BEVLM: Distilling Semantic Knowledge from LLMs into Bird's-Eye View Representations

Thomas Monninger[1,*], Shaoyuan Xie[2,*,†], Qi Alfred Chen[2], and Sihao Ding[1,✉]

[1] Mercedes-Benz Research & Development North America, San Jose, USA
[2] University of California, Irvine, USA
{thomas.monninger, sihao.ding}@mercedes-benz.com
{shaoyux, alfchen}@uci.edu

Abstract. The integration of Large Language Models (LLMs) into autonomous driving has attracted growing interest for their strong reasoning and semantic understanding abilities, which are essential for handling complex decision-making and long-tail scenarios. However, existing methods typically feed LLMs with tokens from multi-view and multi-frame images independently, leading to redundant computation and limited spatial consistency. This separation in visual processing hinders accurate 3D spatial reasoning and fails to maintain geometric coherence across views. On the other hand, Bird's-Eye View (BEV) representations learned from geometrically annotated tasks (*e.g.*, object detection) provide spatial structure but lack the semantic richness of foundation vision encoders. To bridge this gap, we propose **BEVLM**, a framework that connects a spatially consistent and semantically distilled BEV representation with LLMs. Through extensive experiments, we show that BEVLM enables LLMs to reason more effectively in cross-view driving scenes, improving accuracy by 46.0 %, by leveraging BEV features as unified inputs. Furthermore, by distilling semantic knowledge from LLMs into BEV representations, BEVLM significantly improves closed-loop end-to-end driving performance in safety-critical scenarios across UniAD and VAD, with gains of up to 28.2 %.

Keywords: Autonomous Driving · Large Language Models · Bird's-Eye View

1 Introduction

Large Language Models (LLMs) have rapidly advanced in scene understanding and reasoning [1, 10, 62, 70], attracting growing interest for autonomous driving applications [27, 42, 48, 52, 56, 60]. Integrating language foundation models into autonomous driving systems provides a path towards commonsense reasoning, open-world understanding, and enhanced interpretability, capabilities often

(∗) Equal contribution, the order was determined alphabetically.
(†) Work was done during an internship at Mercedes-Benz Research & Development North America.
(✉) Corresponding author.

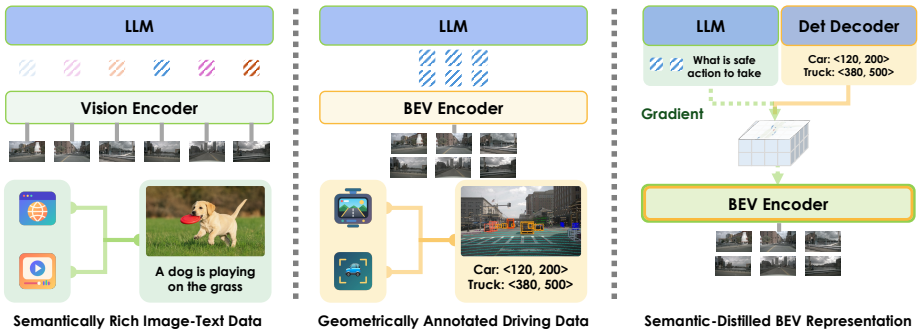


Fig. 1: Representation Comparison: Left: Vision encoders can leverage widely available semantically rich image-text data, but process multi-view images independently. Center: Bird's-Eye View (BEV) encoders provide a spatially consistent scene representation, but are limited to geometrically annotated data. Right (ours): We propose the semantic distillation from LLMs to BEV encoders to build a semantic-enhanced and spatially consistent scene representation.

lacking in conventional perception or end-to-end driving pipelines [64]. Such reasoning ability is particularly critical for handling complex driving scenarios and corner cases, *i.e.*, the "long tail" of the distribution.

To integrate LLMs into autonomous driving, most existing systems leverage Vision Language Models (VLMs) and extract visual tokens independently from multi-view and multi-frame images [25,27,45,47–49,53,65,69]. While this design is straightforward and leverages the large-scale pre-training of VLMs to align vision and language modalities, it introduces two key limitations. As shown in Fig. 1, first, the resulting representations capture each view angle individually and are encoded into different token chunks. This separate processing fails to model spatial consistency, which is crucial for modeling dynamic driving environments [37] and poses a challenge for 3D spatial reasoning of LLMs [17]. Second, this design fails to exploit the temporal correlation, and separate processing makes the computational cost grow proportionally with the number of frames, leading to an inevitable trade-off between capturing long-term temporal information and maintaining computational efficiency [67].

Meanwhile, the Bird's-Eye View (BEV) representation has become a cornerstone of modern autonomous driving systems [37]. BEV provides a unified top-down view of the 3D environment by fusing information from multiple viewpoints, time steps, and even sensor modalities into a compact and spatially consistent grid. This representation enables more effective reasoning about spatio-temporal relationships among the ego-vehicle, dynamic agents, and static surroundings, which is crucial for reliable scene understanding. Owing to these advantages, the BEV grid has become the de facto intermediate representation for object detection [30,35], motion prediction [55,66], and vehicle planning [22,28].

However, despite its compactness and spatial consistency, the BEV representation cannot be pre-trained at scale using semantically rich image–text datasets,

as is possible with foundation visual encoders in VLMs [1, 10, 62, 70]. Large-scale pre-training is essential for learning transferable visual features that generalize to rare and open-world driving scenarios [44]. The lack of such semantic richness forms a fundamental bottleneck, preventing BEV-based representations from being adopted by the advances of LLMs.

In this paper, we conduct the first rigorous experiments showing the advantages of the spatially consistent BEV representation for LLM reasoning in autonomous driving, which we call **Bird's-Eye View Language Model (BEVLM)**. BEV improves scene understanding accuracy by 46.0 % over multi-view inputs and matches foundation vision encoders that are 10× larger, suggesting BEV is a superior scene representation for LLMs. Building on this, we use the BEVLM framework for semantic distillation, framing the LLM as a teacher that supervises the BEV encoder (student) via VQA tasks. This yields a semantic-aware BEV encoder that interacts effectively with language models while preserving spatial structure. Plugging it into two representative end-to-end driving frameworks, UniAD and VAD, both show consistent closed-loop gains in safety-critical scenarios (up to 28.2 % higher safety score and 16.1 % lower collision rate), confirming that the benefits generalize across BEV-based architectures. The project website is at `https://sites.google.com/view/secure-safe-ai/bevlm`.

Our contributions are summarized as follows:

1. We are the first to carry out a representation study that compares individual multi-frame multi-view perspective images and joint BEV representations for LLM spatial reasoning in autonomous driving.
2. We propose BEVLM, a framework to distill semantic information from LLMs into the BEV encoder while preserving the spatial BEV representation.
3. We train an end-to-end driving model from the distilled BEV encoder and find significant improvements in closed-loop evaluation, confirming distillation performance specifically in safety-critical scenarios.

