RefAV: Towards Planning-Centric Scenario Mining

Cainan Davidson, Deva Ramanan, Neehar Peri
Carnegie Mellon University

Abstract

Autonomous Vehicles (AVs) collect and pseudo-label terabytes of multi-modal data localized to HD maps during normal fleet testing. However, identifying interesting and safety-critical scenarios from uncurated driving logs remains a significant challenge. Traditional scenario mining techniques are error-prone and prohibitively time-consuming, often relying on hand-crafted structured queries. In this work, we revisit spatio-temporal scenario mining through the lens of recent vision-language models (VLMs) to detect whether a described scenario occurs in a driving log and, if so, precisely localize it in both time and space. To address this problem, we introduce RefAV, a large-scale dataset of $10,000$ *diverse natural language queries that describe complex multi-agent interactions relevant to motion planning derived from* 1000 *driving logs in the Argoverse 2 Sensor dataset. We evaluate several referential multi-object trackers and present an empirical analysis of our baselines. Notably, we find that naively repurposing off-the-shelf VLMs yields poor performance, suggesting that scenario mining presents unique challenges. Lastly, we discuss our recently held competition and share insights from the community. Our code and dataset are available on Github and Argoverse.*

1. Introduction

Autonomous Vehicle (AV) deployment on public roads has increased significantly in recent years, with Waymo completing more than a million rides per week as of February 2025 [25]. Despite the maturity of such autonomous ride-hailing services, AVs can still sometimes require manual interventions [26, 44] and can be prone to accidents [62]. Although data-driven simulators are an integral component in establishing a safety case [18, 21, 66, 75], validating end-to-end autonomy with real-world operational data remains a critical part of the testing stack. However, identifying interesting and safety-critical scenarios from uncurated real-world driving logs is akin to finding a "needle in a haystack" due to the scale of data collected during normal fleet operations. In this paper, we revisit the task of spatio-temporal scenario mining with recent vision-language models (VLMs) to identify interesting multi-agent interactions using natural language (cf. Fig. 1).

Status Quo. Although language-based 3D scene understanding has been extensively studied in the context of referential multi-object tracking (RMOT) [29, 71, 72], multi-modal visual question answering (VQA) [20, 51, 63], and VLM-based motion planning [42, 56, 58, 68], we argue that spatiotemporal scenario mining presents unique challenges. RMOT extends referential grounding by associating referred objects over time. However, unlike RMOT, scenario mining does not guarantee that referred objects exist in a given driving log. Next, multi-modal VQA extends VQA with additional visual modalities like LiDAR. Although spatio-temporal scenario mining also leverages multi-modal input (e.g. RGB, LiDAR and HD maps), we require that methods output 3D tracks instead of text-based answers. Lastly, VLM-based motion planners directly estimate future ego-vehicle waypoints based on high-level language instructions and past sensor measurements. In contrast, scenario mining methods can reason over the full driving log to identify interactions between non-ego vehicles. Concretely, spatio-temporal scenario mining requires identifying if a described scenario occurs in a driving log from raw sensor measurements (e.g. RGB and LiDAR), and if so, precisely localizing *all* referred objects in both time and space with 3D tracks. To support this task, we propose RefAV, a large-scale dataset of $10,000$ diverse natural language queries designed to evaluate a model's ability to find a visual "needle in a haystack".

RefAV Scenario Mining Dataset. We repurpose the Argoverse 2 (AV2) Sensor dataset [70], which contains $1,000$ driving logs with synchronized LiDAR, 360° ring cameras, HD maps, and 3D track annotations for 30 categories. We curate a set of $10,000$ natural language prompts that describe interesting (and often rare) scenarios (cf. Fig. 2 and Fig. 3) using a combination of manual annotations and procedural generation using large language models (LLMs). Notably, AV2 annotates ground truth tracks at 10 Hz, which allows for fine-grained motion understanding. In contrast, prior referential multi-object tracking benchmarks (the established task most similar to spatio-temporal scenario mining) use nuScenes [4], which annotates ground truth tracks at 2 Hz (cf. Table 1). Argoverse's higher temporal resolution uniquely

