AiSCREAM: Absolute Target Localization with Language-Conditioned Cross-View Alignment for Autonomous Vehicles

Kei Katsumata[1], Jun Piao[2], Naoki Hosomi[2], Kentaro Yamada[2], and Komei Sugiura[1]

[1] Keio University, Japan
{ke59ka77, komei.sugiura}@keio.jp
[2] Honda R&D Co., Ltd., Japan
{jun_piao, naoki_hosomi, kentaro_yamada}@jp.honda

Abstract. Autonomous personal mobility benefits greatly from the ability to predict executable spatial goals from language and front view image alone. In this study, we focus on absolute target position prediction from a language instruction and front view image, which requires referring expression disambiguation and absolute target localization without depth cues. To address this, we propose AiSCREAM, a language-conditioned target localization model based on cross-view vision-language reasoning. AiSCREAM constructs cross-view geometric semantic alignment between the front and an Aerial Semantic view, which is an overhead-view image synthesized from the front image via a text-conditioned image-to-image generation model. By preserving appearance-level scene characteristics, the alignment can leverage language-conditioned semantic cues and enable reliable absolute target localization from a single image. To evaluate AiSCREAM, we constructed DRAMATiST, a benchmark that introduces road and traffic diversity. The experimental results demonstrated that AiSCREAM achieved a mean absolute position error of 4.03 m on DRAMATiST, which outperformed baseline methods and human performance. The project page is available at `https://ai-scream-project-page.vercel.app/` .

Keywords: Vision and Language · Autonomous Driving · Language-conditioned Target Localization

1 Introduction

Professional labor shortages in transportation services increasingly constrain equitable access to mobility. To mitigate these challenges, prior work has explored autonomous personal mobility that assists users with mobility impairments [20, 28, 44]. Such a system is expected to facilitate user-friendly interactions. In particular, comprehending language instructions in complex surroundings is essential for safe and effective navigation. However, most existing methods exhibit limited performance for such instructions.

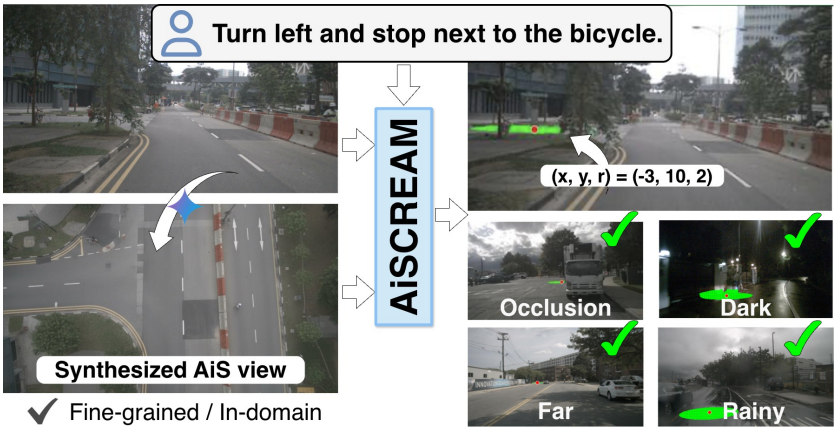

Fig. 1: Overview of AiSCREAM. Given a language instruction and front camera image, AiSCREAM predicts the target position and distribution in an ego-centric coordinate system. AiSCREAM constructs a cross-view geometric semantic alignment between the front view and an Aerial Semantic (AiS) view, an overhead-view image synthesized from the front image via a text-conditioned image-to-image generation.

In this study, we focus on predicting an absolute target position from a language instruction and a front camera image of the mobility. Fig. 1 shows a typical use case for our task. When a user gives the instruction, "Turn left and stop next to the bicycle," it is required to predict the target position and distribution in an ego-centric coordinate system. Our target task is challenging because it requires grounding the language instruction to the visual scene to localize the intended target position, and it also requires absolute target localization under limited geometric and spatial information.

Existing methods [20, 28, 44] including multimodal large language models (MLLMs) often fail in such a task, and even humans exhibit large absolute position error (see Sec. 4). Conventional approaches [20, 28, 44] are insufficient because most of them do not use spatial information effectively. A naive extension would be to use a classic bird's-eye-view (BEV) representation [14, 33, 51, 58], which is typically formed by metrically projecting features from the front view onto the ground plane. However, in such projections, semantics are often reduced to a few coarse categories, which results in weak cross-view cues with the front image. As a result, reliably achieving absolute localization while maintaining consistent instruction-conditioned language grounding remains challenging.

To address this limitation, we propose AiSCREAM, a language-conditioned target localization model for an autonomous personal mobility. AiSCREAM constructs cross-view geometric semantic alignment between the front and an Aerial Semantic (AiS) view, which is an overhead-view image synthesized from the front image via a text-conditioned image-to-image generation model [19, 53]. By preserving appearance-level scene characteristics, this alignment can leverage language-conditioned semantic cues and enables more reliable absolute target localization from a single image. AiSCREAM differs from existing methods in that it performs image-domain alignment between the front and an AiS view

image, rather than a simple semantic-segmentation approach for the front view. This design enables the proposed model to construct geometric semantic cue, which improves the reliability of absolute target position prediction.

Our key contributions are as follows:

- We propose a cross-view geometric semantic alignment representation, constructed from language-conditioned semantic segmentations of the front and AiS view images.
- We introduce the Height-based Cross-view (HC) loss to penalize initial predictions that overlap with obstacle regions, and incorporate MLLM Reshape Localizer to refine the initial predictions.
- We construct DRAMATiST, a benchmark for front view images and navigation instructions that introduces road and traffic diversity beyond the scope of existing benchmarks.
- Experimental results show that AiSCREAM outperformed baseline methods and even human performance in terms of standard metrics on the GRiN-Drive and DRAMATiST benchmarks.

2 Related Work

Vision and language models (VLMs) for autonomous vehicles have been studied actively. Progress in foundation models such as VLMs and MLLMs has introduced strong semantic priors into driving perception [10, 26, 59, 66]. Zhu *et al.* and Jiang *et al.* conducted comprehensive surveys of vision and language modeling for autonomous driving, covering MLLM-based approaches and vision and language action (VLA) models [26, 66].

2.1 Visual Grounding

Visual grounding in driving scenarios has been studied extensively across a range of tasks, including referring expression comprehension (REC) tasks [13, 15, 27, 37, 54, 61, 63, 65], referring expression segmentation (RES) tasks [5–7, 31, 32, 39, 55, 67, 68], and referring navigable region (RNR) tasks [20, 23, 24, 28, 42, 44]. REC aims to localize an object specified by a natural language expression using a bounding box [13, 15, 27, 37, 54, 61, 63, 65]. In autonomous driving contexts, this formulation has been instantiated through datasets such as Talk2Car [12], which introduced REC into real driving scenes [15, 37, 54, 63]. Building on this formulation, in previous studies, image-based 3D grounding was explored, where language queries localize referred objects as 3D boxes in driving scenes from monocular or multi-view observations [22,25,62]. These approaches include BEV-centric formulations that ground language on BEV representations to predict 3D boxes [8, 16, 21].

To overcome the limitations of box-based localization for target grounding at finer spatial granularity, RES was introduced. RES aims to predict a pixel-level mask corresponding to the region referred to by the expression [5–7,31,32,55,68], which provides finer spatial grounding than box-based localization. Among these formulations, the RNR task focuses on segmenting the navigable region specified by a navigation instruction [20,23,24,28,42,44]. PDPC [20] uses a top-down layout that fails to preserve instruction-specified colors, because its representation

is effectively at the line drawing level. LeGo-Drive predicts goal locations from image–instruction inputs using an explicit goal prediction module [42]. GEN-NAV generalizes destination grounding to handle diverse instruction conditions, including the existence and multiplicity of plausible targets [28].