2 Related Work

LLMs for Autonomous Driving. With the rapid advancement of LLMs, there is a growing interest in applying their reasoning and knowledge capabilities to autonomous driving. The key motivation is to leverage the human knowledge and commonsense reasoning embedded in LLMs to better handle corner cases and long-tail scenarios [64]. Existing studies generally follow two directions: (1) using LLM-generated text as high-level guidance for BEV-based end-to-end driving pipelines [15, 27, 42, 49, 52, 60]; and (2) directly generating driving trajectories through LLMs [8, 16, 18, 24, 25, 47, 48, 59, 61, 68, 69]. However, most of these approaches still follow the conventional VLM paradigm, where visual features are extracted independently from individual camera views and frames. This design limits the LLM's ability to capture spatio-temporal consistency and geometric relationships across views. Recent works [4, 54, 67] begin to explore connecting BEV and language modalities. Yet, a systematic study comparing the representational advantages of image-based versus BEV-based inputs for LLM reasoning

has been missing. Also, solutions for addressing the semantic gap between these two representations are still underexplored.

BEV Representation. The BEV representation combines and integrates information from multiple views, time steps, and even sensor modalities [37, 57]. It has become a central intermediate representation for full-stack autonomous driving, enabling perception [23, 30, 35, 40, 66], prediction [14, 22, 66], and planning [22, 28]. However, learning semantically rich BEV representations remains an open challenge [38, 63]. Current BEV learning methods rely heavily on dense geometric supervision, often through object detection [30, 35], map construction [32, 39], or joint end-to-end training [22, 28]. While such supervision provides strong geometric cues, it limits the semantic richness needed for understanding complex, safety-critical scenarios, which is an essential requirement for conditionally, highly, and fully automated driving systems.

Safety-Critical Evaluation. Autonomous driving is inherently a safety-critical task, where unsafe decisions can result in severe consequences [41, 51]. Most existing studies emphasize the robustness of the perception module, particularly under out-of-distribution inputs [29, 57] or adversarial perturbations [7, 46, 58]. In contrast, the safety of the planning module has received relatively limited attention. Recently, several benchmarks have been introduced to assess planning safety [12, 26, 36]. For instance, the NeuroNCAP benchmark [36] generates safety-critical driving scenarios through closed-loop simulation. In this work, we focus on enhancing the semantic understanding of the BEV representation to promote safer decision-making in end-to-end autonomous driving systems.

3 BEV Representation for Spatial Reasoning

To understand how BEV representations affect spatial understanding, this section examines whether LLMs can effectively interpret and reason over BEV inputs, and then compares BEV and image-based representations in supporting spatial reasoning. The results of this study motivate Sec. 4, where we enhance BEV representations with semantic knowledge from LLMs.

3.1 BEV-to-Language Alignment

***Research Question 1**: Can BEV representations be aligned with the language space such that LLMs can reason over them as effectively as task-specific BEV perception modules?*

In this section, we examine whether BEV features can be fully aligned with the language space for LLM-based reasoning. We mainly compare two setups: (1) our BEVLM framework, which uses an LLM receiving BEV features through a learned projector [34], and (2) a task-specific BEV model whose outputs are converted to answers via rule-based logic. By comparing their accuracy in identifying objects in the scene, we can assess whether the BEV-to-language projection

Table 1: BEV Projector Alignment Study. We compare the performance between the BEV-based detected bounding box and the LLM that directly operates on the same BEV grid. We report binary classification accuracy. *Majority class* denotes using the majority choices from the training set for each category. The final score is averaged across all 10 categories, as we only show the top 5 frequent objects. For full results, please refer to the Appendix B.

Model	Modality	cars	peds.	trucks	cones	barriers	**Avg.**
Majority class	-	92.7	81.9	65.0	59.6	54.8	78.2
Linear probe	-	94.6	87.2	84.2	83.8	83.9	88.7
$\text{Detection}_{\text{UniAD}}$	-	91.1	90.9	86.7	**99.0**	**99.6**	92.8
$\text{InternVL3}_{\text{1B}}$	B_{UniAD}	94.7	88.9	85.8	90.1	88.6	90.8
$\text{InternVL3}_{\text{8B}}$	B_{UniAD}	97.7	**94.8**	**89.7**	94.0	95.0	**95.3**
$\text{DeepSeek-VL}_{\text{1B}}$	B_{UniAD}	**95.1**	92.0	84.9	92.6	92.2	92.2

retains sufficient spatial information for the LLM to reason over as effectively as the task-specific decoder (e.g., detection head).

Experimental Setups. We assess whether the projector effectively aligns BEV features with the language space by comparing three results on binary *object-existence* questions from DriveLM-nuScenes dataset [48]. We train the projector on *perception-related* question–answer pairs and evaluate on questions such as *"Is there a moving car in the front left?"*. We adopt the BEV encoder from UniAD [22] and the language component of InternVL3 [70]. The BEV grid is max-pooled to 50×50, producing 2,500 tokens. The original vision encoder is disabled when BEV tokens are used. We choose the language component of a VLM over a pure text-based LLM since we anticipate better spatial understanding from the cross-modal alignment. Additionally, this setup enables us to conduct comparisons across different modalities (Sec. 3.2). Data examples are provided in Appendix C.1. More implementation details can be found in Appendix A.

Results. Tab. 1 reports both overall and class-wise accuracy for the five most frequent object categories. We compare BEVLM against three baselines: (1) a *majority class* prior based on DriveLM-nuScenes training split [48, 56]; (2) a stronger linear probe baseline, implemented as an object-specific linear classifier on max-pooled BEV features of the view specified in the question, to consider the influence of data bias on the learning process; and (3) UniAD's detection head, where answers are derived by matching *class*, *moving state*, and *view angle* to the detector outputs. BEVLM substantially outperforms the majority-class and linear probe baselines across all categories, with especially strong gains on more balanced classes such as trucks, cones, and barriers (near 90 %). This indicates that the projector is not simply capturing dataset priors but preserving meaningful spatial and semantic cues for LLMs to answer correctly.

BEVLM also approaches the accuracy of UniAD's task-specific decoder. Specifically, $\text{InternVL3}_{\text{1B}}$ achieves 90.8 % and $\text{DeepSeek-VL}_{\text{1B}}$ achieves 92.2 %, versus UniAD's 92.8 %. Scaling the InternVL3 to 8B yields 95.3 %, surpassing the detector itself. These results show that a simple MLP projector can effectively map BEV features into the language space. Complete results are in Appendix B.