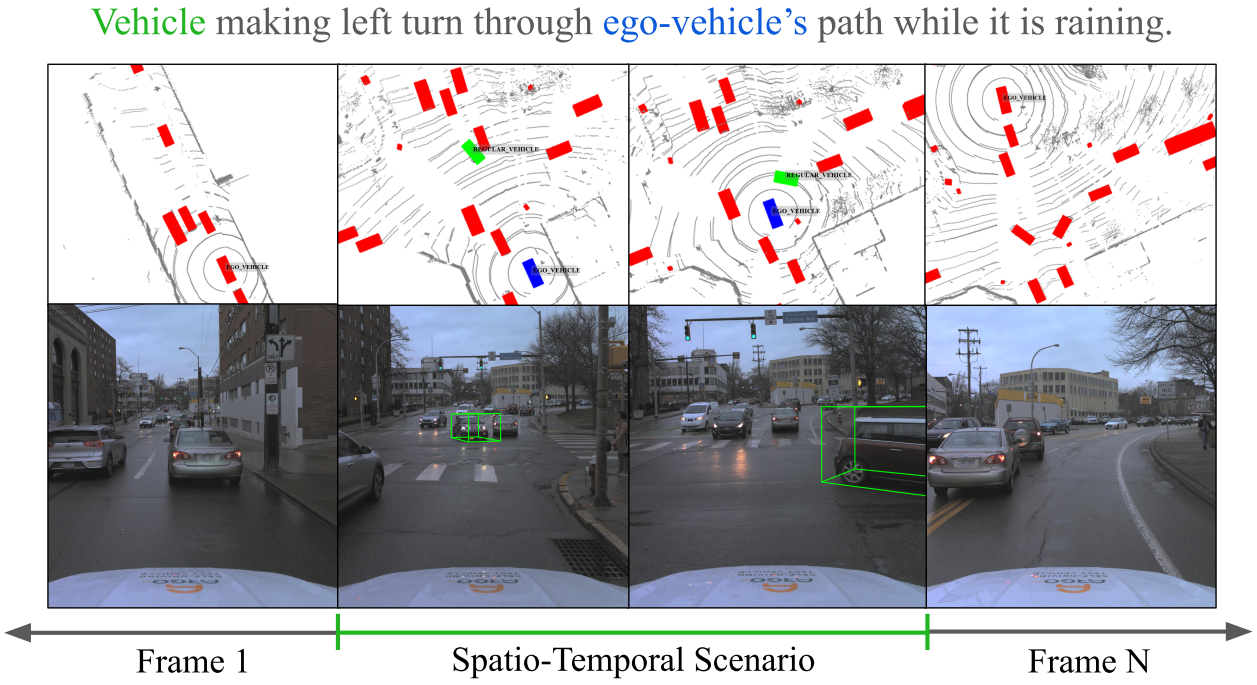


Figure 1. **Scenario Mining Problem Setup.** Given a natural language prompt such as `vehicle making left turn through ego-vehicle's path while it is raining`, our problem setup requires models to determine whether the described scenario occurs within a 20-second driving log, and if so, precisely localize the referred object in 3D space and time from raw sensor data (LiDAR, 360° ring cameras, and HD maps). Based on the example above, a VLM should localize the start and end timestamps and 3D location of the red Mini Cooper executing a "Pittsburgh left" through the ego-vehicle's path with a 3D track. Notably, the "Pittsburgh left" is a regional driving practice where a driver quickly makes a left turn before oncoming traffic proceeds. Although common in Pittsburgh, this maneuver is technically illegal. Therefore, we argue that scenario mining is critical for validating end-to-end autonomy in order to build a comprehensive safety case. Note that `referred objects` are shown in green, `related objects` in blue, and `other objects` in red.

allows RefAV to evaluate dynamic multi-agent interactions (unlike prior work that primarily evaluates referential expressions based on static attributes like vehicle color or relative heading). Interestingly, we find that simply repurposing VLMs for scenario mining yields poor performance.

Referential Tracking by Program Synthesis. Although prior referential trackers and VLMs achieve reasonable performance with simple referential prompts (e.g. `find all cars`), we find that such methods struggle with compositional reasoning and motion understanding (e.g. `find all cars accelerating while changing lanes`). To address this problem, we propose Referential Tracking by Program Synthesis (RefProg), a modular approach that combines off-the-shelf 3D tracks with LLMs. Inspired by recent work in program synthesis [19, 43, 60], our method uses an LLM to break down complex referential expressions into simpler compositional actions. Specifically, we define an API to describe hand-crafted atomic functions (e.g., `turning`, `accelerating`, `changing lanes`) and use an LLM to synthesize a program that composes these atomic actions corresponding to the natural language prompt. We then execute the generated program to filter off-the-shelf tracks to identify the subset of tracks that best match the described scenario (cf. Fig. 4).

Contributions. We present three major contributions. First, we introduce RefAV, a large-scale dataset designed to evaluate VLMs on 3D scene understanding and spatio-temporal localization. Our extensive experiments highlight the limitations of current methods, and demonstrate the effectiveness of our proposed program synthesis-based approach. Lastly, we highlight the results of our recent CVPR 2025 challenge hosted in conjunction with the Workshop on Autonomous Driving.

2. Related Works

Referential Grounding and Tracking are long-standing challenges in vision-language understanding. Early datasets like RefCOCO, RefCOCO+, and RefCOCOg [22, 41, 77] helped popularize the task, which was later adapted for au-

tonomous driving [12, 65]. Although prior work focused on single-frame visual grounding, ReferKITTI [71] formalizes the problem of 2D referential multi-object tracking (RMOT), which extends referential object detection by associating referred objects over time. ReferKITTI-v2 [78] expands ReferKITTI's manual annotations using large language models (LLMs) to improve prompt diversity. Further, LaMOT [33] consolidates 1,660 sequences from four tracking datasets to create a large-scale unified RMOT benchmark.

Referential tracking methods can be broadly classified into two-stage (e.g. separately track and filter predictions based on natural language cues) and end-to-end approaches. iKUN [15] proposes a plug-and-play knowledge unification module to facilitate language grounding for off-the-shelf trackers. Similarly, ReferGPT [6] presents a zero-shot method for filtering off-the-shelf 3D tracks using CLIP [52] scores and fuzzy matching. In contrast, JointNLT [80] proposes a unified visual grounding and tracking framework, and MMTrack [79] reformulates tracking as token generation. OVLM [48] introduces a memory-aware model for RMOT, while ReferFormer [73] leverages language-conditioned object queries. More recently, nuPrompt [72] extends 2D referential tracking to 3D with RGB input. Different from nuPrompt, RefAV addresses 3D *multi-modal* referential tracking and dynamic scene understanding.

Visual Question Answering with VLMs has improved significantly in recent years due to large-scale multi-modal pre-training. Vision-language models (VLMs) like LLaVA [34, 35], BLIP-2 [30], and Qwen2.5-VL [1] show strong generalization across diverse domains [17, 51, 63]. However, state-of-the-art models still struggle with spatial reasoning and grounding [40, 54]. To address this issue, SpatialRGPT [10] and SpatialVLM [7] generate large-scale monocular depth pseudo-labels to improve 3D reasoning. Despite the effectiveness of such methods, zero-shot prompting yields poor compositional reasoning performance. Instead, VisProg [19] and ViperGPT [60] use LLMs to generate executable code and use tools like OwlViT [45], CLIP [52], MiDaS [53], and GLIP [31] for spatial understanding. However, existing program synthesis approaches primarily focus on single-frame reasoning. In contrast, we are interested in understanding dynamic multi-agent interactions from video sequences. We take inspiration from existing methods and apply similar program synthesis-based approaches for spatio-temporal scenario mining.

