2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

VLM-E2E: Enhancing End-to-End Autonomous Driving with Multimodal Driver Attention Fusion

Pei Liu, Haipeng Liu, Haichao Liu, Xin Liu, Jinxin Ni, Jun Ma, *Senior Member, IEEE*

***Abstract*— Human drivers adeptly navigate complex scenarios by utilizing rich attentional semantics, but the current autonomous systems struggle to replicate this ability, as they often lose critical semantic information when converting 2D observations into 3D space. In this sense, it hinders their effective deployment in dynamic and complex environments. Leveraging the superior scene understanding and reasoning abilities of Vision-Language Models (VLMs), we propose VLM-E2E, a novel framework that uses the VLMs to enhance training by providing attentional cues. Our method integrates textual representations into Bird's-Eye-View (BEV) features for semantic supervision, which enables the model to learn richer feature representations that explicitly capture the driver's attentional semantics. By focusing on attentional semantics, VLM-E2E better aligns with human-like driving behavior, which is critical for navigating dynamic and complex environments. Furthermore, we introduce a BEV-Text learnable weighted fusion strategy to address the issue of modality importance imbalance in fusing multimodal information. This approach dynamically balances the contributions of BEV and text features, ensuring that the complementary information from visual and textual modalities is effectively utilized. By explicitly addressing the imbalance in multimodal fusion, our method facilitates a more holistic and robust representation of driving environments. We evaluate VLM-E2E on the nuScenes dataset and achieve significant improvements in perception, prediction, and planning over the baseline end-to-end model, showcasing the effectiveness of our attention-enhanced BEV representation in enabling more accurate and reliable autonomous driving tasks.**

I. Introduction

Autonomous driving has witnessed remarkable progress in recent years [1], [2], [3], with significant advancements in key areas such as perception [4], motion prediction [5], and planning [6]. These developments have laid a solid foundation for achieving more accurate and safer driving decisions. Among these, end-to-end (E2E) autonomous driving has emerged as a transformative paradigm, leveraging large-scale data to demonstrate impressive planning capabilities. By directly mapping raw sensor inputs to driving actions, E2E approaches bypass the need for handcrafted intermediate modules, enabling more flexible and scalable solutions. However, Despite these advancements, traditional end-to-end autonomous driving approaches predominantly predict future trajectories or control signals directly, without explicitly considering the driver's attention to critical information such as traffic dynamics and navigation cues. E2E systems often struggle in complex and ambiguous scenarios due to their limited ability to reason about high-level semantics and contextual cues, such as traffic rules, driver attention, and dynamic interactions. In contrast, human drivers rely on an attentional decision-making process, where attention to both the surrounding traffic environment and navigation guidance plays a critical role [7]. For instance, when approaching an intersection, human drivers naturally prioritize traffic signals, pedestrian movements, and lane markings, dynamically adjusting their focus based on the evolving scene.

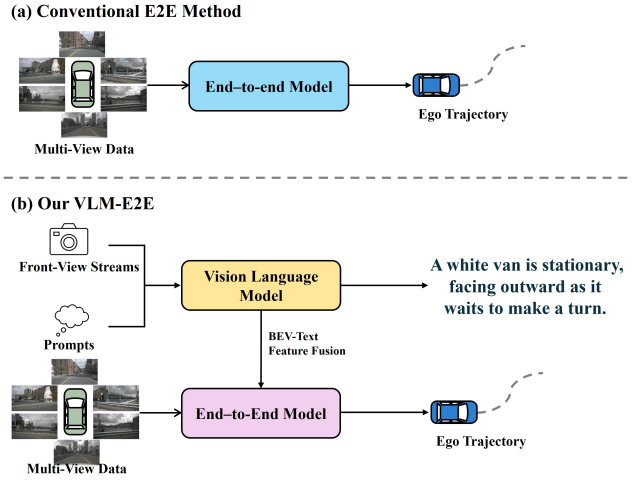


Fig. 1. VLM-E2E augments the end-to-end driving model with semantic textual descriptions during training. These descriptions extract driver attention from VLMs to encourage the model to learn richer attentional semantics.

This limitation has spurred the integration of Vision-Language Models (VLMs) [8], [9] into autonomous driving frameworks. Trained on vast multimodal datasets, VLMs excel at tasks requiring high-level semantic reasoning, such as interpreting complex scenes, predicting dynamic interactions, and generating contextual descriptions. Their ability to leverage commonsense knowledge makes them particularly well-suited for addressing challenges in autonomous driving, such as understanding traffic rules, identifying vulnerable road users, and making safe decisions in ambiguous scenarios. By generating text-based descriptions of critical driving cues, VLMs can explicitly capture and prioritize regions of interest that align with human driver attention. This capability enables more human-like decision-making, partic-

Pei Liu, Haichao Liu, and Xin Liu are with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China (e-mail: {pliu061, hliu369, xliu969}@connect.hkust-gz.edu.cn).

Haipeng Liu is with Li Auto Inc., Shanghai 201800, China (e-mail: liuhaipeng2012@live.com).