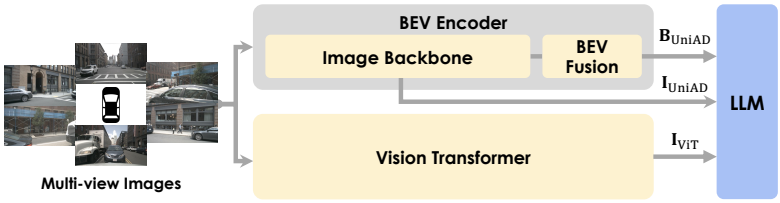


Fig. 2: Representation Study. We compare between (1) I_{ViT}, visual tokens extracted from the Vision Transformer of the original VLM; (2) I_{UniAD}, visual tokens from the backbone before the BEV fusion; and (3) B_{UniAD}, BEV tokens produced from the same backbone after the BEV fusion. The input language question is the same, but not visualized here for simplicity.

3.2 Comparative Study of Visual Representations

Research Question 2: *Are BEV representations more effective than perspective-view inputs for enabling LLMs to perform spatial reasoning?*

Building on the successful BEV-to-language alignment, we next compare different visual representations to examine which one better supports spatial reasoning. Specifically, we evaluate three setups, shown in Fig. 2: (1) I_{ViT}, visual tokens from the ViT of the original VLM [70]; (2) I_{UniAD}, tokens from UniAD's image backbone before BEV fusion; and (3) B_{UniAD}, BEV tokens produced after fusing multi-view features into BEV space [22]. This analysis allows us to isolate the advantages of the BEV representation over conventional image-based inputs in capturing geometric and spatial relationships within driving scenes. More implementation details can be found in Appendix A.

Single-View Reasoning. We start with DriveLM *object-existence* questions, which are designed for single-view reasoning. The quantitative results are presented in Tab. 2a. We first observe that BEV representations show advantages over single-view inputs. Additionally, increasing the size of the LLM from 1B to 8B gives a substantial improvement from 89.8 % to 94.5 % accuracy, showing that a more powerful LLM can better utilize relevant information from the token space. The 1B LLM processing tokens from B_{UniAD} achieves 90.8 % accuracy, which outperforms both I_{ViT} (90.3 %) and I_{UniAD} (89.8 %). This confirms the hypothesized benefits of the BEV representation. Also here, using the larger 8B LLM yields a substantial improvement to 95.3 % accuracy, outperforming the equivalent model on I_{UniAD}. Similarly, DeepSeek-VL_{1B} shows the same trend where the B_{UniAD} surpasses the I_{ViT} and I_{UniAD}.

Therefore, even in the context of single-view reasoning tasks, our empirical findings indicate that the BEV representation exhibits a consistent advantage over perspective-view counterparts. This observation motivates a further investigation into the representation gap when subjected to the demands of complex, cross-view spatial reasoning.

Table 2: DriveLM Results and Generality Analysis. We report accuracy, precision, recall, and F-1 scores. "**I**" denotes image-space, while "**B**" denotes BEV-space. "Proj." indicates training the MLP projector. Metrics are reported as percentages.

(a) InternVL3 Results

Models	Mod.	Proj.	**Acc.↑**	**Prec.↑**	**Recall.↑**	**F1↑**
InternVL3_{1B}	I_{ViT}		74.2	72.2	97.7	83.0
InternVL3_{1B}	I_{ViT}	✓	90.3	90.0	95.5	92.7
InternVL3_{1B}	I_{UniAD}	✓	89.8	90.1	94.6	92.3
InternVL3_{8B}	I_{UniAD}	✓	94.5	94.6	97.0	95.8
InternVL3_{1B}	B_{UniAD}	✓	90.8	90.3	96.1	93.1
InternVL3_{8B}	B_{UniAD}	✓	95.3	95.2	97.7	96.4

(b) DeepSeek-VL Results

Models	Mod.	Proj.	**Acc.↑**	**Prec.↑**	**Recall.↑**	**F1↑**
DS-VL_{1B}	I_{ViT}		61.5	84.1	49.8	62.6
DS-VL_{1B}	I_{ViT}	✓	85.3	86.5	91.5	89.0
DS-VL_{1B}	I_{UniAD}	✓	90.4	89.9	97.3	93.5
DS-VL_{1B}	B_{UniAD}	✓	92.2	90.5	98.2	94.2

Table 3: Ego3D Results. We evaluate cross-view reasoning question types (i.e., "*object-centric*" in Ego3D) and report accuracy for Multiple-Choice-Question (MCQ) and L1 error for numerical distance estimation. We use the InternVL3_{1B} as the LLM. "**Enc. Size**" and "**#Token**" denote the visual encoder size and the number of visual tokens, respectively. Token number details can be found in Appendix A.3.

			MCQ Accuracy (%) ↑				**L1 (m) ↓**
Modality	**Enc. Size**	**#Token**	Loc.	Abs-Dist.	Rel-Dist.	**Avg.**	Dist
I_{ViT}	400M	4,608	13.95	23.81	52.94	28.57	14.15
I_{ViT} w/ ft.	400M	4,608	<u>60.47</u>	**50.00**	**79.41**	**62.19**	<u>7.42</u>
I_{UniAD}	40M	2,250	39.53	26.19	64.71	42.02	9.01
B_{UniAD}	44M	2,500	**67.44**	<u>45.24</u>	<u>73.53</u>	<u>61.34</u>	**7.05**

Cross-View Reasoning. The advantage margin of BEV representation is limited in the DriveLM dataset mainly due to the single-view reasoning questions. Therefore, we show the results on the cross-view reasoning Ego3D dataset [17]. Specifically, we focus on the *object-centric* questions, which ask about the spatial relationship between two objects in different ego camera views. Therefore, the LLMs are required to reason over the entire driving scene beyond a single view as in DriveLM perception data. As shown in Tab. 3, we find the BEV representation surpasses the perspective-view representation by a noticeable margin. Specifically, the MCQ accuracy is improved by 46.0 %, and the L1 error is decreased by 21.8 %. Additionally, B_{UniAD} shows comparable performance to I_{ViT} w/ ft. even though the ViT encoder is 10× larger. The results further confirm the advantages of BEV representation for complex spatial reasoning. Data examples are provided in Appendix C.2.

One concern is that the BEV fusion introduces extra parameters (e.g., spatial attention [30]), confounding comparisons between B_{UniAD} and I_{UniAD}. However, we argue this impact is minor as (1) the module size is only 10 % of the backbone [19], and (2) B_{UniAD} shows comparable performance to I_{ViT} despite the ViT model size being an order of magnitude larger. Thus, the gains mainly stem from BEV's representational advantages rather than added capacity. Therefore,

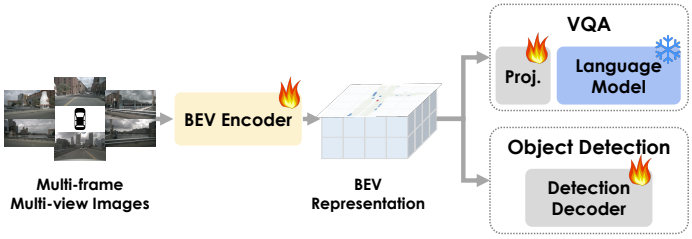


Fig. 3: BEV Semantic Distillation: We distill the knowledge from the language model to the BEV representations by using Visual Question Answering (VQA) tasks while regularizing BEV spatial structure using the original object detection tasks.

we can conclude that BEV representation significantly improves the LLM's spatial reasoning capability, especially for panoramic scene understanding.