Vision-Language Models for 3D Understanding have been extensively explored in the context of representation learning, open-vocabulary 3D perception, and end-to-end driving. SLidR [55] distills 2D features into 3D point clouds for cross-modal learning. SEAL [36] incorporates SAM [24] to produce class-agnostic 3D segments, while SA3D [5] leverages SAM and NeRFs for object segmentation. Similarly, recent work like Anything-3D [57] and 3D-Box-Segment-Anything integrate VLMs and 3D detectors (e.g., VoxelNeXt [9]) for interactive reconstruction and labeling. VLMs have been used extensively in open-vocabulary perception for autonomous driving. UP-VL [46] distills CLIP features into LiDAR data to generate amodal cuboids. Recent work uses 2D VLMs to generate 3D pseudo-labels for open-vocabulary perception, enabling zero-shot LiDAR panoptic segmentation [47, 61] and 3D object detection [23, 39, 49]. While traditional grounding methods [72] struggle with complex instructions, multi-modal LLMs [11, 42, 56] demonstrate impressive visual understanding. However, such methods [3, 14, 68] typically produce scene-level analysis for end-to-end driving rather than precise instance localization. In contrast, our framework combines language grounding with precise geometric localization from offline 3D perception methods for more accurate vision-language reasoning.

3. RefAV: Scenario Mining with Natural Language Descriptions

In this section, we present our approach for curating $10,000$ natural language prompts (Sec 3.1) and describe five zero-shot scenario mining baselines (Sec 3.2).

3.1. Creating RefAV

Unlike prior benchmarks that primarily evaluate referential expressions based on static attributes like vehicle color or relative heading (e.g. `find the red car to the left`), we focus on mining interesting and safety-critical scenarios relevant for motion planning (cf. Fig. 1). We take inspiration from recent planning benchmarks like nuPlan [21] to identify such planning-centric scenarios. In particular, nuPlan introduces a set of 80 scenarios considered relevant for safe motion planning. We use this as a template to identify similar scenarios within AV2.

Why Use the Argoverse 2 (AV2) Sensor Dataset? The AV2 sensor dataset contains 1000 15 – 20 second logs with synchronized sensor measurements from seven ring cameras and two LiDAR sensors. Moreover, the dataset includes HD maps with lane markings, crosswalk polygons, and various lane types (e.g. vehicle, bus, and bike lanes). Notably, AV2 annotates 30 object categories with track-level annotations at 10 Hz. We choose to build RefAV on top of AV2 because prior datasets like nuScenes [4] only annotate ground-truth tracks at a lower temporal resolution (e.g. 2Hz vs. 10Hz), making it more difficult to extract fine-grained motion. We ablate this in the Appendix F. In addition, KITTI [16] and Waymo Open Dataset [59] only label a limited number of categories (e.g. `car`, `pedestrian`, `bicycle`) and do not have HD maps, making it difficult to evaluate diverse multi-agent interactions. Although AV2 annotates objects up to 150m away from the ego vehicle, we clip all object tracks

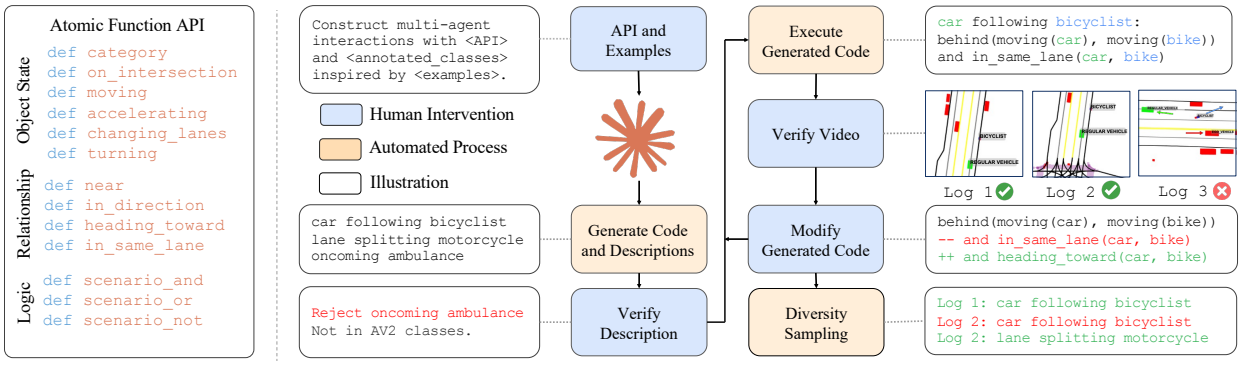


Figure 2. **RefAV Dataset Creation.** First, we define a set of 28 atomic functions that identify the state of an object track, its relationship with other objects (stored in an underlying scene graph), and a set of boolean logical operators to support function composition. Next, we prompt an LLM to permute these atomic functions and generate a program and corresponding natural language description. Finally, we execute the generated code on ground-truth tracks and visualize the referred object track to manually verify that the program output matches the natural language prompt. Code that generates an incorrect video is modified by an annotator and re-executed. We sample valid programs to maximize scenario diversity in our dataset.