Jinxin Ni is with Xiamen University, Xiamen 361102, China (e-mail: nijinxinlxq@outlook.com).

Jun Ma is with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China, and also with The Hong Kong University of Science and Technology, Hong Kong SAR, China (e-mail: jun.ma@ust.hk). *(Corresponding Author: Jun Ma.)*

ularly in safety-critical scenarios where attentional focus is paramount.

Motivated by these challenges, we propose VLM-E2E (illustrated in Fig. 1), a novel framework designed to enhance autonomous driving systems by incorporating a deeper understanding of driver attentional semantics. Our approach addresses three key questions:

How to integrate VLMs with E2E models? While most existing methods integrate VLMs with decision-making modules [10] or other high-level components [11], leveraging their semantic understanding capabilities to enhance decision processes, our approach introduces a novel integration strategy. Instead of limiting VLMs to decision modules, we combine them directly with the BEV module, which is widely used to represent and process spatial information from multiple perspectives in autonomous driving. By integrating VLMs into the BEV module, we enable BEV representations to incorporate both visual and textual features, resulting in richer and more semantic-aware spatial understanding. This integration allows the model to not only perceive geometric structures but also reason about high-level driver attentional semantics.

How to fuse vision and text representations? Existing methods for fusing vision and text representations predominantly rely on attention-based mechanisms [12], [13], such as cross-attention or co-attention modules, to align and enhance interactions between modalities. While effective, these approaches often rely on predefined attention mechanisms that lack flexibility in adapting to varying task requirements and are time and memory-consuming. To address this issue, we propose a BEV-Text learnable weighted fusion strategy, where the importance of each modality is dynamically determined through a learnable weight mechanism. This approach allows the model to adaptively emphasize visual or textual features based on their relevance to the task, leading to a more robust and context-aware multimodal representation. For instance, in scenarios requiring precise localization such as lane keeping, the model can prioritize BEV features, while in scenarios requiring high-level reasoning, such as red lights, it can emphasize text features.

How to represent driver attentional environment? To effectively model a driver-attentional environment, we propose a multimodal framework that leverages vision-language representations. First, we utilize front-view images captured from the driving scene as input to BLIP-2 [14] to generate initial textual descriptions of the environment. These descriptions provide a semantic understanding of key objects and events within the driver's vision scope. To address the hallucination problem of VLMs, we further refine these textual representations using ground truth annotations and high-level maneuvering intentions. This refinement ensures that the generated text is not only accurate but also contextually aligned with the driving task. Finally, the refined text is encoded into a dense representation using a pre-trained CLIP [15] model, which aligns the textual information with visual features in a shared embedding space. This textual representation enables the model to capture driver attentional cues, such as focusing on pedestrians near crosswalks or traffic signals at intersections, leading to more human-like decision-making and better safety performance. The key contributions of this work can be summarized as follows:

- We propose VLM-E2E, a novel framework that leverages VLMs to enrich the training process with attentional understanding. By integrating semantic and contextual information, VLM-E2E explicitly captures driver attentional semantics, which enables more human-like decision-making in complex driving scenarios.
- We introduce a BEV-Text learnable weighted fusion strategy that dynamically balances the contributions of BEV and textual modalities. This adaptive fusion mechanism is computationally efficient, which requires minimal additional overhead while significantly enhancing the model's adaptability and robustness.
- To address the hallucination problem of VLMs, we incorporate semantic refinement of textual annotations generated from front-view images. By leveraging ground truth (GT) labels and high-level maneuvering intentions, we ensure that the textual representations are both accurate and highly relevant to the driving task, enhancing the model's ability to reason about critical driving cues.
- Extensive experiments on the nuScenes dataset demonstrate the superiority of VLM-E2E over existing methods. Our framework achieves significant improvements in handling complex driving scenarios, showcasing its ability to integrate geometric precision with high-level semantic reasoning for safer and more reliable autonomous driving.

II. Related Work

A. End-to-End Autonomous Driving

Recent advances in end-to-end autonomous driving systems have established vision-based frameworks like ST-P3 [16] and UniAD [2], which unify perception, prediction, and planning for improved scene understanding. Follow-up works such as VAD [1] and VADv2 [17] further enhanced dynamic environment handling through vectorized scene representations. Subsequent innovations like Ego-MLP [18] and BEV-Planner [19] introduce modular designs focusing on ego-state modeling and environment interactions.

B. Vision Language Models in Autonomous Driving

VLMs have gained traction in autonomous driving for their ability to enhance reasoning and interpretability in end-to-end systems. Recent studies demonstrate diverse integration approaches. For instance, Drive-with-LLMs [20] processes perception data via transformers for trajectory prediction, while DriveGPT4 [21] generates control signals with natural language explanations. Frameworks like DriveMLM [22] and ELM [23] validate VLM-based planning through simulation and cross-domain pretraining.