4 BEVLM for Semantic Distillation

***Research Question 3**: Can we enhance the semantic awareness of BEV representations, and thereby improve driving safety, by distilling commonsense knowledge from LLMs into BEV encoders?*

The representation study in Sec. 3.2 confirmed that BEV tokens show superior spatial consistency and geometric coherence compared to perspective-view image tokens for enabling better LLM spatial reasoning. However, BEV encoders are typically trained solely on dense geometric annotations (*e.g.*, bounding boxes [30], map elements [32], or ego trajectories [22]), ignoring the broader semantic knowledge and commonsense reasoning essential for safety-critical corner cases. This is the key bottleneck that prevents a wide adoption of BEV representations for LLMs, compared to semantically rich 2D visual representations [47–49, 69].

Therefore, we pioneer the exploration of distilling semantic knowledge from pre-trained LLMs into the BEV encoder. By training the BEV encoder using the BEVLM framework to perform high-level visual question answering (VQA) [2] on safety-related topics, the encoder is *forced* to capture comprehensive, safety-related aspects of the scenario that cannot be covered by geometric supervision alone. This novel semantic distillation process yields a semantically-enhanced BEV representation that is universally beneficial for downstream tasks, including improved performance in LLM-based reasoning [48] and enhanced planning in end-to-end (E2E) driving systems [22, 28].

In this work, we focus on evaluating the benefits of distilled BEV representations on the E2E driving pipeline. Since directly using LLMs for E2E control (*e.g.*, VLAs) remains limited by real-time efficiency [25,69] and reliability [13,56], established E2E pipelines [22, 28] remain the most practical architectures for planning, and demonstrating improvement on them provides the most critical evidence of our method's real-world value. We leave LLM-based control (*e.g.*, VLAs) as future work.

4.1 BEV Distilled from LLM Semantic Space

As illustrated in Fig. 3, we use a shared BEV representation to support both the VQA and object detection tasks. Unlike output-level distillation [21], our method performs representation distillation [20], where the BEV encoder learns to encode semantic cues directly into its feature grid to satisfy the LLM's reasoning requirements. Specifically, we frame the LLM as a fixed semantic teacher that provides supervision signals via VQA tasks. The BEV encoder (student) is distilled to produce features that align with the semantic space learned by the teacher LLMs. For example, when the LLM answers a question such as *"What is the safe action for the ego vehicle to take?"*, the BEV tokens *must* encode safety-relevant scene information.

To maintain the spatial structure of the BEV representation, we jointly train with the original perception tasks, such as object detection, which preserve the geometric structure of the BEV grid. This prevents catastrophic forgetting of spatial relationships that are not directly constrained by the distillation.

Coordinate Conversion. Since BEV tokens require reasoning in an ego-centric spatial frame, we convert the image-plane object locations used by DriveLM [48, 56] into the ego-centric BEV frame. For instance, instead of expressing a location as *"Object appears at (*450 px*,* 360 px*)"*, we reformulate it as *"Object appears 3 meters in front of the ego vehicle and 1.5 meters to the left"*, enabling more intuitive spatial reasoning. We obtain this by projecting the 3D ground-truth boxes onto the image plane and matching them via Intersection-over-Union, retaining only matches above a predefined threshold.

Intuitive Interpretation: BEV as a Semantic Manifold. Unlike standard auxiliary multi-task supervision, where a task-specific head is learned from scratch, we freeze the LLM parameters ϕ so that it acts as a fixed semantic teacher rather than a trainable decoder. The BEV encoder (student) must therefore reshape its feature space to satisfy the LLM's requirements. Concretely, let $\mathcal{X}$ be the input driving scene and $\mathbf{B}_s = E_\theta(\mathcal{X})$ the student BEV feature. For a safety-critical query $\mathbf{q}$, the frozen LLM implicitly requires a specific set of ideal semantic token embeddings $\mathbf{v}^*$, encoding concepts such as "blocked lane" or "unsafe velocity", to produce the correct answer $\mathbf{a}$. The distillation objective forces the student to align its projected features with $\mathbf{v}^*$:

$$\mathcal{L}_{\text{distill}} \approx \|\text{MLP}(E_\theta(\mathcal{X})) - \mathbf{v}^*\|_2^2. \tag{1}$$

Since $\mathbf{v}^*$ cannot be accessed directly, we use the frozen LLM's cross-entropy loss as a differentiable proxy. Therefore, the VQA dataset is not the end goal, but it acts as an *information bottleneck*: by restricting supervision to complex, reasoning-heavy queries, we selectively distill only the high-level semantics that are absent from the BEV encoder's geometric training. Thus, both the teacher LLM (studied in Sec. 5.3) and the VQA data (studied in Sec. 5.5) play a critical role in the proposed framework.

Table 4: Open-Loop and Closed-Loop E2E Planning Results. We compare UniAD and VAD performance using the baseline encoder versus the distilled BEV encoder. Open-loop L2 error is measured on nuScenes [5], while NeuroNCAP score and Collision Rate (CR) are measured on the NeuroNCAP benchmark [36].

		Open-Loop (nuScenes)				NeuroNCAP	
Pipeline	**Method**	L2@1s↓	L2@2s↓	L2@3s↓	**Avg.L2↓**	**Score↑**	**CR↓**
UniAD	Baseline BEV	0.50	0.99	1.67	1.05	2.38±1.52	0.56±0.31
	Rand. aux. head	0.59	1.09	1.79	1.16	1.75±0.96	0.73±0.19
	VLM-AD trans.	0.56	1.02	1.64	1.07	2.02±1.18	0.70±0.23
	VLM-AD CLIP	0.57	1.01	1.64	1.07	2.07±1.13	0.67±0.23
	Distilled BEV_{1B}	**0.46**	**0.91**	**1.55**	**0.97**	2.93±0.69	0.54±0.19
	Distilled BEV_{8B}	0.48	0.94	1.59	1.00	**3.05±1.26**	**0.47±0.30**
VAD	Baseline BEV	0.47	0.79	1.19	0.82	3.24±1.21	0.37±0.24
	Distilled BEV_{8B}	**0.40**	**0.71**	**1.11**	**0.74**	**3.42±0.97**	**0.34±0.20**