The proposed method differs from existing approaches [20, 28, 44] that it introduces cross-view geometric semantic alignment between the front view and an Aerial Semantic views.

2.2 Language Models for Autonomous Driving

Language models for autonomous driving have recently shown promising results by leveraging multimodal understanding and natural interaction with humans [4, 9, 35, 38, 45–47, 69]. Existing approaches often use language models to reason about driving scenes and predict trajectories or control commands [1, 47, 56, 57]. More recent approaches, including VLA models [17, 43, 52], move closer to execution by explicitly connecting language-based reasoning with numerical planning and control outputs and improving the alignment between semantics and actions. For instance, ORION [17] bridges large language model (LLM)-style reasoning and trajectory prediction through a dedicated planner to maintain consistency between high-level intent and motion. SimLingo [43] jointly models driving and language-centric tasks to strengthen language–action alignment, thereby producing behaviors that compatible with query understanding. Typical VLA settings couple language-conditioned scene understanding with action generation under richer and more heterogeneous inputs (e.g., multi-camera, LiDAR, and top-down BEV representations) [17, 43, 52], whereas RNR focuses on grounding the target position from a navigation instruction and front camera image.

2.3 Benchmarks

The surveys [10, 66] summarize commonly used benchmarks and datasets for autonomous driving scene understanding, and highlight representative resources such as Waymo Open Dataset [49], BDD100K [60], KITTI [18], and nuScenes [3]. Waymo Open Dataset provides synchronized camera–LiDAR data with strong 3D detection and tracking annotations, whereas nuScenes offers 360° surround sensing with cameras, radar, and LiDAR, thereby serving as a widely used multi-view 3D detection benchmark. Building on nuScenes, Talk2Car [12] augments data with navigation commands that contain landmark-based referring expressions and annotations for the referred landmarks, whereas GRiN-Drive [28] further extends this setting by providing pixel-wise segmentation masks for intended navigable regions. Complementing these scale-oriented benchmarks, THMA [50] and DRAMA [36] broaden dataset diversity from the perspective of data sources and geographic coverage.

3 Method

3.1 Problem Statement

In this paper, we focus on the Referring Navigable Position Localization (RNPL) task, which involves identifying the absolute target position in an ego-centric coordinate system given a natural language navigation instruction and a front

camera image taken from a moving mobility. In this task, the model is expected to predict the target position and distribution indicating an acceptable region.

Fig. 1 shows an example of the RNPL task. Given a natural-language instruction (e.g., "Turn left and stop next to the bicycle.") and a front camera image, the goal is to predict the target position and distribution in an ego-centric coordinate, visualized as a red point and a green region, respectively (e.g., $(x, y, r) = (-3.0\,[\mathrm{m}], 10.0\,[\mathrm{m}], 2.0\,[\mathrm{m}])$). We assume that the target position is visible within the image. Furthermore, we do not address trajectory prediction toward the target position, because existing lower-level planning and control modules are already capable of generating feasible trajectories with sufficient reliability in practical applications.

We define the terms used in this study as follows: A "navigation instruction" is a natural language instruction that guides the mobility to a destination. The "target position" is the position specified in the given navigation instruction. A "landmark" is to an object used as a reference when the target position is specified.

3.2 Overview

We propose AiSCREAM, a language-conditioned target localization model for an autonomous personal mobility. AiSCREAM extends language-conditioned localization methods [20, 24, 28, 42, 44, 55, 68] with cross-view alignment and prediction-conditioned refinement. In our method, we introduce a cross-view geometric semantic alignment representation that is constructed using semantic segmentation from both the front and Aerial Semantic (AiS) views as structural cues. The AiS view is an overhead image synthesized from the front image using a text-conditioned image-to-image generation model (e.g., Nano Banana [19] and Qwen Image Edit [53]). Because the synthesis preserves appearance-level scene characteristics, the alignment provides language-conditioned structural cues. The framework jointly captures appearance semantics and ego-centric spatial structure through complementary cross-view signals, and is applicable to vision–language grounding models for autonomous vehicles.

Fig. 2 shows the structure of AiSCREAM. AiSCREAM consists of two main modules: AiS Cross-View Representation Alignment Module (AiRAM), and Multi Task Absolute Position Localizer (MTAPL). We define the model input $\boldsymbol{x}$ as follows: $\boldsymbol{x} = \{\boldsymbol{x}_{\mathrm{img}}, \boldsymbol{x}_{\mathrm{inst}}\}$, where $\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{3\times W\times H}$ and $\boldsymbol{x}_{\mathrm{inst}} \in \{0,1\}^{V\times L}$ denote a front camera image and navigation instruction, respectively and W, H, V and L denote the width of the image, height of the image, vocabulary size, and maximum token length, respectively. We pre-process $\boldsymbol{x}_{\mathrm{inst}}$ using Qwen3-Embedding-0.6B [64] to obtain the linguistic features.

3.3 AiS Cross-View Representation Alignment Module

AiRAM introduces a cross-view geometric semantic alignment representation $\boldsymbol{h}_{\mathrm{cv}}$ that is constructed using semantic segmentation from both the front and AiS view images as structural cues. Unlike metric-projection-based BEV or top-down representations, an AiS view is generated in the image domain and preserves appearance-level scene characteristics similar to those in the front view. This

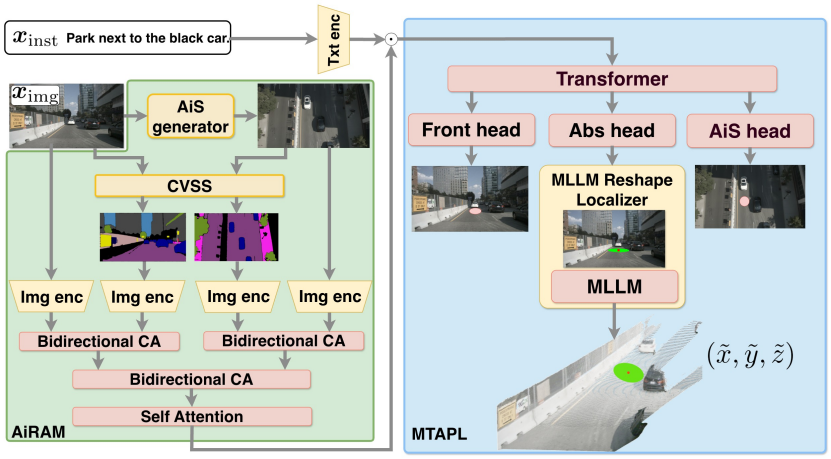

Fig. 2: Overall architecture of AiSCREAM. AiRAM (green) constructs a cross-view geometric semantic alignment representation using semantic segmentations from the front and AiS view images as structural cues. MTAPL (blue) uses a two-stage pipeline that predicts an initial target distribution and then refines it with MLLM Reshape Localizer. "Img enc", "Front head", "Abs head" and "AiS head" represent an image encoder followed by an MLP projection, a front segmentation head, an absolute target prediction head and an AiS view segmentation head, respectively.

allows both views to be processed with a shared image encoder and semantic mask extraction, which enables cross-view semantic alignment through language.

Given $\boldsymbol{x}_{\text{img}}$, we first generate an AiS view image $\boldsymbol{x}_{\text{ais}}$ using the AiS generator. The AiS generator synthesizes an overhead view from a front image using a text-conditioned image-to-image generation model (e.g., Nano Banana [19] and Qwen Image Edit [53]). Then we obtain semantic segmentations for both views using cross-view consistent semantic segmentation (CVSS).