Specialized datasets [24], [25] and hybrid systems such as DriveVLM [26], which refines VLM outputs into precise

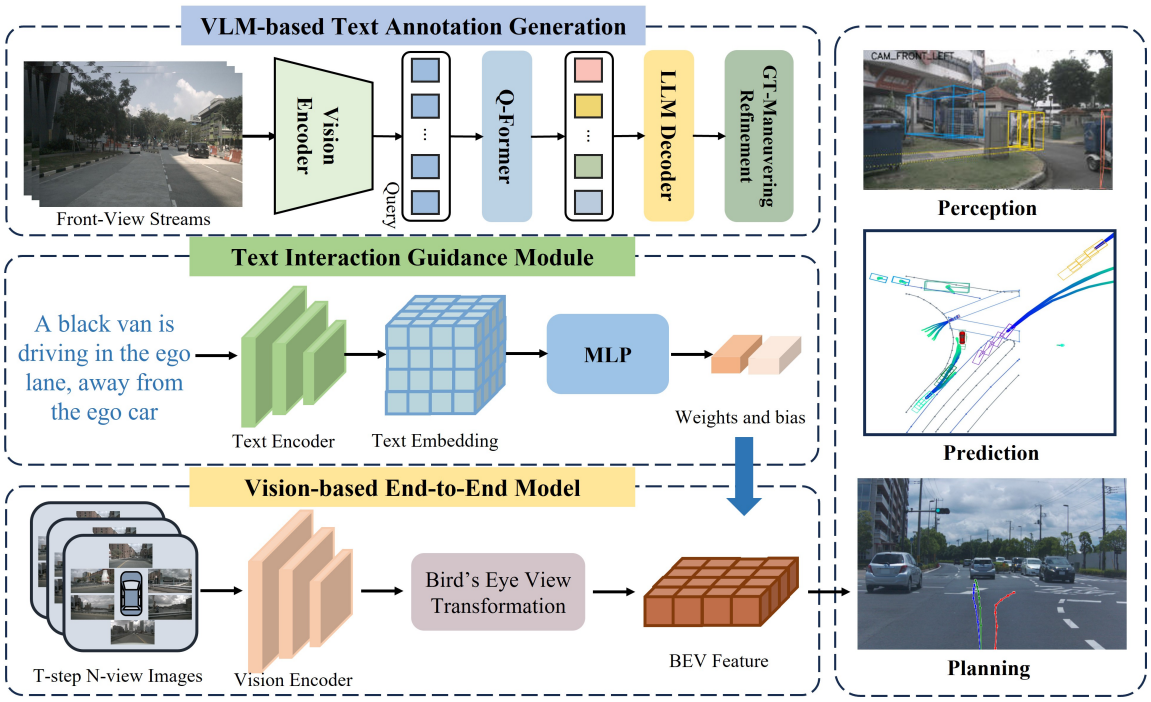


Fig. 2. We present VLM-E2E, a driver attention enhanced end-to-end vision-based framework. VLM-E2E consists of three modules: VLM-based Text Annotation Generation, Text Interaction Guidance Module, and Vision-based End-to-end Model.

trajectories, and VLM-AD [10] and Senna [27], which simplify decision-making through text-based planning, highlight the field's advancements. However, existing methods rarely address the design of driving-specific textual semantics, particularly those capturing driver-attentional cues essential for human-like situational awareness. Our work bridges this gap by integrating attentional textual prompts to prioritize safety-critical scenarios.

III. Methodology

In this section, we provide a detailed introduction to VLM-E2E, as illustrated in Fig. 2. The input scene information includes multi-view image sequences, GT, maneuvering, and user prompts. The front-view image, maneuvering, and user prompts are fed into the VLM-based Text Annotation Generation module to generate descriptive text annotations, while the multi-view images are processed by the visual encoding layer to produce BEV features. These text annotations are then passed to the Text Interaction Guidance Module, where they are encoded into text features using a pre-trained CLIP model. Subsequently, the BEV and text features are fused to support downstream tasks such as perception, prediction, and decision-making. In Section III-A, we introduce the design of VLM-based Text Annotation Generation in detail. Sections III-B and III-C focus on the design of the Text Interaction Guidance Module and Vision-based End-to-End Architecture, respectively.

A. VLM-based Text Annotation Generation

1) Text Annotation: Fig. 2 depicts the proposed pipeline for extracting driver attentional information from visual inputs, leveraging the reasoning capabilities of a pre-trained VLM. The semantic annotation extraction process can be formulated as follows:

$$T = \mathcal{BLIP}_2(P, I_{front}) \tag{1}$$

where $\mathcal{BLIP}_2(\cdot)$ denotes the visual language model BLIP-2, P represents the task-specific prompts, I_{front} is the visual input from the ego vehicle's front camera, and T is the generated textual description providing detailed environment-related information. The goal of this process is to utilize task-specific prompts alongside real-time visual inputs to extract actionable and attentional information from BLIP-2. This approach not only emphasizes critical elements such as pedestrians, traffic signals, and dynamic obstacles but also filters out irrelevant scene details, ensuring that the outputs directly support driving decisions.

In our work, we employ a state-of-the-art vision language model BLIP-2 [14], capable of performing complex reasoning over visual contexts, to generate precise and contextually relevant descriptions. The model interprets visual scenarios guided by prompts and outputs textual descriptions. This method enhances the dataset's richness by providing driver attentional annotations, thereby improving the understanding and decision-making capabilities of downstream driving models.