4.2 Training

Following our previous setup, we use DriveLM-nuScenes [48] training split as the VQA data for the distillation. We leverage the *full* VQA data, including perception, prediction, behavior, and planning. We also conduct an ablation study for different VQA data types in Sec. 5.5. We start from the pre-trained BEVFormer checkpoints [30] because UniAD [22] and VAD [28] share the same BEV encoder as BEVFormer and are initialized with the same weights. For the baseline model, we freeze the pre-trained BEV encoder and only train the task-specific decoder. To demonstrate the benefits of our distilled BEV for final E2E driving performance, we introduce this semantic distillation as an additional step after object detection pretraining. The distillation is performed for 1 epoch with equal loss term weights for distillation and object detection. After distillation, we perform the same E2E training as in [22, 28], freezing the distilled BEV encoder and training all task-specific heads for 20 epochs (UniAD) or 12 epochs (VAD). Thus, we can study the effects of semantic enhancement on the final E2E tasks. All the experiments are performed on 8 NVIDIA A100 80GB GPUs. Distillation for 1 epoch of the full DriveLM dataset takes ∼35 h for the 1B and 100 h for the 8B LLM. E2E training time of the UniAD remains unchanged with 115 h.

5 Experiments

5.1 Experimental Setups

Following prior works [22, 28], we first evaluate the L2 error on the open-loop nuScenes validation set, and further evaluate on the closed-loop NeuroNCAP benchmark [36]. To account for the stochasticity of the closed-loop simulation, all NeuroNCAP results (score and collision rate) are reported as mean±std over 50 random-seed runs of the NeuroNCAP scenarios.

5.2 Open-Loop Evaluation

The open-loop E2E results are evaluated on the nuScenes validation set [5]. The results on the L2 error metric at 1 s, 2 s, and 3 s are shown in Tab. 4.

On UniAD, both distilled models show a clear improvement over the baseline at all time horizons. The same trend holds on VAD, where the distilled encoder reduces the average L2 from 0.82 to 0.74. Together, these results indicate the potential of BEVLM distillation in reducing the error of imitating human trajectories across both pipelines. However, as shown in prior works [6, 11, 12, 31], open-loop evaluation provides limited signal about real-world performance. Additionally, the original nuScenes dataset includes scenarios where the ego is mostly moving straight [31,56]. Therefore, the improvement in closed-loop safety-critical scenarios is not well understood.

5.3 NeuroNCAP Closed-Loop Evaluation

To study the closed-loop performance in safety-critical scenarios, we further evaluate the distilled model using the NeuroNCAP benchmark [36]. The benchmark simulates safety-critical scenarios and calculates a NeuroNCAP score between 0 and 5 based on the impact velocity, with 5 being no collision. The NeuroNCAP score provides a more fine-grained evaluation beyond the collision rate metric.

The results are shown in Tab. 4. On UniAD, distilling with the 1B LLM raises the NeuroNCAP score substantially from the baseline 2.38 to 2.93 while leaving the collision rate almost unchanged, so the improved score reflects a reduction in crash severity, which we confirm via the pre-collision velocity in Appendix D. Scaling the teacher to 8B further lifts the score to 3.05 (a 28.2 % improvement over the baseline) and lowers the collision rate from 0.56 to 0.47, showing that a stronger teacher transfers richer semantics under the same VQA data. To verify that these gains are not specific to UniAD, we transfer the same 8B-distilled BEV encoder to VAD [28] by freezing it and retraining only the VAD decoder: the NeuroNCAP score improves from 3.24 to 3.42 and the collision rate drops from 0.37 to 0.34. This confirms that semantic distillation generalizes across BEV-based end-to-end pipelines and consistently improves safety in closed-loop, safety-critical scenarios.

We visualize the results between the baseline model and our distilled version (8B) in Fig. 4. The distilled model consistently demonstrates safer and more adaptive behavior in complex scenarios. In the first corner case scenario, the ego vehicle turns right into a lane blocked by an excavator. The baseline model shows issues with understanding the scene and proceeds hesitantly, resulting in a collision with the white vehicle approaching from behind. In contrast, the distilled model anticipates the blockage and performs a swift lane change before the white car approaches. In the second corner case scenario, a white car drives on the wrong side of the road, approaching the ego. The baseline model reacts too late and causes a crash while steering into the opposing road. Our distilled model, however, avoids the collision by swiftly changing into the free lane on the right, showing a safety-oriented understanding of the scene. These results

highlight that semantic distillation from LLMs equips the BEV encoder with improved situational understanding and safety awareness beyond simply imitating human trajectories in common scenarios [31]. More qualitative examples on the NeuroNCAP benchmark are given in Appendix D.

5.4 Comparison with Alternative Semantic Supervision

We compare BEVLM against two alternative supervision designs that use the same DriveLM VQA data: a multi-task learning baseline and a VLM-AD-style distillation [60]. Both are trained on the UniAD pipeline and reported in Tab. 4.

Multi-Task Learning Baseline. The key question is whether the closed-loop improvement comes from the semantics of the pretrained LLM, or simply from attaching an auxiliary VQA head that supervises the BEV encoder with a multi-task objective. To answer this, we replace the frozen LLM with a randomly initialized Transformer decoder of identical architecture (∼13M effectively trainable parameters, matched to BEVLM's ∼13M projector) and train it on the same VQA data (*Rand. aux. head* in Tab. 4). Lacking pretrained semantics, this head merely memorizes answer templates and back-propagates no useful semantic gradient to the BEV encoder. As a result, the NeuroNCAP score *drops* from the baseline 2.38 to 1.75, and the collision rate rises from 0.56 to 0.73. This confirms that the gain originates from the pretrained LLM's semantic latent space, not from the auxiliary VQA head itself.

VLM-AD-Style Distillation. We further compare against VLM-AD [60], a representative method that distills semantics into the BEV encoder from text supervision. Since VLM-AD and its supervision data are not publicly available, we follow its original recipe: the BEV features are supervised against a CLIP text embedding of answers auto-labeled by $\text{InternVL3}_{\text{8B}}$ [70] on DriveLM, using a cosine-similarity loss (*VLM-AD CLIP*). We also consider a setup where we use the same auto-labeled answer with a randomly initialized transformer for the distillation (*VLM-AD trans.*). The results are shown in Tab. 4. These two variants of VLM-AD only reach a NeuroNCAP score of 2.02 and 2.07, respectively, below the baseline 2.38 and far below BEVLM's 3.05, improving only the open-loop L2 at the 3 s horizon. We attribute this gap to two factors. First, the distillation is performed in the CLIP latent space, which is less semantically rich than the LLM's latent space and thus transfers weaker driving-relevant semantics to the BEV encoder. Second, aligning BEV features to a fixed CLIP embedding integrates the BEV representation with the language model less seamlessly than BEVLM's token-level distillation, which produces features that the LLM can natively consume.