CVSS uses language-conditioned segmentation models (e.g., SAM3 [5]), using text prompts to specify semantic categories that characterize the scene. The resulting masks remain comparable across viewpoints. The use of CVSS is important because standard semantic segmentation models are trained on non AiS view image distributions, which can introduce view-dependent prediction bias when applied to AiS view geometry. Therefore, we adopt language-conditioned segmentation, which conditions predictions on language-specified categories and yields semantic cues that are shared across viewpoints. Unlike metric-projection-based BEV representations that encode semantics with only a few coarse categories, our approach yields fine-grained per-pixel estimations, which provides stronger cross-view cues for aligning the front and AiS view images.

Next, given $\boldsymbol{x}_{\text{img}}$ and $\boldsymbol{x}_{\text{ais}}$, we obtain $\boldsymbol{h}_{\text{img}}$ and $\boldsymbol{h}_{\text{ais}}$ by encoding each image with an image encoder [41, 48] with an MLP projection. Similarly, given the CVSS segmentation outputs of $\boldsymbol{x}_{\text{img}}$ and $\boldsymbol{x}_{\text{ais}}$, we obtain $\boldsymbol{h}_{\text{seg}}$ and $\boldsymbol{h}_{\text{aisseg}}$ using the same encoder and projection head. To inject structural cues and align the two views, we use bidirectional cross-attention (BiCA), defined as

$$\text{BiCA}(\boldsymbol{a}, \boldsymbol{b}) = \big[\text{CA}(\boldsymbol{a}, \boldsymbol{b}),\ \text{CA}(\boldsymbol{b}, \boldsymbol{a})\big], \tag{1}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two arbitrary vectors, $\mathrm{CA}(\boldsymbol{a}, \boldsymbol{b})$ performs cross-attention using $\boldsymbol{a}$ as queries and $\boldsymbol{b}$ as keys and values, and $[\cdot, \cdot]$ represents concatenation.

We fuse the image and its segmentation results within each view using BiCA, and then align the resulting fused representations across views using BiCA. The final cross-view representation $\boldsymbol{h}_{\mathrm{cv}}$ is computed as

$$\boldsymbol{h}_{\mathrm{cv}} = \mathrm{SA}\Big(\mathrm{BiCA}\big(\mathrm{BiCA}(\boldsymbol{h}_{\mathrm{img}}, \boldsymbol{h}_{\mathrm{seg}}), \mathrm{BiCA}(\boldsymbol{h}_{\mathrm{ais}}, \boldsymbol{h}_{\mathrm{aisseg}})\big)\Big), \tag{2}$$

where $\mathrm{SA}(\cdot)$ denotes self-attention. This design allows the semantic segmentation masks to act as explicit structural anchors, whereas cross-view attention integrates semantic cues with spatial structure across viewpoints to yield $\boldsymbol{h}_{\mathrm{cv}}$ that jointly captures semantics and ego-centric geometry.

3.4 Multi Task Absolute Position Localizer

MTAPL consists of a two-stage prediction pipeline. It first predicts an initial target position and distribution, and then refines this prediction using MLLM Reshape Localizer. For the initial prediction stage, MTAPL predicts an absolute target position $\hat{\boldsymbol{y}}_{\mathrm{ais}}$, which serves as the initial prediction. In the second stage, an MLLM is used to refine the initial prediction. Unlike direct MLLM-based prediction, the localizer performs prediction-conditioned refinement by conditioning on an overlaid point and its predicted numeric value, which enables correction relative to the referenced prediction.

The module takes $\boldsymbol{h}_{\mathrm{cv}}$ and the language embeddings of $\boldsymbol{x}_{\mathrm{inst}}$ as inputs. They are fused via a Hadamard product and encoded by a transformer to obtain a shared latent representation. Given this representation, a front segmentation head, AiS view segmentation head, and absolute target prediction head predict $\hat{\boldsymbol{y}}_{\mathrm{front}} \in \mathbb{R}^3$, $\hat{\boldsymbol{y}}_{\mathrm{ais}} \in \mathbb{R}^3$, and $\hat{\boldsymbol{y}}_{\mathrm{abs}} \in \mathbb{R}^3$, respectively. The front segmentation head and absolute target head serve as primary heads, and provide region-level target cues and direct positional estimates, respectively, whereas the AiS view segmentation head is trained with an auxiliary loss (see Eq. 3) to provide complementary AiS view structural cues.

In the second stage, we further refine the initial prediction using MLLM Reshape Localizer that conditions on a visual prompt constructed by overlaying $\hat{\boldsymbol{y}}_{\mathrm{abs}}$ onto $\boldsymbol{x}_{\mathrm{img}}$. Given the $\hat{\boldsymbol{y}}_{\mathrm{abs}}$ overlaid image and $\boldsymbol{x}_{\mathrm{inst}}$, the localizer assesses whether the predicted target position and distribution are physically and contextually plausible, and predicts a corrected target position and distribution $\tilde{\boldsymbol{y}}_{\mathrm{abs}}$ when the initial prediction is judged implausible. Details of the prompt are provided in the Supplementary Material. Finally, the model outputs $\tilde{\boldsymbol{y}}_{\mathrm{abs}} = (\tilde{x}, \tilde{y}, \tilde{r})$, where $(\tilde{x}, \tilde{y})$ denotes the target position coordinates and $\tilde{r}$ denotes the radius of a circular region representing its target distribution.

3.5 Height-based Cross-View Loss

We introduce the HC loss, which incorporates complementary cross-view supervision terms and penalizes predictions that overlap with obstacle regions. The HC loss is defined as:

$$\mathcal{L}_{\mathrm{HC}} = \lambda_{\mathrm{abs}} \ell_1(\hat{\boldsymbol{y}}_{\mathrm{abs}}, \boldsymbol{y}_{\mathrm{abs}}) + \lambda_{\mathrm{ais}} \ell_1\left(\hat{\boldsymbol{y}}_{\mathrm{ais}}, \boldsymbol{y}_{\mathrm{ais}}\right) + \lambda_{\mathrm{obs}} \mathcal{L}_{\mathrm{obs}}, \tag{3}$$

where $\boldsymbol{y}_{\text{abs}}$ and $\boldsymbol{y}_{\text{ais}}$ denote the ground-truth target position and ground-truth region in $\boldsymbol{x}_{\text{ais}}$. Furthermore, $\mathcal{L}_{\text{obs}}$ penalizes overlapping of the predicted target region and an obstacle mask. The overall loss function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \ell_1(\hat{\boldsymbol{y}}_{\text{front}}, \boldsymbol{y}_{\text{front}}) + \mathcal{L}_{\text{HC}}, \tag{4}$$

where $\ell_1(\cdot, \cdot)$ and $\boldsymbol{y}_{\text{front}}$ denote the ℓ_1 loss and the ground-truth target region in $\boldsymbol{x}_{\text{img}}$, respectively.

4 Experiments

4.1 Experimental Setup

Benchmark. We used the single target subset of the GRiN-Drive [28] benchmark and our newly constructed DRAMATiST benchmark for the RNPL task. The dataset for the RNPL task is expected to include front camera images annotated with natural language instructions and corresponding target positions. The GRiN-Drive benchmark provides such instructions and images with annotations and serves as a standard benchmark for this setting. However, using a single benchmark alone can be insufficient because benchmark-specific data collection can induce dataset biases. For example, the single target subset of GRiN-Drive is built on nuScenes [3], which was collected in only a few urban areas. As a result, the distributions of traffic rules, road layouts, vehicle types, and urban visual appearance can be biased toward those regions, and evaluation on such data may overestimate performance. To enable a controlled assessment under the RNPL annotation format, we constructed a novel DRAMATiST benchmark based on the DRAMA dataset [36]. This benchmark enables the systematic evaluation of cross-regional and cross-traffic-systems.