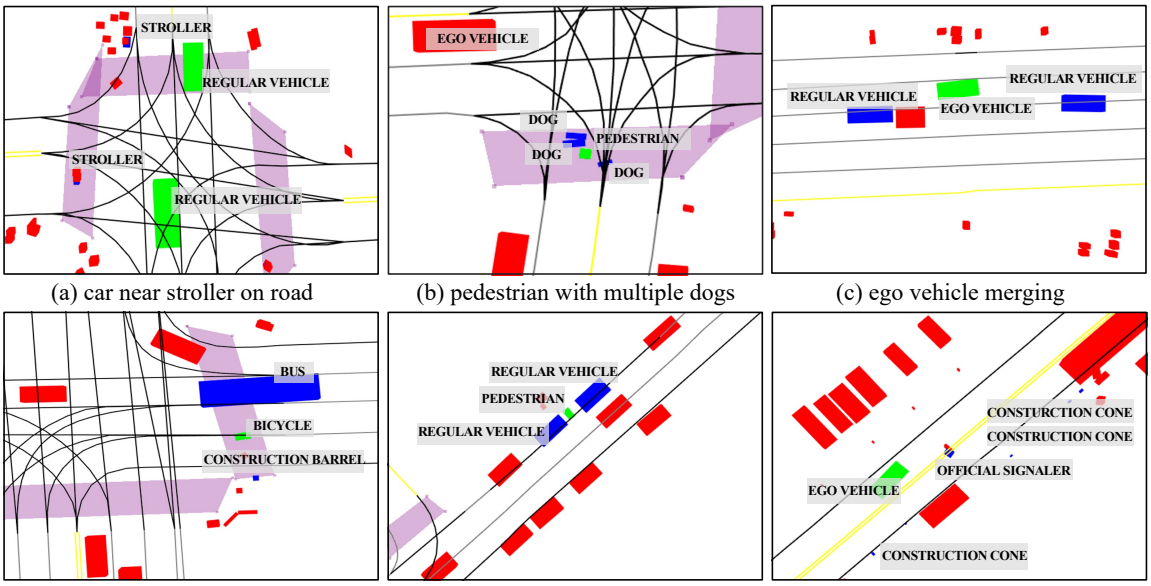
at 50m. We find that current 3D perception models struggle with long-range detection and tracking [50], suggesting that the community is not ready to address spatio-temporal scenario mining at range.

LLM-based Procedural Scenario Generation. Our key insight is that many complex scenarios (e.g. `find all cars accelerating while changing lanes`) can be broken down into simpler atomic actions (e.g. `find cars`, `accelerating`, and `changing lanes`), and a large set of such atomic actions can be composed to generate new diverse scenarios. To this end, we first define 28 atomic functions based on nuPlan's list of planning scenarios (cf. Figure 2 left). We include a full API listing in the Appendix J. To generate a new scenario definition, we provide an LLM (in our case, Claude 3.7 Sonnet) with the full API listing, along with in-context examples of real compositions. We prompt the LLM to generate permutations of the atomic functions and describe the generated code with a natural language description (cf. Figure 2 right). We execute the generated code to filter ground-truth tracks from all 1000 logs to identify true positive matches. We aggregate all true positive log-prompt matches (>50K) and sample 8,000 true positive log-prompt pairs that maximize scenario diversity. Lastly, we randomly sample 2,000 true negative log-prompt pairs. Our automatic scenario generation pipeline has an average success rate of 70%.

Verifying Procedurally Generated Scenarios. Although LLM-based procedural generation allows us to generate diverse scenarios at scale, this process is not perfect and requires extensive manual validation (cf. Figure 2 right). First, we verify that the generated descriptions match the generated code. We then manually review video clips of true positive log-prompt matches to ensure that the generated natural language description accurately describes the localized tracks. Notably, we find two common error modes. First, the LLM often defines the prototypical case of a scenario, but misses or incorrectly includes edge cases due to under or overspecification. For example, given a description `car following a bicyclist`, the generated code identifies instances where a moving car is behind a moving bike while traveling in the same road lane. This definition includes the false positive edge case where the car and bicyclist are traveling in opposite directions. Second, we find that LLMs often reverse the relationship of `referred` and `related` objects. For example, given the description `bicycles in front of the car`, the LLM generates code that corresponds to `car behind the bicycles`. Based on the videos, human annotators make the necessary changes to the scenario definition, and the modified code is re-executed across all logs. It takes about 3 minutes to verify and edit each scenario definition. In total, we spent 200 hours verifying scenario definitions.

Manual Scenario Annotation. Despite the versatility of our procedural generation approach, many interesting interactions cannot be easily identified by composing atomic functions. Therefore, we manually inspect all videos in the validation and test sets to identify interesting driving behaviors (cf. Fig 3). Further, we manually annotate weather (e.g. `clear`, `cloudy`, `rain`, `snow`, and `fog`) and lighting conditions (e.g. `daylight`, `dusk/dawn`, and `night`) in the validation and test sets as these attributes are relevant for establishing a safety case [27] in all driving conditions. See the Appendix D for more details about our annotation workflow.

Dataset Statistics. As shown in Table 1, our dataset uniquely addresses the task of spatio-temporal scenario mining, whereas other datasets focus on object detection, referring multi-object tracking (RMOT), or visual question answering (VQA). Further, RefAV is uniquely built on Argo-



(a) car near stroller on road (b) pedestrian with multiple dogs (c) ego vehicle merging

(d) bicycle passing between bus and construction barrier (e) pedestrian walking between cars (f) ego vehicle approaching construction with traffic controller

Figure 3. **Examples of Multi-Agent Interactions.** We visualize representative examples from RefAV to highlight the diversity of our dataset. In (a), we capture the interactions between vulnerable road users and vehicles at a crowded intersection. Scenario (b) presents an atypical instance of a common multi-agent interaction (e.g. `pedestrian walking a dog`). In (c), we show a complex ego-vehicle trajectory that involves multiple moving vehicles. Scenario (d) illustrates an example of a rare multi-object interaction. In (e), we highlight a scenario that might require evasive maneuvers from the ego-vehicle (e.g. the occluded pedestrian might cross the path of the ego-vehicle). Finally, subfigure (f) visualizes a scenario with a multiple-step relationship (e.g. the official signaler is standing inside of a construction zone). Note that we show `referred objects` in green, `related objects` in blue, and all `other objects` in red.

verse 2 [70], whereas most existing datasets rely on nuScenes [4] or KITTI [16]. Next, RefAV provides track-level annotations at a higher frequency of 10Hz, compared to the typical 2Hz in other datasets, allowing for more fine-grained temporal analysis. While datasets like nuGrounding [29] and nuScenes-QA [51] include large numbers of expressions, RefAV emphasizes a diversity of annotation types, including referring expressions that capture dynamic multi-agent interactions, weather, and lighting conditions. Lastly, RefAV is one of only two datasets to support negative prompts.