5.5 Ablation Studies

BEV Token Downsampling Method. We conduct an ablation study to determine the optimal BEV downsampling method for the projector, with results

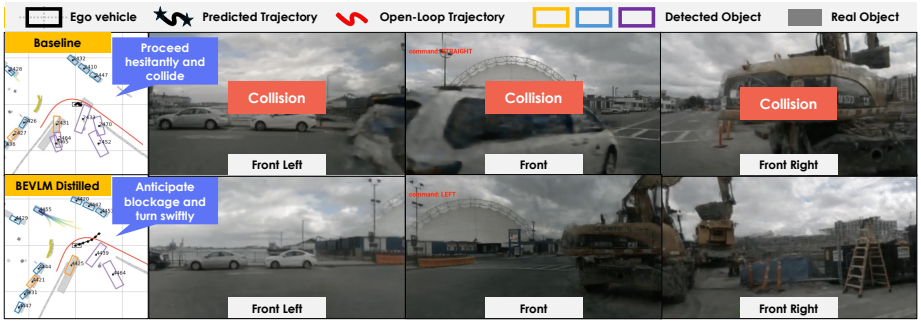
Table 5: Ablation on BEV Token Downsampling. Comparison of pooling and lightweight convolutional methods for BEV downsampling. We evaluate across different tasks and metrics (all scores in %). The metric abbreviations are: **B** = BLEU-4 [43], **R** = ROUGE-L [33], **M** = METEOR [3], **C** = CIDEr [50], **Avg.** = Averaged accuracy. We evaluate "Yes" or "No" questions using Accuracy, while other open-ended questions use language metrics. *Max. Pool* achieves comparable accuracy with no added parameters and is used in all experiments.

Models	Projector	B↑	R↑	M↑	C↑	cars	peds.	trucks	cones	barriers	**Avg.↑**
InternVL3$_{1B}$	Convolution	0.469	**0.686**	**0.610**	3.747	94.4	90.0	84.6	86.9	90.4	91.0
InternVL3$_{1B}$	Depthw. Conv	0.461	0.680	0.604	3.657	94.9	88.2	85.5	89.4	89.3	90.6
InternVL3$_{1B}$	Concat	0.454	0.673	0.597	3.621	93.8	88.2	86.4	89.4	90.0	90.2
InternVL3$_{1B}$	Avg. Pool	0.472	**0.686**	0.609	**3.796**	94.8	89.1	85.2	89.7	90.8	90.9
InternVL3$_{1B}$	Max. Pool	0.468	0.682	0.606	3.765	94.7	88.9	85.8	90.1	88.6	90.8
InternVL3$_{8B}$	Concat	0.468	0.661	0.567	3.521	95.6	91.1	87.3	89.4	91.5	92.4
InternVL3$_{8B}$	Avg. Pool	0.441	0.627	0.559	3.646	97.2	**95.5**	**91.5**	93.6	**96.4**	**95.3**
InternVL3$_{8B}$	Max. Pool	**0.485**	0.673	0.579	3.703	**97.7**	94.8	89.7	**94.0**	95.0	**95.3**

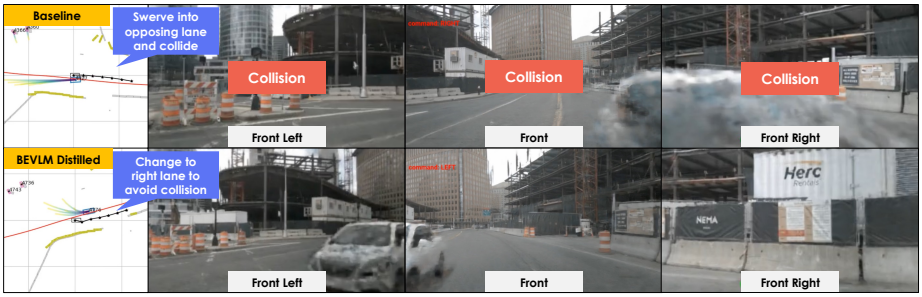
Table 6: VQA Data Ablation. We ablate the performance of the distilled BEV encoder on subsets of the DriveLM dataset. All models have similar collision rates ("CR"), allowing for comparison of the average velocity at impact for collision cases ("Imp. Vel."). The different variants of BEV distillation significantly reduce impact velocities, resulting in higher NeuroNCAP scores. The *Behavior* and *Planning* questions help more than the *Perception* and *Prediction* questions, and the best result is achieved with all questions combined.

Method	Perception	Prediction	Behavior	Planning	L2@1s↓	L2@2s↓	L2@3s↓	avg.L2↓	NeuroNCAP Score↑	CR@0.0s↓	Imp. Vel.↓
Baseline BEV					0.50	0.99	1.67	1.05	2.38±1.52	0.56±0.31	6.11
Distilled$_{1B}$ BEV	✓	✓			0.48	0.95	1.62	1.02	2.77±1.11	0.54±0.25	4.64
Distilled$_{1B}$ BEV			✓	✓	0.48	0.94	1.62	1.01	2.75±1.30	0.54±0.26	**3.97**
Distilled$_{1B}$ BEV	✓	✓	✓	✓	**0.46**	**0.91**	**1.55**	**0.97**	**2.93±0.69**	**0.54±0.19**	4.08

for six variants presented in Tab. 5. Given the objective of a lightweight projector, only simple sampling methods are evaluated, supporting the hypothesis that the BEV tokens are already highly expressive and suitable for direct use. Parameter-free methods include *Avg. Pool*, *Max. Pool*, and a *Concat* operation (equivalent to a pixel unshuffle used in VLMs [70]), which substantially increases the projector's parameter count. Additionally, several learnable methods for downsampling are assessed: a 1-layer standard *Conv*, and a more parameter-efficient *Depthw. Conv* [9]. Same as before, accuracy is measured for existence-based questions. Additionally, we employ various language metrics for a more precise evaluation of the open-ended questions. The results in Tab. 5 demonstrate that while the learnable methods perform well, they offer no significant advantage over simple pooling methods, which are parameter-free and computationally efficient. On the 8B LLM, *Avg. Pool* and *Max. Pool* achieve the same accuracy. We selected *Max. Pool* to create the results in this paper due to slightly better values on the language scores.



(a) Corner Case 1: Right-Turn Conflict with Blocked Lane: In this scenario, the ego vehicle attempts to turn right at an intersection, but the rightmost lane after the turn is blocked by an excavator. Meanwhile, a vehicle traveling straight in the adjacent lane restricts the available space, requiring the ego to adjust its trajectory to avoid a potential collision.



(b) Corner Case 2: Oncoming Car in Ego Lane: In this scenario, the ego vehicle drives on a two-way road with two lanes in each direction. A car from the opposite direction enters the ego vehicle's lane and drives toward it on the wrong side of the road. To avoid a collision, the ego vehicle needs to move to the free right lane in its own direction.