We constructed the DRAMATiST benchmark from the DRAMA dataset [36], which targets joint risk localization and captioning in driving. The DRAMA dataset provides noun phrases that denote candidate landmark entities (e.g., pedestrians or objects) in each frame, along with their associated bounding boxes. Using these noun phrases, we generated navigation instructions by populating predefined templates (e.g., "Stop next to <noun phrase>") and asked human annotators to identify the corresponding target position. Details of the annotation procedure are provided in the Supplementary Materials.

Finally, we converted the target position annotations into 3D target positions using a monocular depth estimator [30] and applied the same conversion procedure to the other benchmarks for consistency. Specifically, we defined the centroid of the annotated binary mask as the 2D target center and the radius r of the largest inscribed circle within the mask as the target region. Using a monocular depth estimation model, we converted the 2D target center into a 3D target position in the egocentric coordinate system. We stored this target position as the ground-truth target for each sample in the benchmark.

The statistics of GRiN-Drive and DRAMATiST benchmarks are summarized as follows: they contain 10,020 and 6,241 samples, and have vocabulary sizes of 2,736 and 573, total word counts of 106,972 and 108,771, and average sentence lengths of 10.68 and 17.43 words, respectively. The target positions of

the DRAMATiST benchmark were provided by 218 annotators, averaging 28.6 samples per annotator.

We divided the GRiN-Drive benchmark into training, validation, and test sets, which contained 8,349, 1,163, and 508 samples, respectively. We divided the DRAMATiST benchmark into training, validation, and test sets, which contained 4,362, 949, and 930 samples, respectively. We followed the original split [28, 36] because adopting a different dataset split would lead to an unfair comparison with the baseline methods. We used each training set to train the models, each validation set for adjusting their hyperparameter, and the corresponding test set to evaluate the models. We resized the original images to 224×224 to reduce computational cost.

Baselines. We used Rufus *et al.* [44], GSVA [55], GENNAV [28], SAM3 [5], SimLingo [43], ORION [17], Alpamayo-R1 [52], Qwen3-VL-8B [2] and GPT-5.2 [40] as the baseline methods. We used each method as a baseline method for the following reasons: We selected Rufus *et al.*, GSVA, GENNAV, and SAM3 as the baseline methods because of its successful application to the GRNR task, which is closely related to the RNPL task. Moreover, GPT-5.2 and Qwen3-VL are representative multimodal LLMs that have been pre-trained on large-scale datasets and have demonstrated outstanding performance on various vision and language tasks [2,40]. Finally, we selected ORION, SimLingo, and Alpamayo-R1, an end-to-end VLA method, because they have been reported to achieve strong performance in autonomous driving trajectory generation. In our comparative experiments, we considered three baselines and standardized their evaluation protocols as follows:

Segmentation-based baselines: We fine-tuned Rufus *et al.*, GSVA, GENNAV, and SAM3 using the hyperparameter settings reported in the original papers [5, 28, 44, 55]. Because they are 2D segmentation-based approaches and do not directly output a metric target position, we converted their predicted masks into target positions in metric ego-centric coordinates. Specifically, we (i) computed the centroid of the predicted binary mask as the 2D target center and (ii) calculated the radius r as that of the largest inscribed circle within the mask to obtain a target position. Following the same procedure used in dataset construction, we used a monocular depth model for the conversion into the target position in the ego-centric coordinate system.

Zeroshot VLA baselines: To compare the capabilities of VLA-style policies for autonomous driving, we used ORION [17], SimLingo [43], and Alpamayo-R1 [52] in a zero-shot setting. We executed them on a single front image without temporal history, conditioned on the given instruction, and used its output for comparison.

Zeroshot MLLM baselines: As a baseline comparison with representative MLLMs, we conducted zero-shot evaluation using GPT-5.2 and Qwen3-VL. Prior works reported that such MLLMs can predict approximate metric distances from images, which motivated their inclusion as baselines [11, 29]. We first ran preliminary experiments with more than ten prompt variants and selected the prompt that achieved the best validation performance. Then for robustness, we repeated inference five times by varying the temperature parameter. Details of the prompts are provided in the Supplementary Material.

Table 1: Quantitative comparison between the proposed method and baseline methods on each test set of the GRiN-Drive and DRAMATiST benchmark. The best score for each metric is in **bold** and the second-highest score is underlined.

Method	GRiN-Drive			DRAMATiST		
	RMSE [m]↓	AED [m]↓	mIoU [%]↑	RMSE [m]↓	AED [m]↓	mIoU [%]↑
Rufus *et al.* [44]	35.30 ± 3.51	18.57 ± 1.24	2.64 ± 0.26	37.91 ± 4.75	19.72 ± 1.60	0.98 ± 0.02
GSVA [55]	25.58 ± 0.42	21.83 ± 0.15	2.53 ± 0.41	22.15 ± 0.34	18.15 ± 0.25	0.84 ± 0.02
GENNAV [28]	10.51 ± 0.52	6.35 ± 0.35	7.25 ± 0.31	8.56 ± 0.52	6.45 ± 0.60	2.83 ± 0.49
SAM3 [5]	74.22 ± 0.75	44.67 ±0.43	0.49 ± 0.01	33.85 ± 0.27	16.37 ± 0.14	0.71 ± 0.01
SimLingo [43]	25.64 ± 0.28	21.79 ± 0.19	-	22.19 ± 0.06	18.84 ± 0.19	-
ORION [17]	30.27 ± 0.00	28.34 ± 0.00	-	30.67 ± 0.00	28.22 ± 0.00	-
Alpamayo-R1 [52]	27.15 ± 0.04	23.23 ± 0.04	-	21.42 ± 0.03	17.72 ± 0.02	-
Qwen3-VL-8B [2]	27.05 ± 0.94	23.00 ± 1.23	0.42 ± 0.02	18.25 ± 0.07	13.79 ± 0.06	0.76 ± 0.08
GPT-5.2 [40]	12.44 ± 0.15	9.07 ± 0.13	6.26 ± 0.16	11.29 ± 0.24	7.91 ± 0.08	3.22 ± 0.17
AiSCREAM (ours)	**4.72 ± 0.14**	**3.95 ± 0.09**	**11.89 ± 0.69**	**5.00 ± 0.05**	**4.03 ± 0.02**	**6.16 ± 0.37**
Human	7.45 ± 0.50	4.00 ± 0.20	4.82 ± 0.56	7.00 ± 1.62	4.69 ± 1.20	6.50 ± 0.63

Metrics. We used root mean squared error (RMSE), average euclidean distance (AED), and mean intersection over union (mIoU), which are standard evaluation metrics for regression and segmentation tasks, with RMSE as the primary metric. Details of the metrics are provided in the Supplementary Material.