3.2. Scenario Mining Baselines

We present five referential tracking baselines for scenario mining, including filtering by referred class, ReferGPT [6], image embedding similarity, LLM APIs as a black box, and referential tracking by program synthesis. We repurpose 3D trackers from the AV2 End-to-End Forecasting Challenge

Filtering by Referred Class. We propose a simple "blind" baseline that does not explicitly reason about multi-agent interactions. For a given natural language prompt, we use an LLM to parse the `referred object` class and only keep 3D tracks from this class. For example, for the natural language query `find all cars turning left`, we remove all predicted tracks except `cars`.

ReferGPT. ReferGPT [6] classifies the output of off-the-shelf 3D trackers into `referred objects` and `other objects` using CLIP text embeddings. First, ReferGPT projects the predicted 3D bounding box of each track onto the 2D image plane at each timestamp to get an image crop of each object. Next, ReferGPT parses the object's coordinates, velocity, yaw, yaw rate, and distance from the ego-vehicle.

Table 1. **Comparison to Other Benchmarks.** Language-based 3D scene understanding has been extensively studied in the context of referential multi-object tracking (RMOT) and multi-modal visual question answering (VQA). Different from prior work, we address the problem of spatio-temporal scenario mining. Specifically, RefAV is based on Argoverse 2, which provides 3D track-level annotations for 30 categories at 10 Hz. Although RefAV does not include as many referential expressions as prior work (e.g. OmniDrive [68] and nuGrounding [29]), our referential annotations focus on capturing diverse multi-agent interactions. Lastly, RefAV includes negative prompts, which allows us to more accurately measure scenario mining performance.

Dataset	Base Data	Task	View	Anno. Freq.	# Expressions	# Frames	Neg. Prompts	Human Anno.
Talk2Car [12]	nuScenes	Detection	Front	Per Video	12k	9k	✗	Referring Expression
nuScenes-QA [51]	nuScenes	VQA	360°	2Hz	460k	34k	✓	None
DriveLM [58]	nuScenes	VQA	360°	2Hz	443k	5k	✗	QA Pairs
OmniDrive [68]	nuScenes	VQA	360°	2Hz	200k	34k	✗	None
Refer-KITTI [71]	KITTI	RMOT	Front	2Hz	818	7k	✗	Referring Expression
nuGrounding [29]	nuScenes	RMOT	360°	2Hz	2.2M	34k	✗	Object Attribute
nuPrompt [72]	nuScenes	RMOT	360°	2Hz	407k	34k	✗	Object Color
RefAV	Argoverse 2	RMOT & Scenario Mining	360°	10Hz	10k	155k	✓	Referring Expression, Weather, Lighting

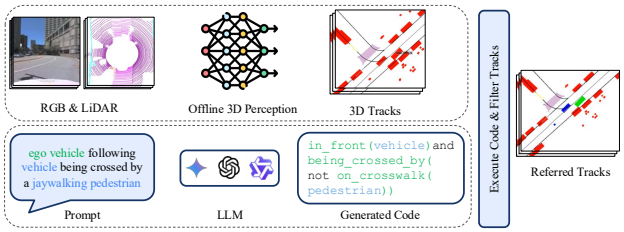


Figure 4. **Method Overview.** RefProg is a dual-path method that independently generates 3D perception outputs and Python-based programs for referential grounding. Given raw LiDAR and RGB inputs, RefProg runs a offline 3D perception model to generate high quality 3D tracks. In parallel, it prompts an LLM to generate code to identify the referred track. Finally, the generated code is executed to filter the output of the offline 3D perception model to produce a final set of `referred objects`, `related objects`, and `other objects`.

Using both the object image crop and bounding box descriptor, ReferGPT uses a VLM to generate a descriptive caption. Finally, ReferGPT uses the text cosine-similarity between the generated caption and referential prompt to score the relevance of the object track. We use GPT-5-mini as our captioning VLM and SigLIP2 [64] to compute text embeddings. We make several modifications to ReferGPT to adapt it for the spatio-temporal scenario mining problem. Since an object may be visible in multiple cameras at a single timestamp, we only caption the 2D bounding box with the largest area at each timestamp. ReferGPT selects `referred objects` by clustering tracks with the highest similarity scores. Since RefAV contains also many negative prompts, we consider a track to be a `referred object` if the similarity score between the caption text embeddings and prompt text embedding are above a fixed threshold for more than 50% of the predicted track length.

Image-Embedding Similarity. We take inspiration from ReferGPT to classify the output of off-the-shelf 3D trackers into `referred objects` and `other objects` by directly using CLIP [52] image features. We extract the track image crops in the same way as ReferGPT and compute CLIP image features for each track at each timestep. Next, we compute the cosine similarity between the per-timestep CLIP image embeddings and the CLIP text embedding of the referential prompt. We consider a track to be a `referred object` if the similarity score between the image and text embeddings are above a fixed threshold for more than 50% of the track length.

LLM API as a Black Box. We evaluate GPT-5 as a black-box scenario mining algorithm. For each log in AV2, we ask the OpenAI Responses API to identify which predictions match the referential prompt, and provide map data, city to ego-vehicle transforms, and 3D track information from an off-the-shelf tracker. GPT-5 responds with a CSV listing all `referred objects` from the initial set of predictions. Interestingly, analyzing the thinking traces of GPT-5 shows that it writes code from scratch to parse the input files and reason about multi-agent interactions.

Referential Tracking by Program Synthesis (RefProg). RefProg takes a similar approach as LLM API as a Black Box, but provides more code scaffolding to improve accuracy. Specifically, we prompt an LLM to generate code using an API listing of atomic actions for a given referential query. We execute the generated code on off-the-shelf 3D tracks to identify `referred objects`, `related objects` and `other objects` (cf. Fig 4). In addition to using API listings for estimating an object's state and relationships from 3D tracks, RefProg is given access to visual tools such as SigLIPv2 to identify visual attributes of tracked objects. Despite the similarity of RefProg and our LLM-based procedural scenario generator, we highlight two key differences. First, RefProg generates code based on the referential query, while our LLM-based procedural scenario generator synthesizes the referential query based on the generated code. We posit that the role of the scenario generator is considerably easier since there are many valid natural language prompts to describe a generated program, but significantly fewer valid programs corresponding to each natural language prompt. Lastly, code from our scenario generator is extensively modified by human annotators for correctness, while code from RefProg is executed without verification. Notably, we also evaluate RefProg on nuPrompt to demonstrate that the atomic functions generalize beyond RefAV.