We encountered a challenge in determining the visual input. That is, selecting the right images from multiple cameras that can cover 360 degrees of the ego vehicle. Considering that we want to capture the driver's attentional semantics when driving, the front view images usually contain the most relevant information required for most driving tasks. All-view images contain more distracting information that affects

the system's decision-making, so we choose to use only the front-view images to extract the attentional information. In addition, considering that the ego vehicle and its surroundings are in dynamic motion and the hallucination problem inherent in large models, we use the GT and maneuvering to refine the annotations of dynamic objects, inspired by [28].

B. Text Interaction Guidance Module

The driver's attentional text descriptions preserve rich visual semantic cues. It is complementary to the BEV features that mainly represent the 3D geometric information. Hence, BEV-Text fusion is proposed for comprehensive scene understanding from the BEV perspective.

1) Text Encoder: Given a text input T that provides semantic features to guide the BEV-Text fusion network toward achieving a specified fusion result, the text encoder and embedding within the text interaction guidance architecture are responsible for transforming this input into a text embedding. Among various VLMs, we adopt CLIP [15] due to its lightweight architecture and efficient text feature extraction capabilities. Compared to other VLMs, CLIP is computationally less demanding and produces text embeddings with relatively small text tokens of 77, which significantly enhances the efficiency of subsequent BEV-Text feature fusion. We freeze the text encoder from CLIP to preserve its consistency and leverage its pretrained knowledge. This process can be formally expressed as:

$$f_t = \mathcal{CLIP}_e(T) \tag{2}$$

where $\mathcal{CLIP} \in \mathbb{R}^{N\times L}$ denotes the CLIP model with wights frozen. f_t is the text semantic representations.

In different but semantically similar texts, the extracted features should be close in the reduced Euclidean space. Furthermore, we use the MLP F_m^i to mine this connection and further map the text semantic information and the semantic parameters. Therefore, it can be obtained:

$$\gamma_m = F_m^1(f_t), \beta_m = F_m^2(f_t) \tag{3}$$

where F_m^1 and F_m^2 are the chunk operations to form the text semantic parameters.

As depicted in Fig. 2, in the text interaction guidance module, textual parameters interact through feature interaction and BEV features s_t, to obtain the effect of guidance. The feature modulation consists of scale scaling and bias control, which adjust the features from two perspectives, respectively. In particular, a residual connection is used to reduce the difficulty of network fitting, inspired by [29]. For simplicity, it can be described as:

$$x_t = (1 + \gamma_m) \odot s_t + \beta_m \tag{4}$$

where $\odot$ denotes the Hadamard product. x_t denotes the fused BEV feature, and s_t denotes the BEV feature defined in Section III-C.1.

C. Vision-based End-to-End Model

1) Spatial Temporal BEV Perception: In our framework, the BEV representation is constructed from multi-camera images. The input multi-camera images $\{I_t^1, \cdots, I_t^n\}, n = 6$ at time t are first passed through a shared backbone network, EfficientNet-b4 [30], to extract high-dimensional feature maps. For each camera image k at time t, we get its encoder features $e_t^k \in \mathbb{R}^{C\times H_e\times W_e}$ and depth estimation $d_t^k \in \mathbb{R}^{D\times H_e\times W_e}$ with C denotes the number of feature channels, D is the number of discrete depth values and (H_e, W_e) depicts the spatial feature size. Implicit depth estimation is applied to infer the depth information for each pixel, enabling the construction of a 3D feature volume. Since the depth values are estimated, we take the outer product of the features with the depth estimation.

$$\hat{e}_t^k = e_t^k \otimes d_t^k \tag{5}$$

where $\hat{e}_t^k \in \mathbb{R}^{C\times D\times H_e\times W_e}$. Then, to transform the 2D perspective features into a 3D space, we employ a feature lifting module. This module uses camera intrinsic and extrinsic parameters to project the 2D features into a 3D voxel space. The 3D feature volume is then collapsed into a 2D BEV representation by aggregating features along the vertical axis to form the BEV view features $b_t \in \mathbb{R}^{C\times H\times W}$, with (H, W) denoting the spatial size of BEV feature. This is achieved through attention-based aggregation, which preserves the most salient features while maintaining spatial consistency. The resulting BEV map provides a top-down view of the scene, encapsulating both geometric and semantic information.