Implementation Details. Our model had approximately 41.5M trainable parameters and 662M multiply-add operations. We trained our model on a GeForce RTX 4090 with 24 GB of GPU memory and an Intel Core i9-13900KF with 64 GB of RAM. The training time for the proposed model was approximately 3 hours on the GRiN-Drive dataset. The inference time of the proposed model itself was approximately 9 milliseconds per sample. By contrast, the end-to-end inference time was approximately 10.3 seconds, which was largely attributed to pre-processing, particularly the AiS generator (approximately 8 seconds). Experimental settings of the proposed method are provided in Supplementary Materials. In AiSCREAM, we employed the following pre-trained models. In AiRAM, we used Nano Banana [19], SAM3 [5], and DINOv3 [48] as the AiS generator, CVSS, and the image encoder, respectively. Details of the prompt used for the AiS generator are provided in the Supplementary Materials. In MTAPL, we used GPT-5.2 [40] as the MLLM for MLLM Reshape Localizer.

4.2 Quantitative Results

Table 1 shows the quantitative comparison between the proposed method and baseline methods on each test set of the GRiN-Drive benchmark and DRAMATiST benchmark. The best score for each metric is in **bold** and the second-highest score is underlined. The reported values are the mean and standard deviation over five trials.

Table 1 shows that AiSCREAM outperformed the baseline methods for RMSE, AED, and mIoU on the GRiN-Drive and DRAMATiST benchmarks. Specifically, AiSCREAM resulted in an RMSE of 4.72 m on GRiN-Drive and 5.00 m on DRAMATiST, whereas the best baseline reached 10.51 m and 8.56 m, respectively. This corresponds to RMSE reductions of 5.79 m on GRiN-Drive and

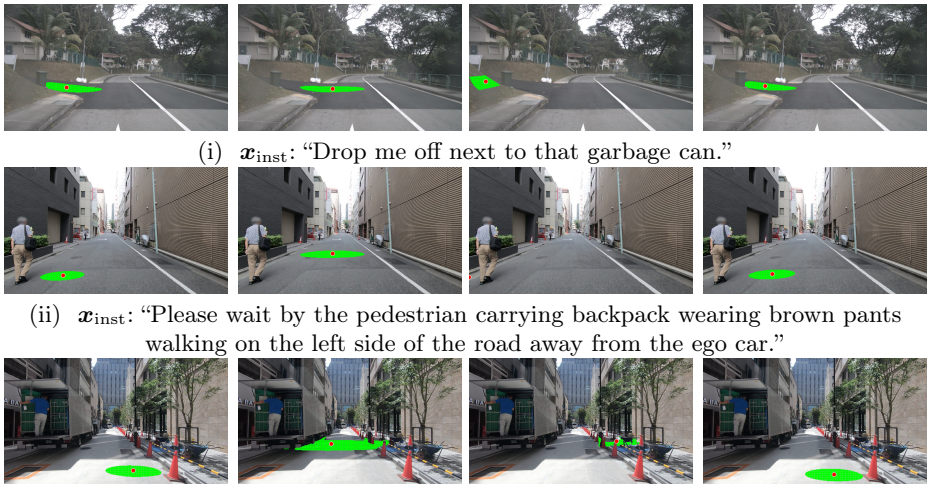

Fig. 3: Qualitative results of the AiSCREAM and baseline methods. Columns (a)–(d) show the ground-truth target position and distribution, and the corresponding predictions by GENNAV, GPT-5.2, and AiSCREAM. Example (i) corresponds to GRiN-Drive sample, whereas (ii) and (iii) come from DRAMATiST. The red point and green region indicate the target position and distribution, respectively.

3.56 m on DRAMATiST. To contextualize these errors, a distance of 4–5 m corresponds to roughly a few walking steps for a user. By contrast, the strongest baseline in Table 1 remains at an error scale of 10 m, which is consistent with lane-level (or larger) mislocalization. This constitutes a substantial improvement. Moreover, AiSCREAM consistently outperformed the baselines across the other evaluation metrics. The improvement achieved by the proposed method was statistically significant with $p < 0.01$.

We also conducted a human subject experiment to evaluate human performance. From the GRiN-Drive and DRAMATiST benchmarks, we randomly sampled 250 instances from each dataset, yielding 500 samples in total. Each sample was evaluated by three unique participants, leading to an average of 150 predictions per participant and a total of 1,500 predictions across 10 participants. We asked participants to predict the target position specified by the navigation instruction using the provided front image.

Human performance, measured by RMSE and AED, was 7.00 m and 4.69 m on DRAMATiST, as shown in Table 1. By comparison, AiSCREAM achieved an RMSE and AED of 5.00 m and 4.03 m on DRAMATiST. Although the human RMSE and AED values were relatively large, this was mainly because of the experimental protocol, which restricted the input to only a single front image. Human participants are likely to achieve better performance in real-world settings with full environmental context and interactive perception.

$\boldsymbol{x}_{\text{inst}}$: "That bus has stopped; drive around it and continue straight."
(a) GT (b) GENNAV (c) GPT-5.2 (d) AiSCREAM

Fig. 4: Qualitative results for a representative failure case. Columns (a)–(d) show the ground-truth target position and distribution, and the corresponding predictions by GENNAV, GPT-5.2, and AiSCREAM. The red point and green region indicate the target position and distribution, respectively.

4.3 Qualitative Results

Fig. 3 shows successful cases of AiSCREAM. Columns (a)–(d) show the ground-truth target position and distribution, and the corresponding predictions by GENNAV, GPT-5.2, and AiSCREAM. Example (i) corresponds to GRiN-Drive samples, whereas Columns (ii) and (iii) come from DRAMATiST. The red point and green region indicate the target position and distribution, respectively.

In Fig. 3 (i), $\boldsymbol{x}_{\text{inst}}$ was "Drop me off next to that garbage can." The models were required to predict the target position next to the garbage can on the left side of the image. AiSCREAM successfully predicted the target position and distribution next to the garbage can. Although the baseline methods also predicted positions close to the landmark, the predicted targets were often displaced or located in non-road regions, which indicates that the relative spatial relationship to the landmark was not properly modeled.

Fig. 3 (ii) presents an example where $\boldsymbol{x}_{\text{inst}}$ was "Please wait by the pedestrian carrying backpack wearing brown pants walking on the left side of the road away from the ego car." Because the instruction explicitly refers to that pedestrian, the target position should be predicted near the pedestrian on the left side. Our proposed method appropriately predicted the target position, whereas GENNAV failed to model the relative position between the referenced landmark and the ego vehicle and GPT-5.2 predicted a potentially hazardous target position by disregarding the relative position between the landmark and the ego vehicle.

Fig. 3 (iii) illustrates the case in which $\boldsymbol{x}_{\text{inst}}$ was, "Pull over next to the orange traffic cones on the right side of the road." In this example, the target position should be predicted on the left side of the orange traffic cones positioned along the right side of the road. The baseline method, GENNAV, predicted a target on the left side of the road and produced an overly broad distribution that extended into the truck region, which led to an invalid prediction. GPT-5.2 predicted a position near the cones; however, the predicted point lay on the opposite, non-navigable side and was therefore inaccessible to the vehicle. By contrast, AiSCREAM appropriately localized both the target position and its distribution within the drivable road region adjacent to the orange cones on the right side of the image.

Fig. 4 illustrates a qualitative example of a failure case. Columns (a)–(d) show the ground-truth target position and distribution, and the corresponding

Table 2: Quantitative results of the ablation studies of modules on GRiN-Drive and DRAMATiST. RL, AiS, and CVSS denote MLLM Reshape Localizer, Aerial Semantic view, and cross-view consistent semantic segmentation, respectively. The best score for each metric is in **bold**.