4. Experiments

In this section, we briefly describe our evaluation metrics, provide an empirical analysis of our baselines, ablate the impact of different LLMs on RefProg's spatio-temporal scenario mining accuracy, and evaluate RefProg on nuPrompt to demonstrate the robustness of our approach.

Metrics. We evaluate scenario mining methods using HOTA-Temporal and HOTA-Track (variants of HOTA [38]) to measure referential tracking performance and balanced accuracy to measure information retrieval accuracy. HOTA is a unified metric that explicitly balances detection accuracy, association, and localization, making it better aligned with human visual evaluations of tracking performance. HOTA-Temporal extends the standard HOTA metric by only considering the referred timestamps within a track as true positives. HOTA-Track is similar to HOTA-Temporal, but it does not penalize methods for incorrectly predicting the start and end of a referred action. For example, given the prompt `car turning right`, HOTA-Temporal only considers the timestamps where the car is turning right in the full referred track as true positives, whereas HOTA-Track considers all timestamps in the referred track as true positives. We

Table 2. **Experimental Results.** We evaluate several zero-shot referential tracking baselines. We find that RefProg significantly outperforms all other zero-shot baselines. Notably, filtering by referred class is a particularly strong baseline, outperforming image-embedding similarity. Interestingly, directly using LLM APIs as a black box outperforms ReferGPT's hand-crafted approach. All winning submissions to our challenge build upon RefProg.

Method	HOTA ↑	HOTA-Temporal ↑	HOTA-Track ↑	Log Bal. Acc. ↑	Timestamp Bal. Acc. ↑
Ground Truth	100.0	13.3	20.5	50.0	50.0
Filter by Referred Class					
Ground Truth	100.0	21.4	33.9	52.9	57.2
LE3DE2E [8]	74.4	19.2	30.0	53.4	55.0
Image-Embedding Similarity					
Ground Truth	100.0	24.6	28.3	54.7	57.0
Le3DE2E [8]	74.4	17.2	24.4	51.1	51.1
ReVoxelDet [28]	63.1	17.1	21.2	52.3	54.4
TransFusion [2]	63.6	16.0	18.2	52.8	55.3
ReferGPT [6]					
Le3DE2E [8]	74.4	20.2	30.8	57.0	57.1
LLM APIs as a Black Box					
Le3DE2E [8]	74.4	37.2	39.2	58.4	62.3
Referential Tracking by Program Synthesis					
Ground Truth	100.0	64.8	68.7	81.1	80.7
Le3DE2E [8]	74.4	50.1	51.1	**71.8**	74.6
ReVoxelDet [28]	63.1	43.9	45.1	70.7	69.8
TransFusion [2]	63.6	48.1	46.6	70.7	68.7
Valeo4Cast [74]	69.2	41.7	44.1	62.7	69.1
Challenge Submissions					
Zeekr UMCV	74.4	**53.4**	51.0	66.3	76.6
Mi3 UCM	74.4	52.4	**51.5**	65.8	**77.5**
ZXH	74.4	52.1	50.2	66.5	76.1

also include HOTA results to contextualize standard tracking performance with the referential tracking performance of all benchmarked trackers. Additionally, we use timestamp balanced accuracy as a timestamp-level classification metric and log balanced accuracy as a log-level classification metric to evaluate how well methods can identify which timestamps and logs contain objects that match the referential prompt, respectively. Balanced accuracy is used over binary classification metrics such as F1 due to the imbalance of positive and negative prompts.

For each prompt, we categorize all objects into three groups: `referred objects`, which are the primary objects specified by the prompt; `related objects`, which are objects that interact with the referred object; and `other objects`, which are neither referred to nor relevant to the prompt. We compute both HOTA-Temporal and HOTA-Track metrics exclusively for the `referred object` class in the main paper. We report the performance on `related objects` and `other objects` in the Appendix E.

4.1. Empirical Analysis of Results

Table 2 presents several simple baselines for addressing spatio-temporal scenario mining. We evaluate each of the locally run zero-shot baselines using oracle ground-truth track annotations to understand how tracking accuracy and language grounding independently influence final performance. We do not evaluate *ReferGPT* and *LLM APIs as a Black Box* with the ground-truth track annotations to avoid dataset leakage. Evaluating every track as `referred object` without considering language prompts (first row) yields a HOTA-Temporal of 13.3% and HOTA-Track of 20.5%. Predicting the nearest object class that corresponds to the prompt with *Filter by Referred Class* improves referential tracking accuracy by 8%. However, *Filter by Referred Class* with LE3DE2E tracks results in a slight performance drop over ground truth tracks. This performance drop is perhaps lower than expected because Le3DE2E has exceptionally high tracking accuracy for `REGULAR VEHICLE`, which dominates the AV2 dataset. The *Image-Embedding Similarity* oracle outperforms the *Filter by Referred Class* oracle, suggesting that CLIP-based filtering can capture richer semantics than class names alone. However, *Image-Embedding Similarity* underperforms *Filter by Referred Class* when using predicted tracks (e.g., LE3DE2E scores 17.2 vs. 19.2 HOTA-Temporal). This likely stems from CLIP being much more sensitive to the predicted 3D location of the bounding box. *ReferGPT* slightly outperforms *Filter by Referred Class* and *Image-Embedding Similarity* across all reported metrics. *ReferGPT*, unlike the other methods, uses information about

Table 3. **Impact of Different LLMs on Code Synthesis Quality.** We evaluate the impact of using different LLMs for program synthesis in the RefProg pipeline. Interestingly, we find that Claude 3.5 Sonnet has the lowest program failure rate (e.g. 99.5% of Claude 3.5 Sonnet's generated programs were valid, compared to 81.9% for Qwen 2.5B Instruct), while Claude 3.7 Sonnet achieves the highest HOTA-Temporal.

