SM-Agent Solution for AV2 2025 Scenario Mining Challenge

Dubing Chen[1,2]*, Huan Zheng[1], Wencheng Han[1],
Runzhou Tao[2], Zhongying Qiu[2], Jianfei Yang[2], Jianbing Shen[1] †
[1]SKL-IOTSC, CIS, University of Macau
[2]Zhejiang ZEEKR Automobile Research & Development Co., Ltd.

Abstract

This report presents our solution, SM-Agent, for the Argoverse (AV) 2 Scenario Mining (SM) Challenge at the Workshop on Autonomous Driving (WAD), CVPR 2025. Our framework leverages large language models (LLMs) as code-generating agents to translate natural language queries into executable scripts for scenario mining. Our core contribution is a systematic investigation into different agentic architectures to optimize code generation. We introduce a Global Context-Aware Generation method that processes all queries simultaneously to improve consistency and a Multi-Agent Refinement system where a dedicated Refiner agent iteratively debugs and enhances the generated code. Our method achieved the 1st place in the Argoverse 2 Scenario Mining Challenge at the CVPR 2025 WAD.

1. Introduction

The ability to automatically identify specific, complex, and rare scenarios from massive-scale sensor data is critical for the development and validation of autonomous driving systems. Traditional methods for scenario mining often rely on rigid, hard-coded rules and struggle to interpret nuanced, high-level descriptions of events. To address these limitations, the Argoverse 2 [8] Scenario Mining Competition reframes SM as a natural-language-to-code translation task.

Our approach tackles this challenge by utilizing a framework where LLMs act as intelligent agents to convert human-written scenario descriptions into programmatic scripts. These scripts, composed of predefined atomic functions that operate on object trajectory data, are then executed to precisely filter and extract the desired scenarios, combining the intuitive flexibility of natural language with the rigor and precision of formal code. Specifically, we design and implement this framework by systematically exploring several agent architectures, from a single-query baseline to more advanced designs. Our contributions include a *Global Context-Aware Generation* strategy that leverages long-context reasoning across the entire query set and a Multi-Agent Refinement process, featuring a dedicated *Refiner* agent and an automated error-correction loop, to iteratively improve code quality. We conduct extensive experiments on the Argoverse 2 Scenario Mining benchmark, demonstrating that our proposed methods achieve first-place results in the competition.

2. Method

2.1. Overall Framework

Given a natural language descriptor for a specific scenario and a variety of sensor inputs (*e.g.*, LiDAR, images), the core task is to mine and extract the corresponding scenarios, which are represented as a set of object trajectories. Our framework decomposes this complex task into three primary stages:

- **Tracking and Prediction:** We follow the official competition repository to leverage the tracking prediction results from Le3DE2E [1], a state-of-the-art method for 3D detection and tracking on the Argoverse 2 dataset.
- **Code Generation:** For all given scenario descriptions, we employ an LLM-based agent to translate the natural language texts into the programmatic scripts. The script is constructed by calling a set of predefined atomic functions.
- **Code Execution and Track Filtering:** The generated script is then executed to query and filter the predicted tracking data, thereby mining the specific trajectories that match the scenario description.

The central contribution of this work is an in-depth investigation into the code generation component, which is detailed in the following sections.

2.2. Code Generation Agent

As illustrated in Fig. 1, we explored several architectural designs for our code generation agent. These include a baseline single-query generator, a global context-aware ap-

*This work was done during internship at Zeekr.

†Corresponding author: *Jianbing Shen*.