In addition to the BEV construction pipeline described above, we further incorporate temporal modeling to enhance the dynamic understanding of the scene. Specifically, given the current timestamp t and its h historical BEV features $\{b_{t-h}, \cdots, b_{t-1}, b_t\}$, we first align the historical features to the current frame's coordinate system using a temporal alignment module. This process leverages the relative transformation and rotation matrix $M_{t-i\to t} \in \mathbb{R}^{4\times 4}$ between adjacent frames. The past BEV feature b_{t-i} is then spatially transformed as:

$$\hat{b}_{t-i} = \mathcal{W}(b_{t-i}, M_{t-i\to t}), \quad i = 1, 2 \tag{6}$$

where $\mathcal{W}(\cdot)$ denotes the pose-based BEV feature warping operation, and $\hat{b}_{t-i}$ represents the aligned historical features. Subsequently, the aligned BEV features from the h frames are concatenated to form the spatiotemporal input $\hat{b} = [\hat{b}_{t-h}, \cdots, \hat{b}_{t-1}, \hat{b}_t] \in \mathbb{R}^{h\times C\times H\times W}$. To capture long-term dependencies in dynamic scenes, we use a spatiotemporal transformer module F_s.

$$s_t = F_s(\hat{b}_{t-h}, \cdots, \hat{b}_{t-1}, \hat{b}_t) \tag{7}$$

where $s_t \in \mathbb{R}^{h\times C\times H\times W}$ is the spatiotemporally fused BEV feature. F_s is a spatiotemporal convolutional unit with cross-frame self-attention. Our spatial-temporal BEV representation explicitly models the static and dynamic evolution of the scene, enabling the BEV representation to encode geometric structures and temporal continuity simultaneously.

D. Attention Guided Future Planning

The primary objective of the proposed motion planning is to generate trajectories that ensure safety, comfort, and efficient progress toward the goal. To achieve this, we employ a motion planner. The planner generates a set of kinematically feasible trajectories, each of which is evaluated using a learned scoring function, inspired by [31], [32], [33], [16].

To ensure real-time performance, the set of sampled trajectories must remain sufficiently small. However, this set must also represent various possible maneuvers and actions to avoid encroaching obstacles. To strike this balance, we employ a sampling strategy that is aware of the lane structure, ensuring that the sampled trajectories effectively capture a diverse range of driving behaviors while remaining computationally feasible.

In particular, we follow the trajectory sampling method proposed in [34], [35], where trajectories are generated by combining longitudinal motion with lateral deviations relative to specific lanes, such as the current SDV lane or adjacent lanes. This approach allows the planner to sample trajectories that adhere to lane-based driving principles while incorporating variations in lateral motion. These variations enable the motion planner to handle a wide array of traffic scenarios.

To ensure the planned trajectory adheres to driver attention on traffic regulations and routes, we utilize a temporal refinement module that dynamically integrates traffic regulations. Leveraging front-view camera features e_{front} from the encoder, we initialize a GRU-based refinement network to iteratively adjust the initially selected trajectory. The front-view features explicitly encode traffic regulations semantics, enabling the model to halt at red lights or proceed through green signals. The recurrent architecture ensures smooth transitions between trajectory points, mitigating abrupt steering or acceleration changes.

IV. Experiments

A. Experimental Settings

Dataset. We evaluate our method on the nuScenes dataset [36], a large-scale autonomous driving benchmark comprising 1,000 diverse driving scenes, each spanning 20 seconds with annotations provided at 2 Hz. The dataset features a 360° multi-camera rig comprising six synchronized cameras with minimal field-of-view overlap. Precise camera intrinsic and extrinsic are provided for each frame to ensure accurate spatial alignment.

All labels are transformed into the ego vehicle's reference frame using GT future ego-motion, ensuring temporal consistency across frames. Besides, maneuvering is used to correct annotations from VLMs.

Implementation Details. Our framework processes 1 second of historical sensor data, equivalent to 3 frames at 2Hz, to predict trajectories over a 3 s horizon spanning 6 future frames. Inputs include synchronized multi-camera RGB streams with 224×480 pixel resolution and BEV grid maps covering a 100m×100m area at 0.5m per pixel resolution. Training employs the Adam optimizer with a learning rate of 2.0×10^{-3} for 20 epochs, using mixed precision on 4 NVIDIA A6000 GPUs and a batch size of 6.

TABLE I

Perception results. We report the semantic segmentation IoU (%) in BEV.

Method	Drivable Area	Lane	Vehicle	Pedestrian
VED [37]	60.82	16.74	23.28	11.93
VPN [38]	65.97	17.05	28.17	10.26
PON [39]	63.05	17.19	27.91	13.93
Lift-Splat [40]	72.23	19.98	31.22	15.02
IVMP [41]	**74.70**	20.94	34.03	17.38
FIERY [42]	71.97	33.58	38.00	17.15
ST-P3 [16]	74.38	38.47	38.79	14.06
VLM-E2E	74.69	**39.33**	**39.08**	**17.49**

TABLE II

Prediction results. We report the semantic and instance segmentations in BEV for 2s in the future.

Method	IoU ↑	PQ ↑	SQ ↑	RQ ↑
FIERY [42]	36.20	27.80	-	-
ST-P3 [16]	36.89	29.10	**69.77**	41.71
VLM-E2E	**38.54**	**29.83**	69.56	**42.88**

TABLE III

Planning results. We report the L2 (m) and Collision Rate CR (%) across 1s, 2s, 3s.