Model	Condition			GRiN-Drive			DRAMATiST		
	RL	AiS	CVSS	RMSE [m]↓	AED [m]↓	mIoU [%]↑	RMSE [m]↓	AED [m]↓	mIoU [%]↑
(i)	✓	✓	✓	**4.72 ± 0.14**	**3.95 ± 0.09**	**11.89 ± 0.59**	**5.00 ± 0.05**	**4.03 ± 0.02**	**6.16 ± 0.37**
(ii)		✓	✓	4.95 ± 0.08	4.18 ± 0.06	10.48 ± 0.37	5.19 ± 0.07	4.29 ± 0.08	5.90 ± 0.58
(iii)			✓	6.60 ± 0.10	5.61 ± 0.10	8.24 ± 0.72	5.57 ± 0.13	4.47 ± 0.16	3.35 ± 0.35
(iv)		✓		5.80 ± 0.10	4.83 ± 0.06	9.50 ± 0.33	5.34 ± 0.06	4.31 ± 0.07	3.43 ± 0.23

predictions by GENNAV, GPT-5.2, and AiSCREAM. This sample is taken from the GRiN-Drive benchmark. The red point and green region indicate the target position and distribution, respectively. In this example, $\boldsymbol{x}_{\text{inst}}$ was: "That bus has stopped; drive around it and continue straight." The target should have been predicted in front of the bus; however, GENNAV, GPT-5.2, and AiSCREAM predicted a target position behind the bus. This failure is likely to have arisen because of the difficulty of resolving an object-centered frame of reference, which requires interpreting spatial relations relative to the object's orientation.

4.4 Ablation Study

Tables 2 and 3 show the results of the ablation studies. RL, AiS, CVSS, $\mathcal{L}_{\text{ais}}$ and $\mathcal{L}_{\text{obs}}$ represent MLLM Reshape Localizer, Aerial Semantic view, cross-view consistent semantic segmentation, AiS view segmentation term and obstacle penalty term, respectively. To examine whether the performance gap between AiSCREAM and the baseline methods is due to the AiS view images alone, we conducted another ablation study. The results are provided in the Supplementary Material.

MLLM Reshape Localizer ablation: We omitted the MLLM Reshape Localizer to analyze its contribution. Compared with the full Model (i), Model (ii) yielded a longer RMSE by 0.23 m on GRiN-Drive and 0.19 m on DRAMATiST. This suggests that prediction-aware visual prompting helped the MLLM to refine occasional implausible target predictions.

AiS ablation: To evaluate the contribution of the AiS view independently of the MLLM Reshape Localizer, we compare with Model (ii). Removing the AiS view images (iii) increased the RMSE by 1.65 m on GRiN-Drive and 0.38 m on DRAMATiST relative to (ii). This indicates that the AiS view provided complementary cues that helped to reduce the localization error.

CVSS ablation: We assessed the contribution of CVSS by removing it from Model (ii). Model (iv) achieved RMSEs of 5.80 m and 5.34 m on the GRiN-Drive and DRAMATiST benchmarks, which were 0.85 m and 0.15 m larger than those for Model (ii), respectively. This supports that CVSS contributed fine-grained structural cues for target grounding, which improved cross-view alignment.

HC loss ablation: Since MLLM Reshape Localizer is a second-stage refinement module, we ablate HC loss terms in the RL-disabled setting (Model (a)) to directly assess their impact on the initial localization. We evaluated the effectiveness of the HC loss by omitting each component term. Table 3 shows that

Table 3: Quantitative results of the ablation studies of HC loss terms on GRiN-Drive and DRAMATiST. $\mathcal{L}_{\text{ais}}$ and $\mathcal{L}_{\text{obs}}$ denote the AiS-view segmentation term and the obstacle penalty term, respectively. The best score for each metric is in **bold**.

Model	HC loss		GRiN-Drive			DRAMATiST		
	$\mathcal{L}_{\text{ais}}$	$\mathcal{L}_{\text{obs}}$	RMSE [m]↓	AED [m]↓	mIoU [%]↑	RMSE [m]↓	AED [m]↓	mIoU [%]↑
(a)	✓	✓	**4.95 ± 0.08**	**4.18 ± 0.06**	**10.48 ± 0.37**	**5.19 ± 0.07**	**4.29 ± 0.08**	**5.90 ± 0.58**
(b)		✓	5.17 ± 0.06	4.26 ± 0.04	9.86 ± 0.48	5.24 ± 0.02	4.35 ± 0.03	5.30 ± 0.29
(c)	✓		5.27 ± 0.16	4.57 ± 0.14	9.93 ± 0.46	5.24 ± 0.06	4.30 ± 0.05	4.06 ± 0.69

Table 4: Quantitative results of the additional ablation studies on the use of AiS-view and front-view inputs on GRiN-Drive and DRAMATiST. Simply using either the AiS view alone or the front view alone did not improve performance to the same extent as the full model. The best score for each metric is in **bold**. For AiSCREAM , we used the initial-stage results for a fair comparison.

Model	Condition		GRiN-Drive			DRAMATiST		
	Front	AiS	RMSE [m]↓	AED [m]↓	mIoU [%]↑	RMSE [m]↓	AED [m]↓	mIoU [%]↑
(A-i) Qwen3-VL [2]		✓	25.73 ± 0.13	21.04 ± 0.13	0.42 ± 0.08	20.26 ± 0.08	16.05 ± 0.06	0.16 ± 0.02
(A-ii) Qwen3-VL [2]	✓		27.05 ± 0.94	23.00 ± 1.23	0.42 ± 0.02	18.25 ± 0.07	13.79 ± 0.06	0.76 ± 0.08
(A-iii) GPT-5.2 [40]		✓	14.46 ± 0.23	10.60 ± 0.20	4.51 ± 0.55	17.45 ± 0.15	13.73 ± 0.14	1.06 ± 0.06
(A-iv) GPT-5.2 [40]	✓		12.44 ± 0.15	9.07 ± 0.13	6.26 ± 0.16	11.29 ± 0.24	7.91 ± 0.08	3.22 ± 0.17
(A-v) AiSCREAM		✓	7.70 ± 0.23	6.50 ± 0.25	4.91 ± 0.91	6.93 ± 0.14	5.70 ± 0.11	2.25 ± 0.54
(A-vi) AiSCREAM	✓		6.60 ± 0.10	5.61 ± 0.10	8.24 ± 0.72	5.57 ± 0.13	4.47 ± 0.16	3.35 ± 0.35
(A-vii) AiSCREAM	✓	✓	**4.95 ± 0.08**	**4.18 ± 0.06**	**10.48 ± 0.37**	**5.19 ± 0.07**	**4.29 ± 0.08**	**5.90 ± 0.58**

removing the AiS view segmentation term (Model (b)) resulted in RMSE increases of 0.22 m on GRiN-Drive and 0.05 m on DRAMATiST. The removal of the obstacle penalty term (Model (c)) led to RMSE increases of 0.32 m and 0.05 m on the same benchmarks, respectively. These results support the use of both HC loss components for cross-view consistency and reducing the error.

View-input ablation. To examine whether the performance gap between AiSCREAM and the baseline methods is due to the AiS view images alone, we conducted another ablation study. In this experiment, we replaced the input of the baseline methods from a front view image to an AiS view image.

Table 4 shows the quantitative results of an additional ablation study on GRiN-Drive and DRAMATiST, where we evaluate the contribution of AiS view and front view inputs by evaluating AiSCREAM and representative MLLM baselines using either the synthesized AiS view image or the original front view image. For AiSCREAM , we used the initial-stage results for a fair comparison. The best score for each metric is in **bold**. The reported values are the mean and standard deviation over five trials.