Fig. 4: Qualitative NeuroNCAP Results. Two representative closed-loop planning scenarios are presented for comparison between the baseline and semantically distilled models. The distilled model demonstrates improved decision-making under safety-critical scenarios, successfully performing a *safe right turn* in corner case 1 and an *evasive lane change to the free right lane* in corner case 2 to avoid potential collisions, where the baseline model fails.

Distillation VQA Data. We further conduct an ablation study from the VQA data perspective. Specifically, we study what type of question can best benefit the BEV semantic learning by separating the questions in the DriveLM dataset, including *Perception*, *Prediction*, *Behavior*, and *Planning*. Since the different task data samples vary, we train the less frequent subset (*i.e.*, behavior and planning) for 3 epochs while training the more frequent subset (*i.e.*, perception and prediction) for one epoch to match the overall iterations (shown in Appendix Tab. A.1).

The results are presented in Tab. 6. Both question subsets individually yield a substantial improvement over the baseline with no distillation, reaching comparable NeuroNCAP scores of 2.77 (*Perception* and *Prediction*) and 2.75 (*Behavior* and *Planning*). Combining all four question types achieves the best NeuroNCAP

score of 2.93, indicating that the two subsets provide complementary semantic supervision.

The collision rate is similar for all variants (around 0.54), so the higher NeuroNCAP score indicates that the velocity at impact is much lower for the models trained from distilled BEV encoders, indicating safer and more anticipatory behavior in such corner cases. This is confirmed when comparing the average impact velocity. The UniAD baseline trained without distillation has $6.11\,\mathrm{m\,s^{-1}}$. Training on the *Perception* and *Prediction* subset reduces the average impact velocity to $4.64\,\mathrm{m\,s^{-1}}$, while training on the *Behavior* and *Planning* subset is even more impactful, reaching the lowest value of $3.97\,\mathrm{m\,s^{-1}}$. The variant distilled from all DriveLM questions reaches $4.08\,\mathrm{m\,s^{-1}}$ while attaining the highest overall NeuroNCAP score, showing that both subsets complement each other. More qualitative results together with the corresponding velocity profiles are discussed in Appendix D.3.

6 Discussion & Conclusion

Limitations. In this work, we mainly use the DriveLM-nuScenes [48] dataset for the experiments. We select this dataset due to its (1) compatibility with nuScenes [5] and (2) curated and high-quality ground truth. However, this should not be a major concern, as the distillation from this dataset alone already brings significant improvement. Evaluating the distilled BEV representation on more diverse and semantically rich VQA data to confirm the scaling of the framework is left for future work. More discussion is in Appendix E.

Conclusion. In this work, we presented a framework that unifies the spatial structure of BEV representations with the semantic reasoning capabilities of LLMs. We show that BEV features can be effectively tokenized for LLM processing and that they enable stronger spatial reasoning than image-based tokens in multi-view settings. Building on this insight, we introduced BEVLM, which distills semantic knowledge from LLMs into BEV encoders. This semantic enrichment leads to consistent gains in end-to-end driving, including up to 28.2 % improvement in safety on the NeuroNCAP benchmark. These results highlight the promise of integrating BEV with LLM reasoning to improve the safety and reliability of autonomous driving systems.

Acknowledgements

We thank Mukesh Ghimire and Joona Hellmuth for their help with running the NeuroNCAP closed-loop simulation. This work was primarily supported by Mercedes-Benz Research & Development North America, Inc. Shaoyuan Xie and Qi Alfred Chen were supported in part by (1) the National Science Foundation under grants CNS-2145493 and CNS-2413877, and (2) the U.S. Department of Transportation under Grant 69A3552348327 through the CARMEN+ University Transportation Center.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 Technical Report. arXiv preprint arXiv:2303.08774 (2023)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
3. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Brandstaetter, F., Schuetz, E., Winter, K., Flohr, F.: BEV-LLM: Leveraging Multimodal BEV Maps for Scene Captioning in Autonomous Driving. arXiv Preprint arXiv:2507.19370 (2025)
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A Multimodal Dataset for Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. Cao, W., Hallgarten, M., Li, T., Dauner, D., Gu, X., Wang, C., Miron, Y., Aiello, M., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., Chitta, K.: Pseudo-Simulation for Autonomous Driving. In: Conference on Robot Learning (CoRL) (2025)
7. Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q.A., Liu, M., Li, B.: Invisible for Both Camera and LiDAR: Security of Multi-Sensor Fusion Based Perception in Autonomous Driving Under Physical-World Attacks. In: IEEE Symposium on Security and Privacy (SP). pp. 176–194. IEEE (2021)
8. Chen, Y., Wang, Y., Zhang, Z.: DrivingGPT: Unifying Driving World Modeling and Planning with Multi-Modal Autoregressive Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26890–26900 (2025)
9. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv Preprint arXiv:2507.06261 (2025)
11. Dauner, D., Hallgarten, M., Geiger, A., Chitta, K.: Parting with Misconceptions About Learning-Based Vehicle Motion Planning. In: Conference on Robot Learning. pp. 1268–1281. PMLR (2023)
12. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: NavSim: Data-Driven Non-Reactive Autonomous Vehicle Simulation and Benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
13. Ding, S., Vasa, S., Ramadwar, A.: Explanation-Driven Counterfactual Testing for Faithfulness in Vision-Language Model Explanations. In: NeurIPS 2025 Workshop on Regulatable ML (2025)
14. Fadadu, S., Pandey, S., Hegde, D., Shi, Y., Chou, F.C., Djuric, N., Vallespi-Gonzalez, C.: Multi-View Fusion of Sensor Data for Improved Perception and

Prediction in Autonomous Driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)
15. Feng, B., Mei, Z., Li, B., Ost, J., Girgis, R., Majumdar, A., Heide, F.: VERDI: VLM-Embedded Reasoning for Autonomous Driving. arXiv preprint arXiv:2505.15925 (2025)
16. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24823–24834 (2025)
17. Gholami, M., Rezaei, A., Weimin, Z., Mao, S., Zhou, S., Zhang, Y., Akbari, M.: Spatial Reasoning with Vision-Language Models in Ego-Centric Multi-View Scenes. arXiv preprint arXiv:2509.06266 (2025)
18. Guo, Z., Gubernatorov, K., Asfaw, S., Yagudin, Z., Tsetserukou, D.: VDT-Auto: End-to-End Autonomous Driving with VLM-Guided Diffusion Transformers. arXiv preprint arXiv:2502.20108 (2025)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
20. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A Comprehensive Overhaul of Feature Distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1921–1930 (2019)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531 (2015)
22. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-Oriented Autonomous Driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
23. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: BEVDet: High-Performance Multi-Camera 3D Object Detection in Bird's-Eye-View. arXiv Preprint arXiv:2112.11790 (2021)
24. Huang, Z., Tang, T., Chen, S., Lin, S., Jie, Z., Ma, L., Wang, G., Liang, X.: Making Large Language Models Better Planners with Reasoning-Decision Alignment. In: European Conference on Computer Vision (ECCV). Springer (2024)
25. Hwang, J.J., Xu, R., Lin, H., Hung, W.C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., Zhou, Y., Guo, J., Anguelov, D., Tan, M.: EMMA: End-to-End Multimodal Model for Autonomous Driving. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=kH3t5lmOU8
26. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-to-End Autonomous Driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
27. Jiang, B., Chen, S., Liao, B., Zhang, X., Yin, W., Zhang, Q., Huang, C., Liu, W., Wang, X.: Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. arXiv preprint arXiv:2410.22313 (2024)
28. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
29. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3D: Towards Robust and Reliable 3D Perception Against Corruptions. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)

30. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In: European Conference on Computer Vision (ECCV). Springer (2022)
31. Li, Z., Yu, Z., Lan, S., Li, J., Kautz, J., Lu, T., Alvarez, J.M.: Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
32. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
33. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. Advances in Neural Information Processing Systems **36** (2023)
35. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2023)
36. Ljungbergh, W., Tonderski, A., Johnander, J., Caesar, H., Åström, K., Felsberg, M., Petersson, C.: NeuroNCAP: Photorealistic Closed-Loop Safety Testing for Autonomous Driving. In: European Conference on Computer Vision (ECCV). Springer (2024)
37. Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., Manocha, D., Zhu, X.: Vision-Centric BEV Perception: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
38. Min, C., Zhao, D., Xiao, L., Zhao, J., Xu, X., Zhu, Z., Jin, L., Li, J., Guo, Y., Xing, J., et al.: DriveWorld: 4D Pre-Trained Scene Understanding via World Models for Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
39. Monninger, T., Anwar, M.Z., Antol, S., Staab, S., Ding, S.: AugMapNet: Improving Spatial Latent Structure via BEV Grid Augmentation for Enhanced Vectorized Online HD Map Construction. arXiv preprint arXiv:2503.13430 (2025)
40. Monninger, T., Zhang, Z., Mo, Z., Anwar, M.Z., Staab, S., Ding, S.: MapDiffusion: Generative diffusion for vectorized online hd map construction and uncertainty estimation in autonomous driving. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4099–4106. IEEE (2025)
41. Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., De Albuquerque, V.H.C.: Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. IEEE Transactions on Intelligent Transportation Systems **22**(7), 4316–4336 (2020)
42. Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., Ren, L.: VLP: Vision Language Planning for Autonomous Driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
43. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision. In: International Conference on Machine Learning (ICML). PMLR (2021)

45. Renz, K., Chen, L., Arani, E., Sinavski, O.: SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11993–12003 (2025)
46. Sato, T., Shen, J., Wang, N., Jia, Y., Lin, X., Chen, Q.A.: Dirty Road Can Attack: Security of Deep Learning Based Automated Lane Centering Under Physical-World Attack. In: 30th USENIX Security Symposium (USENIX Security) (2021)
47. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
48. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: DriveLM: Driving with Graph Visual Question Answering. In: European Conference on Computer Vision (ECCV). Springer (2024)
49. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In: Conference on Robot Learning. pp. 4698–4726. PMLR (2024)
50. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-Based Image Description Evaluation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
51. Wan, Z., Shen, J., Chuang, J., Xia, X., Garcia, J., Ma, J., Chen, Q.A.: Too Afraid to Drive: Systematic Discovery of Semantic DoS Vulnerability in Autonomous Driving Planning Under Physical-World Attacks. In: ISOC Network and Distributed Systems Security (NDSS) Symposium (2022)
52. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
53. Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al.: Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. arXiv preprint arXiv:2511.00088 (2025)
54. Winter, K., Azer, M., Flohr, F.B.: BEVDriver: Leveraging BEV Maps in LLMs for Robust Closed-Loop Driving. arXiv Preprint arXiv:2503.03074 (2025)
55. Wu, P., Chen, S., Metaxas, D.N.: MotionNet: Joint Perception and Motion Prediction for Autonomous Driving Based on Bird's Eye View Maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
56. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. In: IEEE/CVF International Conference on Computer Vision (ICCV) (October 2025)
57. Xie, S., Kong, L., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Benchmarking and Improving Bird's Eye View Perception Robustness in Autonomous Driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5) (2025). https://doi.org/10.1109/TPAMI.2025.3535960
58. Xie, S., Li, Z., Wang, Z., Xie, C.: On the Adversarial Robustness of Camera-Based 3D Object Detection. Transactions on Machine Learning Research (2024)
59. Xie, Y., Xu, R., He, T., Hwang, J.J., Luo, K., Ji, J., Lin, H., Chen, L., Lu, Y., Leng, Z., et al.: S4-Driver: Scalable Self-Supervised Driving Multimodal Large Language Model with Spatio-Temporal Visual Representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)

60. Xu, Y., Hu, Y., Zhang, Z., Meyer, G.P., Mustikovela, S.K., Srinivasa, S., Wolff, E.M., Huang, X.: VLM-AD: End-to-End Autonomous Driving Through Vision-Language Model Supervision. arXiv preprint arXiv:2412.14446 (2024)
61. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: DriveGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model. IEEE Robotics and Automation Letters **9**(10) (2024)
62. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 Technical Report. arXiv Preprint arXiv:2505.09388 (2025)
63. Yang, H., Zhang, S., Huang, D., Wu, X., Zhu, H., He, T., Tang, S., Zhao, H., Qiu, Q., Lin, B., et al.: UniPAD: A Universal Pre-Training Paradigm for Autonomous Driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
64. Yang, Z., Jia, X., Li, H., Yan, J.: LLM4Drive: A Survey of Large Language Models for Autonomous Driving. arXiv preprint arXiv:2311.01043 (2023)
65. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X., Guo, N.: FutureSightDrive: Thinking Visually with Spatio-Temporal COT for Autonomous Driving. arXiv preprint arXiv:2505.17685 (2025)
66. Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., Lu, J.: BEVerse: Unified Perception and Prediction in Bird's-Eye-View for Vision-Centric Autonomous Driving. arXiv Preprint arXiv:2205.09743 (2022)
67. Zhou, X., Liang, D., Tu, S., Chen, X., Ding, Y., Zhang, D., Tan, F., Zhao, H., Bai, X.: HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation. In: IEEE/CVF International Conference on Computer Vision (ICCV) (October 2025)
68. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: OpenDriveVLA: Towards End-to-End Autonomous Driving with Large Vision Language Action Model. arXiv preprint arXiv:2503.23463 (2025)
69. Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J.: AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning. arXiv preprint arXiv:2506.13757 (2025)
70. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv Preprint arXiv:2504.10479 (2025)