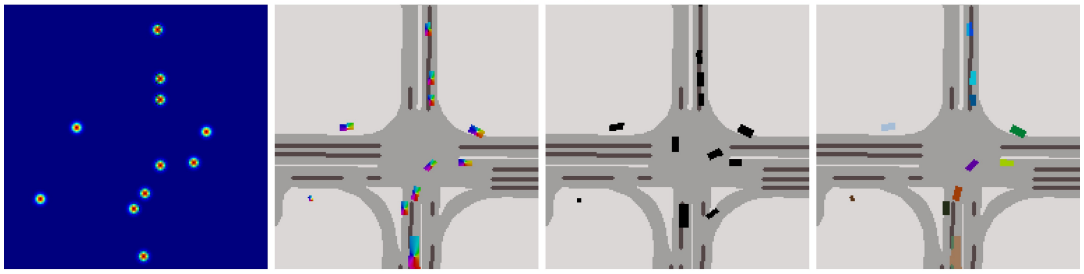
Method	L2 (m) ↓				CR (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Vanilla [43]	0.50	1.25	2.80	1.52	0.68	0.98	2.76	1.47
NMP [44]	0.61	1.44	3.18	1.74	0.66	0.90	2.34	1.30
Freespace [6]	0.56	1.27	3.08	1.64	0.65	0.86	1.64	1.05
ST-P3 [16]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [2]	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD [1]	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
GenAD [45]	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19
Senna [27]	0.37	0.54	0.86	0.59	0.09	0.12	0.33	0.18
VLM-E2E	**0.28**	**0.50**	**0.80**	**0.53**	**0.01**	**0.06**	**0.20**	**0.09**

B. Quantitative Results

Perception. Table I presents the perception performance of various methods across four key categories: Drivable Area, Lane, Vehicle, and Pedestrian. Our proposed VLM-E2E model demonstrates significant improvements over existing approaches, achieving the best results in three out of four categories. Specifically, VLM-E2E outperforms ST-P3 in lane detection with a 2.24% relative improvement, vehicle detection with an 0.75% increase, and Pedestrian detection with a 24.40% boost on the nuScenes validation set. While IVMP achieves the highest score in drivable area detection, VLM-E2E closely follows with a score of 74.69, demonstrating preserved geometric reasoning in BEV space. These results demonstrate that integrating driver-attentional features enhances critical perception tasks without compromising foundational scene understanding.

Prediction. Table II presents the prediction performance of various methods on the task of semantic and instance segmentations in BEV for a future horizon of 2.0 seconds.

(a) Input Images



(b) Instance Center (c) Instance Offset (c) Future Flow (c) Instance Segmentation

Fig. 3. Qualitative analysis on prediction. (a) shows the multi-view input images. (b) shows the heatmap (blue to red) which illustrates the probability distribution of instance centers within the scene, with warmer colors indicating higher confidence regions. (c) represents the vehicles' segmentation, which effectively distinguishes individual instances in the complex traffic scenario. (d) reveals the directional vectors pointing towards the corresponding instance centers for each pixel, demonstrating the model's understanding of spatial relationships. (e) exhibits consistency within each instance, reflecting the characteristic rigid-body motion of vehicles.

TABLE IV

ABLATIONS OF TEXT ENCODERS. WE REPORT THE SEMANTIC SEGMENTATION IOU (%) IN BEV.

Method	Drivable Area	Lane	Vehicle	Pedestrian
Bert	73.44	38.19	38.53	15.90
CLIP	**74.69**	**39.33**	**39.08**	**17.49**

TABLE V

ABLATIONS OF INPUT TEXT VIEWS. WE REPORT THE SEMANTIC SEGMENTATION IOU (%) IN BEV.

Text View	Drivable Area	Lane	Vehicle	Pedestrian
All View	73.50	37.98	38.79	**17.51**
Front View	**74.69**	**39.33**	**39.08**	17.49

We evaluate the methods using IoU, PQ, SQ, and RQ. VLM-E2E achieves the best performance across IoU, Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ). Specifically, VLM-E2E attains an IoU of 38.54, representing a 4.47% improvement over ST-P3. In terms of PQ, VLM-E2E achieves a PQ of 29.83, demonstrating superior instance segmentation capabilities compared to ST-P3 and FIERY. These results highlight the effectiveness of VLM-E2E in capturing both semantic and instance-level information in dynamic driving scenarios.

Planning. As shown in Table III, our VLM-E2E achieves the best overall performance among all methods. It achieves the lowest average L2 error, 0.53 m, and substantially reduces the collision rate to 0.09%. Compared to previous state-of-the-art approaches, such as GenAD, which achieves 0.52 meters and 0.19 percent, and Senna, which reaches 0.59 m and 0.18%, VLM-E2E delivers notable improvements in both trajectory accuracy and safety. Moreover, the gains over traditional baselines like Vanilla and NMP are even more significant. These results clearly demonstrate the effectiveness of incorporating VLM-guided attention into end-to-end autonomous driving frameworks.