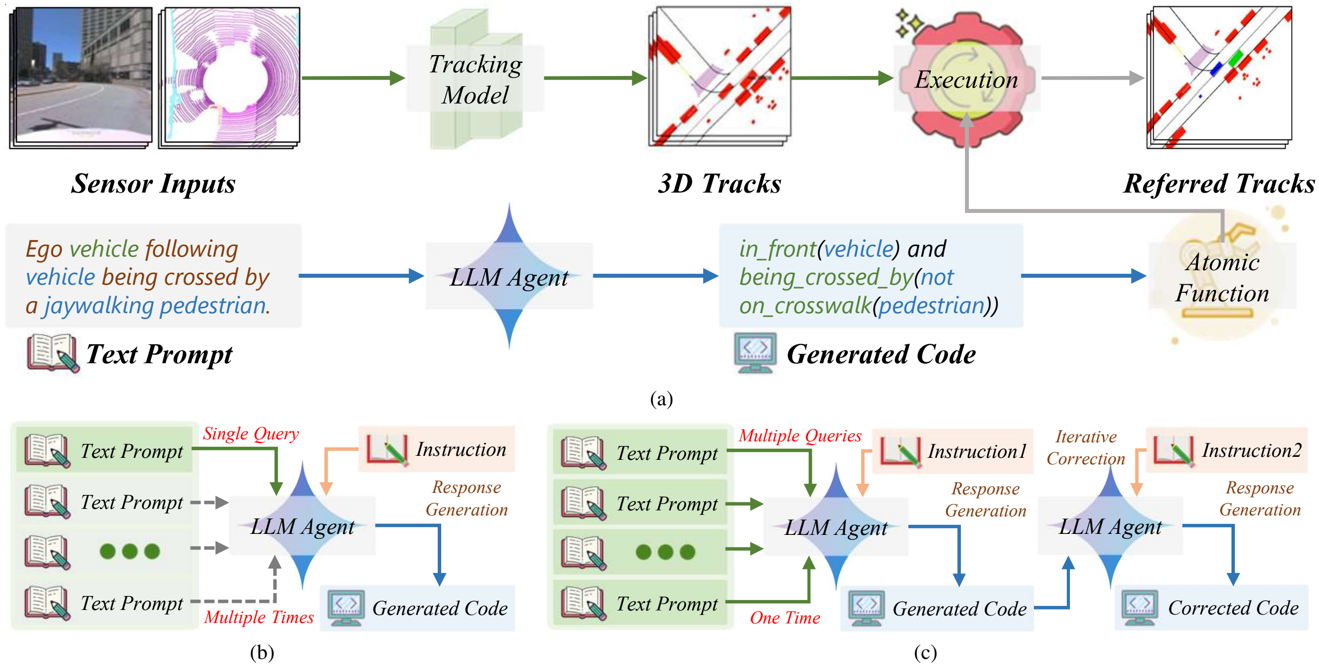

Figure 1. (a): Illustration of the framework of SM-Agent. (b): Single-query code generation. (c): Global context-aware code generation.

proach to leverage long-context reasoning, and a multi-agent system featuring a dedicated code *Refiner* and an iterative correction loop.

2.2.1. Single-Query Code Generation

As shown in Fig. 1b, our initial approach employs a standard LLM as a direct code generator. The model is prompted with a comprehensive set of predefined atomic functions, a list of all possible object categories, and a single natural language descriptor. Its task is to generate a self-contained scenario mining script corresponding to that descriptor. We experimented with various state-of-the-art LLMs, including DeepSeek [4], Gemini [6], and Claude. To enhance performance, we incorporated several few-shot examples into the prompt to facilitate in-context learning [3].

2.2.2. Global Context-Aware Code Generation

A key limitation of the single-query approach is the limited context, which curtails the model's ability for broader reasoning. To leverage the reflective capabilities [5] of modern LLMs, we developed the *Global Context-Aware Generation* method. Instead of processing descriptors individually, we bundle all descriptors from the dataset into a single long prompt. This allows the model to analyze the entire set of tasks at once, perform internal cross-validation on its own generated code, and maintain greater consistency [7], leading to more robust and accurate results.

2.2.3. Multi-Agent Refinement and Iterative Correction

To further improve code quality, we introduced a code *Refiner* within a multi-agent framework. This system consists of a primary *Generator* agent and a secondary *Refiner* agent. The *Refiner* is an LLM specifically tasked with inspecting, debugging, and improving the code produced by the Generator, fostering a collaborative process to achieve superior results. Moreover, we implemented an automated feedback loop. If the generated script fails to execute, encounters a runtime error, or exceeds a predefined execution time limit, the system automatically triggers a regeneration request to the LLM agent. This iterative process continues until a valid and executable script is successfully produced.

3. Experiments

3.1. Dataset and Evaluation Metrics

Our experiments are conducted on the Argoverse 2 Scenario Mining benchmark [2, 8]. This benchmark extends the official Argoverse 2 Sensor dataset, which comprises 1000 scenes (750 for training, 150 for validation, and 150 for testing), by incorporating 10,000 planning-centric natural language queries. The primary evaluation metric is HOTA-temporal, a comprehensive tracking metric that jointly assesses detection and association accuracy for the referred objects over time. We also report metrics including HOTA-Track, Timestamp-level Binary Accuracy (Timestamp BA), and Log-level Binary Accuracy (Log BA), which evaluate classification performance at the timestamp and full-log (*i.e.*, scenario) levels, respectively.