LLM	Failure Rate ↓	HOTA-Temporal ↑	HOTA-Track ↑	Log Bal. Acc. ↑	Timestamp Bal. Acc. ↑
Qwen-2.5-7B-Instruct	18.1	31.6	34.4	62.1	62.0
gemini-2.0-flash	2.6	45.2	46.6	**72.1**	**74.6**
gemini-2.5-flash-preview-04-17	15.4	47.8	47.6	71.0	73.8
claude-3.5-sonnet-20241022	**0.5**	46.1	47.5	71.8	71.8
claude-3.7-sonnet-20250219	2.9	**50.1**	**51.1**	71.8	**74.6**

Table 4. **Zero-Shot Evaluation on nuPrompt.** We evaluate RefProg on nuPrompt and achieve state-of-the-art accuracy, highlighting the strong generalization of our approach. Importantly, we do not modify RefProg's atomic action definitions.

Method	Decoder	AMOTA ↑	AMOTP ↓	Recall ↑	MOTA ↑
CenterPoint Tracker [76]	PETR	0.178	1.650	0.291	0.197
DQTrack [32]	DETR3D	0.186	1.641	0.307	0.208
DQTrack [32]	Stereo	0.198	1.625	0.309	0.214
DQTrack [32]	PETRv2	0.234	1.545	0.332	0.269
ADA-Track [13]	PETR	0.249	1.538	0.353	0.270
PromptTrack [72]	PETR	0.259	1.513	0.366	0.280
RefProg (Ours)	PETR	0.265	1.278	0.498	0.274
RefProg (Ours)	StreamPETR [67]	**0.321**	**1.238**	**0.504**	**0.329**

the current position and velocity of an object relative to the ego vehicle. Notably, the *LLMs as a Black Box* baseline significantly improves over prior approaches with a HOTA-Temporal of 37.2%. Finally, our proposed *RefProg* baseline outperforms all other zero-shot methods. Notably, RefProg with LE3DE2E tracks achieves a 13.8% improvement in HOTA-Temporal over *LLMs as a Black Box* with Le3DE2E tracks.

Scenario Mining Challenge. We hosted a challenge at CVPR 2025 to encourage broad community involvement in addressing spatio-temporal scenario mining. Our competition received submissions from eight teams. Notably, four teams beat our best baseline. We present the top three entries at the bottom of Table 2, summarize their contributions in Appendix H and include a link to full technical reports and code here. The best performing team beats our baseline by 3.3% achieving 53.4 HOTA-Temporal.

Impact of LLMs on Code Generation. We evaluate the impact of using different LLMs on code generation quality in RefProg. We use Le3DE2E's tracks and a consistent prompt for all experiments in this ablation. We find that Claude 3.7 Sonnet performs the best (cf. Table 3), achieving a HOTA-Temporal of 50.1. We posit that this is because it has been explicitly tuned for code generation and instruction following. Gemini 2.0 Flash performs considerably worse. Interestingly, we find that Gemini-2.5-Flash and Qwen-2.5-Instruct struggle to generate valid programs, and instead try to generate full programs using invalid import statements. Failure Rate indicates the percentage of generated programs that throw an exception while running.

Evaluating RefProg on nuPrompt. The atomic functions used in RefAV's dataset generation and RefProg represent fundamental motion primitives that are generalizable across datasets. To this end, we evaluate RefProg on nuPrompt [72], a popular referential-tracking dataset based on nuScenes in Table 4. Notably, we find that our zero-shot compositional approach outperforms current state-of-the-art methods trained on domain-specific data. We make minimal adaptations to RefProg for the nuScenes dataset. We swap out the AV2 annotated class list for the nuScenes annotated class list and remove atomic functions that require access to an HD map. Importantly, we make zero modifications to our atomic action definitions, and expect that dataset-specific modifications can improve performance further.

Analysis of Failure Cases. While RefProg outperforms *Image-Embedding Similarity* and *ReferGPT*, each method has unique failure cases. We find that *Image-Embedding Similarity* performs poorly because the CLIP image embedding for each track lacks the context required to understand multi-agent interactions. Specifically, the 2D projected image only includes the tracked object, and does not contain information about past or future frames. *ReferGPT* similarly suffers from a lack of temporal context. Lastly, we find that RefProg fails in cases where the API listing is not expressive enough for a given prompt (e.g., prompts involving weather and lighting conditions). Although one can always add new atomic actions, this strategy is not scalable.

Limitations and Future Work. The quality of our scenario mining dataset is limited by the quality of the ground truth 3D perception labels in AV2. For example, jittery tracks can lead to poor motion classification over short horizons. We address this issue by significantly post-processing and manually verifying all generated scenarios to mitigate label noise. Although AV2 includes many interesting scenarios (cf. Fig. 3), it is still relatively small compared to industry-scale datasets. Future datasets should be explicitly curated to address spatiotemporal scenario mining.

5. Conclusion

In this paper, we introduce RefAV, a large-scale benchmark designed to evaluate scenario mining. Unlike prior language-based 3D scene understanding tasks, we find that scenario mining poses unique challenges in identifying complex multi-agent interactions. Notably prior referential tracking baselines struggle on this challenging benchmark, demonstrating the limitations of existing methods. Future work should develop models capable of reasoning over complex, multi-modal temporal data.