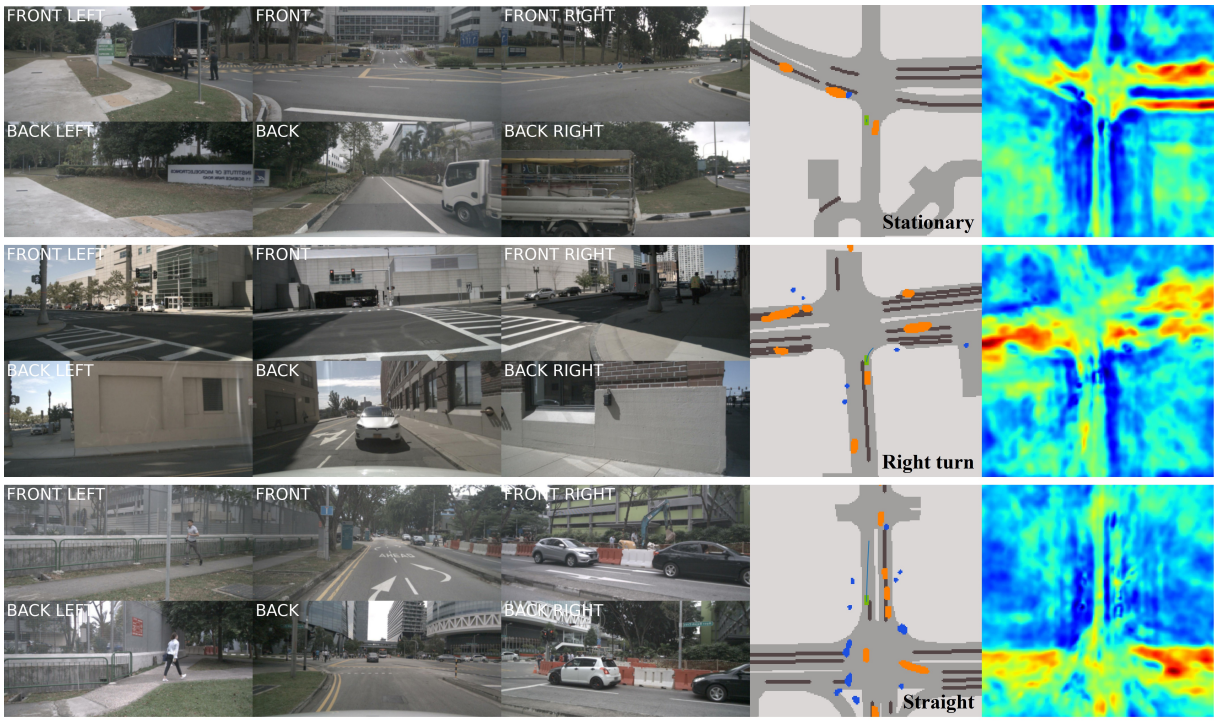
C. Qualitative Analysis

Fig. 3 demonstrates the generated outputs, including instance segmentation, instance center, instance offset, and future flow. Fig. 3(b) features a heatmap highlighting detected objects, while Fig. 3(c) displays the instance segmentation results, where each segment is color-coded to represent different objects. The offset is a vector pointing to the center of the instance in Fig. 3(d). The future flow Fig. 3(e) is a displacement vector field of the dynamic agents. These visualizations enhance the understanding of spatial relationships and the distribution of elements within the environment, underscoring the model's capability to accurately perceive and segment critical features essential for autonomous driving applications

Fig. 4 illustrates examples of planning scenarios. In the upper scene, the model accurately predicts the route when provided with turning instructions, effectively navigating through crowded environments in a manner similar to human demonstrations. The bottom scene demonstrates the model's predictions when instructed to proceed straight at an intersection, further highlighting its ability to handle diverse driving scenarios with precision. These examples emphasize the model's advanced planning capabilities in complex and dynamic environments.

D. Ablation Study

Text Encoder. Table IV compares the performance of text encoders, Bert [46] and CLIP, across four detection categories. Bert shows a consistent performance decline, with



(a) Input Images (b) Planning (c) Cost Heatmap

Fig. 4. Qualitative analysis on planning. (a) shows the multi-view input images. (b) shows the planned trajectory (blue). (d) presents the learned costmap with a warmer color indicates a lower cost.

decreases of 1.67%, 2.90%, 1.41%, and 9.09%, respectively, highlighting its limitations in perception tasks. In contrast, CLIP demonstrates superior capability in encoding textual annotations, achieving higher accuracy across all categories. The results validate that CLIP's joint embedding space effectively mitigates modality gaps between text annotations and BEV representations.

Input Text View. Table V evaluates the impact of text view selection on BEV semantic segmentation performance. When transitioning from all view to front view text inputs, we observe performance improvements of 1.62% for drivable area, 3.55% for lane, and 0.75% for vehicle, respectively. This improvement primarily stems from eliminating interference caused by redundant textual context in all views, where excessive multi-view descriptions introduce semantic noise that distracts the model from task-critical features. The results validate our design choice to adopt the front view as the default configuration, as it optimally balances semantic specificity and noise suppression for scene understanding tasks.

V. Conclusion

We propose VLM-E2E, an E2E autonomous driving framework that leverages VLMs to enhance driver-attentional semantic understanding. We introduce a BEV-Text learnable weighted fusion strategy balancing geometric-semantic features to address modality imbalance and semantic underutilization in existing systems, along with a spatiotemporal coherence mechanism for temporal consistency, and an attention-guided trajectory refinement for probabilistic prediction. Future work will extend integration with advanced E2E architectures for improved generalization.