3.2. Ablation Study

As shown in Tab. 1, we evaluated several state-of-the-art LLMs, including DeepSeek-v3, claude-opus-4, and gemini-

Experiment	HOTA-temporal (↑)	HOTA-Track (↑)	Timestamp BA (↑)	Log BA (↑)
Baselines				
deepseek-v3-250324	40.96	43.46	70.30	65.23
claude-opus-4-20250514	48.12	47.81	73.57	**69.04**
gemini-2.5-pro-preview-06-05	50.57	50.58	73.65	68.07
Ablations on gemini-2.5-pro-preview-06-05				
+ Planner-Coder	48.25	47.79	71.69	64.34
+ Voting Ensemble	51.28	49.22	74.64	66.25
+ Voting + Refiner	52.18	50.15	74.64	65.91
Ours (gemini-2.5-pro-preview-06-05)				
+ Global Context + Refiner	**53.38**	**51.05**	**76.62**	66.34

Table 1. Main results on the Argoverse 2 Scenario Mining benchmark. Our proposed method, combining *Global Context-Aware Generation* and a *Refiner* agent on top of the gemini-2.5-pro-preview-06-05 baseline, achieves the highest performance. Higher values are better for all metrics. Best results are in **bold**.

Rank	Participant Team	HOTA-Temporal (↑)	HOTA-Track (↑)	Timestamp BA (↑)	Log BA (↑)
1	**Zeekr_UMCV (Ours)**	**53.38**	51.05	76.62	66.34
2	Mi3 UCM_AV2	52.37	51.53	77.48	65.82
3	zxh	52.09	50.24	76.12	66.52
4	LiDAR_GPT_VLM	51.92	51.91	76.90	66.74
5	Host_55079_Team	50.15	51.13	74.03	68.31
6	PKUMM	34.31	41.53	66.94	67.95

Table 2. Final leaderboard of the CVPR 2025 Argoverse 2 Scenario Mining Challenge. Our team (**Zeekr_UMCV**) achieved the top rank. The primary ranking metric is HOTA-Temporal. Higher is better for all metrics.

2.5-pro, within our framework. To validate our design choices, we conducted a series of ablation studies. We first explored a two-stage *Planner-Coder* architecture, where one LLM generates a high-level plan that a second LLM translates into code. This approach, however, led to a performance decrease compared to the direct generation baseline. We also experimented with a *Voting Ensemble*, where an additional LLM selects the best script from multiple proposals, which yielded no significant improvement over the best-performing single model.

Our proposed methods, *Global Context-Aware Generation* and the *Refiner* agent, proved most effective. The combination of these two techniques on top of the powerful gemini-2.5-pro baseline delivered the best results, achieving a HOTA-temporal score of **53.38**, a significant improvement over all other configurations. This underscores the value of providing broader context and incorporating an iterative refinement loop for complex code generation tasks.

3.3. Leaderboard Results

The final leaderboard standings are presented in Tab. 2. Our submission secured first place with a leading score of 53.38 on the primary HOTA-Temporal metric. The results highlight a highly competitive field, with the top entries achieving very close scores. This result validates the effectiveness of our agentic framework, particularly the *Global Context-Aware Generation* and Multi-Agent Refinement strategies, in producing robust and accurate scenario mining scripts.

4. Conclusion

In this report, we detailed our first-place solution for the CVPR 2025 Argoverse 2 Scenario Mining Challenge. The core of our approach is a framework that employs LMM as code-generating agents, enhanced by two key architectural innovations: a *Global Context-Aware Generation strategy* for improved reasoning and a Multi-Agent Refinement system for iterative correction. Our experiments demonstrated that this sophisticated agentic design significantly outperforms simpler baselines, underscoring the value of contextual understanding and refinement in complex, domain-specific tasks. This work validates a promising direction for building more powerful and intuitive tools for autonomous vehicle safety analysis, with future research poised to grant these agents even greater autonomy.

References

[1] Feng Chen, Kanokphan Lertniphonphan, Yaqing Meng, Ling Ding, Jun Xie, Kaer Huang, and Zhepeng Wang. Le3de2e so-

lution for av2 2024 unified detection, tracking, and forecasting challenge. 1

[2] Cainan Davidson, Deva Ramanan, and Neehar Peri. Refav: Towards planning-centric scenario mining. *arXiv preprint arXiv:2505.20981*, 2025. 2

[3] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2

[4] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

[5] Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024. 2

[6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[7] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[8] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 1, 2