Technical Report for CVPR 2025 Workshop on Argoverse2 Scenario Mining

Jin-Hee Lee, Jae-Keun Lee, Youngho Cheon, Soon Kwon
Daegu Gyeongbuk Institute of Science and Technology
{jhlee07,lejk8104,yhcheon,soonyk}@dgist.ac.kr

Abstract

Automatically identifying safety-critical and meaningful scenarios within terabyte-scale multimodal data streams is a significant challenge. RefAV addresses this problem by first feeding natural language scenario descriptions into an Large Language Model (LLM) to generate a sequence of Atomic Functions, then searching the tracking results for sub-tracks that satisfy the conditions defined by those functions. However, this approach has two key limitations. First, ambiguity or complex conditions in natural language may not be consistently converted into Atomic Functions, leading to missed or misinterpreted scenarios. Second, when visual judgment is essential, language information alone struggles to guide the correct combination of functions. In this report, we overcome these limitations of the purely LLM-based RefAV by (1) leveraging state-of-the-art LLM (GPT-o3) to improve the quality of Atomic Function generation, (2) integrating Vision-Language Model (VLM) module to directly interpret images and point clouds and thus select the appropriate Atomic Functions when visual reasoning is required, and (3) redesigning the LLM input prompts and category guidelines to be more specific and clear, encouraging the model to produce Atomic Functions reliably. Through these three enhancements, we complement natural language understanding with visual information analysis, substantially improving the accuracy of sub-track extraction for safety-critical scenarios.

1. Method

1.1. Overviews

In this challenge, we propose a framework that takes as input multiple LiDAR point-cloud frames and multi-view camera images, and predicts for each object its category (e.g., REFERRED OBJECT, RELATED OBJECT, OTHER OBJECT), 3D bounding-box parameters (position, size, heading), and track ID. As illustrated in Figure 1, the framework consists of a 3D Object Detector, a Tracker, a LLM, and an Atomic Function module. Detailed descriptions of each component follow in the subsequent sections.

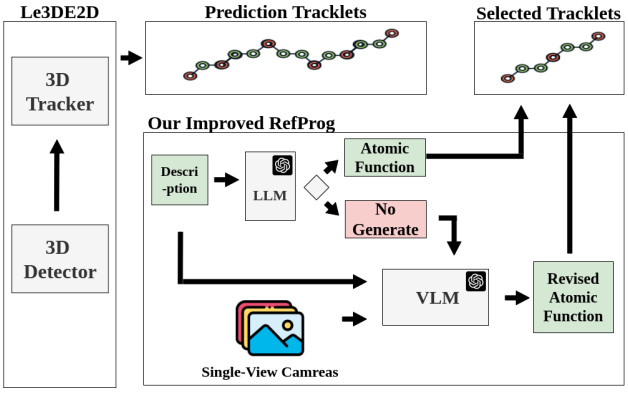

Figure 1. **Overview of the our framework for scenario mining.**

1.2. 3D Detector and Tracker

We adopt Le3DE2E [3], the winner model of the 2024 CVPR Argoverse2 tracking challenge, as our baseline. The 3D Object Detector module processes each frame of LiDAR point cloud to compute object class, confidence score, and bounding box parameters (center x, center y, center z, width, length, height, yaw). These detections are then passed to the AB3DMOT tracker [4], which performs association across temporal frames to assign consistent track IDs and statements.

1.3. LLM and RefProg

Following RefAV [1], we feed natural language scenario descriptions into an LLM to generate combinations of predefined Python-based Atomic Functions for each scenario. The resulting code blocks are applied to the tracker outputs to extract the subset of tracks most similar to the described scenario. To improve scenario mining accuracy, we introduce three core enhancements:

Leveraging state-of-the-art LLM: Scenario mining accuracy heavily depends on the LLM's ability to understand and reason about Atomic Functions [1]. Therefore, we employ state-of-the-art models provided by OpenAI, such as GPT-o3, as our baselines for generating the most appropri-

ate atomic functions from natural language descriptions.

Integrating VLM module: In scenarios requiring visual judgment, the LLM alone may fail to generate suitable Atomic Functions. To address this, we incorporate a VLM module that processes camera images to perform necessary visual reasoning. The outputs from the VLM guide the selection and combination of Atomic Functions, compensating for the LLM's limitations.

Redesign of prompt and guideline We redesign the LLM input prompts and category guidelines to be more concrete and precise, steering the model toward consistent and reliable Atomic Function generation.

These complementary improvements significantly boost the accuracy of sub-track extraction corresponding to input scenarios.

2. Dataset and Metrics

2.1. Dataset

We use the RefAV dataset [1], which extends the Argoverse 2 Sensor Dataset. It comprises 1,000 sequences: 700 for training, 150 for validation, and 150 for testing. Each sequence includes LiDAR point clouds collected at 10 Hz, precise 3D annotations, synchronized multi-view camera images, HD map data, and natural language queries describing the scene.

2.2. Metric

The RefAV dataset defines 26 object classes, categorized as REFERRED OBJECT (directly mentioned by the prompt), RELATED OBJECT (interacting with the referred object), and OTHER OBJECT (unrelated to the prompt). Performance is measured using two variants of the HOTA metric [2], extended to the temporal and tracking domains: HOTA-Temporal and HOTA-Track. Both metrics evaluate only the REFERRED OBJECT class. Final challenge rankings are determined by the HOTA-Temporal score.

3. Experiment

Participant Team	HOTA-Temp.	HOTA-Track
Zeekr_UMCV	**53.38**	51.05
Mi3 UCM_AV2	52.37	51.53
zxh	52.09	50.24
LiDAR_GPT_VLM (Ours)	51.92	**51.91**
Baseline (RefProg)	50.15	51.13
PKUMM	34.31	41.53

Table 1. Results on the test dataset for the 2025 Argoverse2 Scenario Mining Competition Leaderboard.

Table 1 shows the evaluation results of our proposed model on the Argoverse2 test set. Our model achieves a HOTA-Temporal score of 51.92 (4th place) and a HOTA-Track score of 51.91 (1st place), securing 4th overall in the 2025 Argoverse2 Scenario Mining challenge.

References

[1] Cainan Davidson, Deva Ramanan, and Neehar Peri. Refav: Towards planning-centric scenario mining. *arXiv preprint arXiv:2505.20981*, 2025. 1, 2

[2] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 2

[3] Zhepeng Wang, Feng Chen, Kanokphan Lertniphonphan, Siwei Chen, Jinyao Bao, Pengfei Zheng, Jinbao Zhang, Kaer Huang, and Tao Zhang. Technical report for argoverse challenges on unified sensor-based detection, tracking, and forecasting. *arXiv preprint arXiv:2311.15615*, 2023. 1

[4] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.08063*, 2020. 1