Acknowledgments

This work was supported in part by the NSF GRFP (Grant No. DGE2140739).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 7, 17

[3] Yifan Bai, Dongming Wu, Yingfei Liu, Fan Jia, Weixin Mao, Ziheng Zhang, Yucheng Zhao, Jianbing Shen, Xing Wei, Tiancai Wang, et al. Is a 3d-tokenized llm the key to reliable autonomous driving? *arXiv preprint arXiv:2405.18361*, 2024. 3

[4] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 3, 5

[5] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[6] Tzoulio Chamiti, Leandro Di Bella, Adrian Munteanu, and Nikos Deligiannis. Refergpt: Towards zero-shot referring multi-object tracking. *arXiv preprint arXiv:2504.09195*, 2025. 3, 5, 7, 13

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 3

[8] Feng Chen, Kanokphan Lertniphonphan, Yaqing Meng, Ling Ding, Jun Xie, Kaer Huang, and Zhepeng Wang. Le3de2e solution for av2 2024 unified detection, tracking, and forecasting challenge. 7, 17, 18

[9] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 3

[10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024. 3

[11] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024. 3

[12] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2car: Taking control of your self-driving car. *arXiv preprint arXiv:1909.10838*, 2019. 3, 5

[13] Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15184–15194, 2024. 8

[14] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2024. 3

[15] Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19135–19144, 2024. 3

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3, 5

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[18] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36:7730–7742, 2023. 1

[19] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 2, 3

[20] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 930–938, 2024. 1

[21] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, et al. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 629–636. IEEE, 2024. 1, 3

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014. 2

[23] Mehar Khurana, Neehar Peri, Deva Ramanan, and James Hays. Shelf-supervised multi-modal pre-training for 3d object detection. *arXiv preprint arXiv:2406.10115*, 2024. 3

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment any-

thing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[25] Kirsten Korosec. Waymo has doubled its weekly robotaxi rides in less than a year. *TechCrunch*, 2025. 1

[26] Kirsten Korosec. A waymo robotaxi got trapped in chick-fil-a drive-through, 2025. Accessed: 2025-04-10. 1

[27] H. Lee, M. Kang, K. Hwang, and Y. Yoon. The typical av accident scenarios in the urban area obtained by clustering and association rule mining of real-world accident reports. *Heliyon*, 10(3):e25000, 2024. 4

[28] Jae-Keun Lee, Jin-Hee Lee, Joohyun Lee, Soon Kwon, and Heechul Jung. Re-voxeldet: Rethinking neck and head architectures for high-performance voxel-based 3d detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7503–7512, 2024. 7, 17

[29] Fuhao Li, Huan Jin, Bin Gao, Liaoyuan Fan, Lihui Jiang, and Long Zeng. Nugrounding: A multi-view 3d visual grounding framework in autonomous driving. *arXiv preprint arXiv:2503.22436*, 2025. 1, 5

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. 3

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 3

[32] Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, and Jose Alvarez. End-to-end 3d tracking with decoupled queries. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 8

[33] Yunhao Li, Xiaoqiong Liu, Luke Liu, Heng Fan, and Libo Zhang. Lamot: Language-guided multi-object tracking. *arXiv preprint arXiv:2406.08324*, 2024. 3

[34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 3

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[36] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36: 37193–37229, 2023. 3

[37] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 17, 18

[38] 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. 6

[39] Yechi Ma, Neehar Peri, Shuoquan Wei, Wei Hua, Deva Ramanan, Yanan Li, and Shu Kong. Long-tailed 3d detection via 2d late fusion. *arXiv preprint arXiv:2312.10986*, 2023. 3

[40] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:19547–19560, 2024. 3

[41] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2

[42] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 1, 3

[43] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. *arXiv preprint arXiv:2502.06787*, 2025. 2

[44] Dan Mihalascu. Cruise av robotaxi gets stuck in wet concrete in san francisco, 2023. Accessed: 2025-04-10. 1

[45] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 3

[46] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8602–8612, 2023. 3

[47] Aljosa Osep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixe. Better call sal: Towards learning to segment anything in lidar. In *ECCV*, 2024. 3

[48] Bastian Pätzold, Jan Nogga, and Sven Behnke. Leveraging vision-language models for open-vocabulary instance segmentation and tracking. *arXiv preprint arXiv:2503.16538*, 2025. 3

[49] Neehar Peri, Achal Dave, Deva Ramanan, and Shu Kong. Towards Long Tailed 3D Detection. *CoRL*, 2022. 3

[50] Neehar Peri, Mengtian Li, Benjamin Wilson, Yu-Xiong Wang, James Hays, and Deva Ramanan. An empirical analysis of range for 3d object detection. *arXiv preprint arXiv:2308.04054*, 2023. 4

[51] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 1, 3, 5, 16

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 3, 6

[53] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[54] Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-vl: A multi-domain object detection benchmark for vision-language models, 2025. 3

[55] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022. 3

[56] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 1, 3

[57] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 3

[58] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pages 256–274. Springer, 2024. 1, 5

[59] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3

[60] Didac Suris, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[61] Ayca Takmaz, Cristiano Saltori, Neehar Peri, Tim Meinhardt, Riccardo de Lutio, Laura Leal-Taixe, and Aljosa Osep. Towards learning to complete anything in lidar. *ArXiv*, abs/2504.12264, 2025. 3

[62] Trisha Thadani. Woman stuck under cruise self-driving car after getting hit by a driver, 2023. Accessed: 2025-04-10. 1

[63] Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *arXiv preprint arXiv:2504.03164*, 2025. 1, 3

[64] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 6

[65] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[66] Arun Balajee Vasudevan, Neehar Peri, Jeff Schneider, and Deva Ramanan. Planning with adaptive world models for autonomous driving. *arXiv preprint arXiv:2406.10714*, 2024. 1

[67] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 8

[68] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv:2405.01533*, 2024. 1, 3, 5

[69] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020. 17, 18

[70] 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, 5

[71] Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14633–14642, 2023. 1, 3, 5, 16

[72] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *AAAI 2025*, 2025. 1, 3, 5, 8, 16

[73] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 3

[74] Yihong Xu, Éloi Zablocki, Alexandre Boulch, Gilles Puy, Mickael Chen, Florent Bartoccioni, Nermin Samet, Oriane Siméoni, Spyros Gidaris, Tuan-Hung Vu, et al. Valeo4cast: A modular approach to end-to-end forecasting. *arXiv preprint arXiv:2406.08113*, 2024. 7, 17, 18

[75] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *CVPR*, 2023. 1

[76] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 8

[77] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2

[78] Yani Zhang, Dongming Wu, Wencheng Han, and Xingping Dong. Bootstrapping referring multi-object tracking. *arXiv preprint arXiv:2406.05039*, 2024. 3

[79] Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Toward unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2125–2135, 2024. 3

[80] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification, 2023. 3