"active vehicle at stop sign."* The cyan rectangle is the active vehicle and the red octagon the stop sign, rendered over the Argoverse 2 map. Gemini's verdict: *"The stop sign controls traffic from the perpendicular road. The active vehicle is not on the lane it regulates, so the scenario does not hold.* ***Answer: No.****"*

3. Results and Discussion

Table 1 reports our results on the validation and test splits, where Argonaut outperforms both 2026 official baselines across all metrics. As LLMs and VLMs continue to improve, we anticipate further progress along this direction.

References

[1] Cainan Davidson, Deva Ramanan, and Neehar Peri. Refav: Towards planning-centric scenario mining. *arXiv preprint arXiv:2505.20981*, 2025. 1

[2] Google DeepMind. Gemini 3.5 flash model card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026. Accessed: 2026-05-29. 2

[3] OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, 2026. Accessed: 2026-05-29. 1

[4] Qwen Team. Qwen3.5 technical report. *arXiv preprint arXiv:2604.15804*, 2026. 2

[5] Zhepeng Wang, Feng Chen, Kanokphan Lertniphonphan, et al. Le3de2e solution for av2 2024 unified detection, tracking, and forecasting. Technical report, CVPR Workshop on Autonomous Driving, 2024. 1

[6] Benjamin Wilson, Wei Qi, Tanmay Agarwal, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1