Table 4 shows that, for AiSCREAM, using only the AiS view (Model (A-v)) yields RMSEs of 7.70 m on GRiN-Drive and 6.93 m on DRAMATiST, whereas using only the front view (Model (A-vi)) yields lower RMSEs of 6.60 m and 5.57 m, respectively. Similarly, for GPT-5.2 and Qwen3-VL, simply using the AiS view image as input did not yield improvements sufficient to outperform AiSCREAM. These results indicate that simply replacing the front view with the AiS view is insufficient, and highlight the importance of cross-view geometric semantic alignment for using AiS view.

Table 5: Quantitative comparison with BEV-based alternatives.

BEV variants	GRiN-Drive			DRAMATiST		
	RMSE [m]↓	AED [m]↓	mIoU [%]↑	RMSE [m]↓	AED [m]↓	mIoU [%]↑
GaussianLSS [34]	6.48	5.47	8.99	5.47	4.45	3.25
BEVFormer v2 [58]	6.74	5.72	8.51	5.71	4.64	3.17
AiS view (ours)	**4.72**	**3.95**	**11.89**	**5.00**	**4.03**	**6.16**

BEV variants ablation: To verify whether AiS view provides an advantage over conventional BEV-style representations, we added direct comparisons with two BEV-based variants, in which the AiS view is replaced with BEV representations obtained from GaussianLSS [34] and BEVFormer v2 [58]. As shown in Table 5, AiS view achieved lower RMSE than the BEV variants on both benchmarks. These results support the advantage that the proposed generated-aerial-view formulation truly offers a clear advantage over BEV for this task.

AiS views quality and sensitivity: Since ground truth is unavailable, we cannot directly evaluate their quality. Instead, we assess sensitivity by injecting Gaussian noise into AiS, where σ denotes the standard deviation in normalized pixel space. As shown in Fig. 5, the localization error increases gradually, and our method remains better than the w/o AiS variant in Table 2 (RMSE = 6.60).

Moreover, Table 1 shows that even w/o AiS outperforms GPT-5.2 (RMSE = 12.44) and GENNAV (RMSE = 10.51). These results indicate that the framework benefits from AiS views, but is not overly dependent on high-fidelity AiS generation.

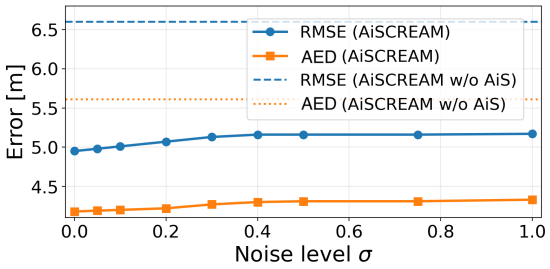

Fig. 5: Sensitivity to AiS degradation

5 Conclusion

In this paper, we studied the RNPL task, which predicts an absolute target position from a language instruction and a front camera image. We introduced a cross-view geometric semantic alignment with front view and the Aerial Semantic view. We further incorporated MLLM Reshape Localizer that performs prediction-conditioned refinement using a visual prompt with an overlaid prediction and an explicit numeric cue. We incorporated the Height-based Cross-view loss and combined it with a front segmentation loss across the three prediction heads. In addition, we constructed DRAMATiST, a benchmark of front view images and navigation instructions that increases road and traffic diversity beyond existing datasets. Our approach outperformed the baseline methods on both GRiN-Drive and DRAMATiST under standard evaluation metrics. Future work will include the acceleration of AiS view synthesis via lighter and faster generation models (e.g., distillation and low-step inference) and a redesign of the pipeline to avoid explicit generation at inference time.

References

1. Arai, H., Miwa, K., Sasaki, K., et al.: CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving. In: WACV. pp. 1933–1943 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., et al.: Qwen3-VL Technical Report. arXiv:2511.21631 (2025)
3. Caesar, H., Bankiti, V., Lang, A., Vora, S., Liong, V., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving. In: CVPR. pp. 11621–11631 (2020)
4. Cao, X., Zhou, T., Ma, Y., Ye, W., Cui, C., Tang, K., Cao, Z., Liang, K., et al.: MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding. In: CVPR. pp. 21819–21830 (2024)
5. Carion, N., Gustafson, L., Hu, Y., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., et al.: SAM 3: Segment Anything with Concepts. In: ICLR (2026)
6. Chen, Y., Li, W., et al.: SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In: ECCV. pp. 323–340 (2024)
7. Cheng, Z., Li, K., Jin, P., Li, S., Ji, X., Yuan, L., Liu, C., Chen, J.: Parallel Vertex Diffusion for Unified Visual Grounding. In: AAAI. pp. 1326–1334 (2024)
8. Cho, H., Ivanovic, B., Cao, Y., Schmerling, E., Wang, Y., Weng, X., et al.: Language-Image Models with 3D Understanding. In: ICLR (2025)
9. Cui, C., Ma, Y., Cao, X., Ye, W., et al.: A Survey on Multimodal Large Language Models for Autonomous Driving. In: WACV. pp. 958–979 (2024)
10. Cui, Y., Lin, H., Yang, S., et al.: Chain-of-Thought for Autonomous Driving: A Comprehensive Survey and Future Prospects. arXiv:2505.20223 (2025)
11. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al.: MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs. In: ICCV. pp. 7395–7408 (2025)
12. Deruyttere, T., Grujicic, D., et al.: Talk2car: Predicting Physical Trajectories for Natural Language Commands. IEEE Access **10**, 123809–123834 (2022)
13. Deruyttere, T., Vandenhende, S., Grujicic, D., Gool, L., Moens, M.: Talk2Car: Taking Control of Your Self-Driving Car. In: EMNLP. pp. 2088–2098 (2019)
14. Drews, F., Feng, D., et al.: DeepFusion: A Robust and Modular 3D Object Detector for Lidars, Cameras and Radars. In: IROS. pp. 560–567 (2022)
15. Du, Y., Lei, C., Zhao, Z., Su, F.: iKUN: Speak to Trackers without Retraining. In: CVPR. pp. 19135–19144 (2024)
16. Etchegaray, D., Huang, Z., et al.: Find n' Propagate: Open-Vocabulary 3D Object Detection in Urban Environments. In: ECCV. pp. 133–151 (2024)
17. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., et al.: ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. In: ICCV. pp. 24823–24834 (2025)
18. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: CVPR. pp. 3354–3361 (2012)
19. Google DeepMind: Nanobanana: Gemini 2.5 flash image model (2025), `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`, accessed: 2026-02-20
20. Grujicic, D., et al.: Predicting Physical World Destinations for Commands Given to Self-driving Cars. In: AAAI. pp. 715–725 (2022)
21. Guan, R., Zhang, R., Ouyang, N., Liu, J., Man, K., Cai, X., Xu, M., Smith, J., Lim, E., et al.: Talk2Radar: Bridging Natural Language with 4D mmWave Radar for 3D Referring Expression Comprehension. In: ICRA. pp. 10884–10891 (2025)