References

[1] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "VAD: Vectorized Scene Representation for Efficient Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

[2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-Oriented Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

[3] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries," in *Conference on Robot Learning*, pp. 180–191, PMLR, 2022.

[4] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, *et al.*, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," 2022.

[5] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "ViP3D: End-to-End Visual Trajectory Prediction via 3D Agent Queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5496–5506, 2023.

[6] A. Prakash, K. Chitta, and A. Geiger, "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7077–7087, 2021.

[7] D. Badre, "Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 193–200, 2008.

[8] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, S. XiXuan, *et al.*, "CogVLM: Visual Expert for Pretrained Language Models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121475–121499, 2025.

[9] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, *et al.*, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

[10] Y. Xu, Y. Hu, Z. Zhang, G. P. Meyer, S. K. Mustikovela, S. Srinivasa, E. M. Wolff, and X. Huang, "VLM-AD: End-to-End Autonomous Driving through Vision-Language Model Supervision," *arXiv preprint arXiv:2412.14446*, 2024.

[11] Y. Ma, T. Wei, N. Zhong, J. Mei, T. Hu, L. Wen, X. Yang, B. Shi, and Y. Liu, "LeapVAD: A Leap in Autonomous Driving via Cognitive Perception and Dual-Process Thinking," *arXiv preprint arXiv:2501.08168*, 2025.

[12] A. Vaswani, "Attention is All you Need," *Advances in Neural Information Processing Systems*, 2017.

[13] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A Survey on Vision Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *International Conference on Machine Learning*, pp. 19730–19742, PMLR, 2023.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[16] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning," in *European Conference on Computer Vision*, pp. 533–549, Springer, 2022.

[17] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang, "VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning," *arXiv preprint arXiv:2402.13243*, 2024.

[18] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang, "Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes," *arXiv preprint arXiv:2305.10430*, 2023.

[19] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14864–14873, 2024.

[20] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving," in *2024 IEEE International Conference on Robotics and Automation*, pp. 14093–14100, IEEE, 2024.

[21] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model," *IEEE Robotics and Automation Letters*, 2024.

[22] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, *et al.*, "DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving," *arXiv preprint arXiv:2312.09245*, 2023.

[23] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied Understanding of Driving Scenarios," in *European Conference on Computer Vision*, pp. 129–148, Springer, 2024.

[24] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "DriveLM: Driving with Graph Visual Question Answering," in *European Conference on Computer Vision*, pp. 256–274, Springer, 2024.

[25] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4542–4550, 2024.

[26] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models," *arXiv preprint arXiv:2402.12289*, 2024.

[27] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving," *arXiv preprint arXiv:2410.22313*, 2024.

[28] S. Moon, H. Woo, H. Park, H. Jung, R. Mahjourian, H.-g. Chi, H. Lim, S. Kim, and J. Kim, "VisionTrap: Vision-Augmented Trajectory Prediction Guided by Textual Descriptions," in *European Conference on Computer Vision*, pp. 361–379, Springer, 2024.

[29] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024.

[30] M. Tan and E. Le Q V, "rethinking model scaling for convolutional neural networks. 2019," *arXiv preprint arXiv:1905.11946*, 1905.

[31] S. Casas, A. Sadat, and R. Urtasun, "MP3: A Unified Model To Map, Perceive, Predict and Plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14403–14412, 2021.

[32] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, Predict, and Plan: Safe Motion Planning Through Interpretable Semantic Representations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pp. 414–430, Springer, 2020.

[33] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-To-End Interpretable Neural Motion Planner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8660–8669, 2019.

[34] A. Sadat, M. Ren, A. Pokrovsky, Y.-C. Lin, E. Yumer, and R. Urtasun, "Jointly Learnable Behavior and Trajectory Planning for Self-Driving Vehicles," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3949–3956, IEEE, 2019.

[35] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a fren x00e9; t frame," in *2010 IEEE International Conference on Robotics and Automation*, pp. 987–993.

[36] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

[37] C. Lu, M. J. G. Van De Molengraft, and G. Dubbelman, "Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder–Decoder Networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, 2019.

[38] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-View Semantic Segmentation for Sensing Surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[39] T. Roddick and R. Cipolla, "Predicting Semantic Map Representations From Images Using Pyramid Occupancy Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11138–11147, 2020.

[40] J. Philion and S. Fidler, "Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pp. 194–210, Springer, 2020.

[41] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning Interpretable End-to-End Vision-Based Motion Planning for Autonomous Driving with Optical Flow Distillation," in *2021 IEEE International Conference on Robotics and Automation*, pp. 13731–13737, IEEE, 2021.

[42] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15273–15282, 2021.

[43] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the Limitations of Behavior Cloning for Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

[44] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by Cheating," in *Conference on Robot Learning*, pp. 66–75, PMLR, 2020.

[45] W. Zheng, R. Song, X. Guo, C. Zhang, and L. Chen, "GenAD: Generative End-to-End Autonomous Driving," in *European Conference on Computer Vision*, pp. 87–104, Springer, 2024.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.