22. Guo, K., Huang, Y., Sun, S., Song, X., Feng, M., Liu, Z., Song, H., Wang, T., Li, J., Akhtar, N., Mian, S.: Beyond Human Perception: Understanding Multi-Object World from Monocular View. In: CVPR. pp. 3751–3760 (2025)
23. Hosomi, N., Hatanaka, S., Iioka, Y., Yang, W., Kuyo, K., Misu, T., Yamada, K., Sugiura, K.: Trimodal Navigable Region Segmentation Model: Grounding Navigation Instructions in Urban Areas. IEEE RA-L **9**(5), 4162–4169 (2024)
24. Hosomi, N., Iioka, Y., Hatanaka, S., Misu, T., Yamada, K., Tsukamoto, N., Kobayashi, S., Sugiura, K.: Multimodal Target Localization With Landmark-Aware Positioning for Urban Mobility. IEEE RA-L **10**(1), 716–723 (2025)
25. Huang, R., Zheng, H., et al.: Training an Open-Vocabulary Monocular 3D Object Detection Model without 3D Data. In: NeurIPS. pp. 72145–72169 (2024)
26. Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., et al.: A Survey on Vision-Language-Action Models for Autonomous Driving. In: ICCV. pp. 4524–4536 (2025)
27. Kamath, A., Singh, M., LeCun, Y., et al.: MDETR-Modulated Detection for End-to-end Multi-modal Understanding. In: ICCV. pp. 1780–1790 (2021)
28. Katsumata, K., Iioka, Y., Hosomi, N., Misu, T., Yamada, K., Sugiura, K.: GEN-NAV: Polygon Mask Generation for Generalized Referring Navigable Regions. In: CoRL. pp. 5195–5217 (2025)
29. Lee, P., Je, J., et al.: Perspective-Aware Reasoning in Vision-Language Models via Mental Imagery Simulation. In: ICCV. pp. 9241–9251 (2025)
30. Lin, H., Chen, S., Liew, J., Chen, D., Li, Z., Zhao, Y., et al.: Depth Anything 3: Recovering the Visual Space from Any Views. In: ICLR (2026)
31. Liu, C., Ding, H., Jiang, X.: GRES: Generalized Referring Expression Segmentation. In: CVPR. pp. 23592–23601 (2023)
32. Liu, J., Ding, H., Cai, Z., et al.: PolyFormer: Referring Image Segmentation as Sequential Polygon Generation. In: CVPR. pp. 18653–18663 (2023)
33. Liu, Z., Tang, H., Amini, A., Yang, X., et al.: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In: ICRA (2023)
34. Lu, S., Tsai, H., Chen, T.: Toward Real-world BEV Perception: Depth Uncertainty Estimation via Gaussian Splatting. In: CVPR. pp. 17124–17133 (2025)
35. Ma, Y., Cui, C., Cao, X., Ye, W., Liu, P., Lu, J., Abdelraouf, A., et al.: LaMPilot: An Open Benchmark Dataset for Autonomous Driving with Language Model Programs. In: CVPR. pp. 15141–15151 (2024)
36. Malla, S., Choi, C., Dwivedi, I., Choi, H., Li, J.: DRAMA: Joint Risk Localization and Captioning in Driving. In: WACV. pp. 1043–1052 (2023)
37. Nguyen, P., Quach, K., Kitani, K., Luu, K.: Type-to-Track: Retrieve Any Object via Prompt-based Tracking. In: NeurIPS. pp. 3205–3219 (2023)
38. Nie, M., Peng, R., Wang, C., et al.: Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving. In: ECCV. pp. 292–308 (2024)
39. Nishimura, T., Kuyo, K., Kambara, M., Sugiura, K.: Object Segmentation from Open-Vocabulary Manipulation Instructions Based on Optimal Transport Polygon Matching with Multimodal Foundation Models. In: IROS. pp. 9549–9556 (2024)
40. OpenAI: Introducing GPT-5.2. `https://openai.com/index/introducing-gpt-5-2/` (2025), accessed: 2026-06-20
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 (2023)
42. Paul, P., et al.: LeGo-Drive: Language-enhanced Goal-oriented Closed-Loop End-to-End Autonomous Driving. In: IROS. pp. 10020–10026 (2024)
43. Renz, K., Chen, L., et al.: SimLingo: Vision-only Closed-Loop Autonomous Driving with Language-Action Alignment. In: CVPR. pp. 11993–12003 (2025)

44. Rufus, N., Jain, K., Nair, K., et al.: Grounding Linguistic Commands to Navigable Regions. In: IROS. pp. 8593–8600 (2021)
45. Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Kochenderfer, M., Choi, C., Dariush, B.: Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In: WACV. pp. 7513–7522 (2024)
46. Shao, H., Hu, Y., Wang, L., et al.: LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In: CVPR. pp. 15120–15130 (2024)
47. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., et al.: DriveLM: Driving with Graph Visual Question Answering. In: ECCV. pp. 256–274 (2024)
48. Siméoni, O., Vo, H., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., et al.: DINOv3. arXiv:2508.10104 (2025)
49. Sun, P., Kretzschmar, H., Dotiwalla, X., et al.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In: CVPR. pp. 2446–2454 (2020)
50. Tang, K., Cao, X., Cao, Z., Zhou, T., Li, E., Liu, A., Zou, S., Liu, C., Mei, S., Sizikova, E., et al.: THMA: Tencent HD Map AI System for Creating HD Map Annotations. In: AAAI. pp. 15585–15593 (2023)
51. Wang, S., Caesar, H., Nan, L., Kooij, J.: UniBEV: Multi-modal 3D Object Detection with Uniform BEV Encoders for Robustness against Missing Sensor Modalities. In: IEEE IV. pp. 2776–2783 (2024)
52. Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., et al.: Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. arXiv:2511.00088 (2025)
53. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., et al.: Qwen-Image Technical Report. arXiv:2508.02324 (2025)
54. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring Multi-Object Tracking. In: CVPR. pp. 14633–14642 (2023)
55. Xia, Z., Han, D., Han, Y., Pan, X., et al.: GSVA: Generalized Segmentation via Multimodal Large Language Models. In: CVPR. pp. 3858–3869 (2024)
56. Xu, Z., Bai, Y., Zhang, Y., Li, Z., Xia, F., Wong, K., Wang, J., Zhao, H.: DriveGPT4-V2: Harnessing Large Language Model Capabilities for Enhanced Closed-Loop Autonomous Driving. In: CVPR. pp. 17261–17270 (2025)
57. Xu, Z., Zhang, Y., et al.: DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. IEEE RA-L **9**(10), 8186–8193 (2024)
58. Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., et al.: BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. In: CVPR. pp. 17830–17839 (2023)
59. Yang, Z., Jia, X., Li, H., Yan, J.: LLM4Drive: A Survey of Large Language Models for Autonomous Driving. In: NeurIPS Workshop (2024)
60. Yu, F., Chen, H., Wang, X., Xian, W., et al.: BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In: CVPR. pp. 2636–2645 (2020)
61. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: MOTR: End-to-End Multiple-Object Tracking with Transformer. In: ECCV. pp. 659–675 (2022)
62. Zhang, H., Jiang, H., Yao, Q., Sun, Y., Zhang, R., Zhao, H., Li, H., Zhu, H., Yang, Z.: Detect Anything 3D in the Wild. In: ICCV. pp. 5048–5059 (2025)
63. Zhang, Y., Wu, D., Han, W., Dong, X.: Bootstrapping Referring Multi-Object Tracking. arXiv:2406.05039 (2024)
64. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176 (2025)
65. Zhang, Y., Wang, C., Wang, X., et al.: FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. IJCV **129**(11), 3069–3087 (2021)

66. Zhou, X., Liu, M., Yurtsever, E., Zagar, B., et al.: Vision Language Models in Autonomous Driving: A Survey and Outlook. IEEE T-IV pp. 1–20 (2024)
67. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., et al.: SeqTR: A Simple yet Universal Network for Visual Grounding. In: ECCV. pp. 598–615 (2022)
68. Zhu, L., et al.: POPEN: Preference-based Optimization and Ensemble for LVLM-based Reasoning Segmentation. In: CVPR. pp. 30231–30240 (2025)
69. Zhu, Y., Wang, S., Zhong, W., Shen, N., Li, Y., Wang, S., et al.: A Survey on Large Language Model-empowered Autonomous Driving. arXiv:2409.14165 (2024)