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

FASIONAD: Adaptive Uncertainty-Gated Fast–Slow Fusion Framework for Safe Autonomous Driving

Ziang Luo[1,*] Sicong Jiang[2,*] Kangan Qian[1,*] Zilin Huang[3]
Tianze Zhu[1] Jiao Siwen[4] Jinyu Miao[1] Zheng Fu[1] Yang Zhong[5] Yunlong Wang[1]
Hao Ye[5] Mengmeng Yang[1,†] Kun Jiang[1,†,‡] Diange Yang[1,†]

Abstract— **Previous fast–slow system architectures demonstrated that pairing a reactive E2E planner with a deliberative vision-language model (VLM) can address these long-tail scenarios. However, these dual-system models that query the slow module at fixed intervals are computationally inefficient and introduce unnecessary latency during normal operation. To bridge this gap, we introduce FASIONAD, an adaptive fast–slow framework for autonomous driving that selectively integrates E2E planning and VLM reasoning. A lightweight fast planner manages general control, while a slow reasoner is activated only when a Laplace-based uncertainty gate detects changed uncertainty. Rather than overriding control, the VLM provides concise planning states and high-level plans. These inform the planner through an information bottleneck and high-level action guidance, enhancing interpretability and safety.**

Evaluated on the nuScenes, Bench2Drive, and CARLA Town05 closed-loop benchmarks, FASIONAD lowers the average trajectory error by 6.7% and the collision rate by 28.1% compared with strong E2E baselines, while also markedly reducing computational overhead relative to always-on fast–slow dual systems. These results demonstrate that adaptive fast–slow fusion is a practical route to safer, more reliable, and more efficient autonomous driving.

I. INTRODUCTION

As technology advances, autonomous driving holds the potential to transform transportation by enhancing efficiency, reducing human workload, and minimizing accidents [1]. Traditional autonomous driving systems typically follow a modular design consisting of perception, prediction, and planning [1], [2], [3]. Although such modular approaches provide interpretability, they can be rigid in nature (*e.g.*, rule-based controllers) and may struggle to handle complex, dynamic real-world scenarios [2]. In contrast, End-to-End (E2E) learning methods have recently gained attention, aiming to learn driving policies directly from sensory inputs [4], [5]. However, purely E2E models often exhibit insufficient generalization and reliability, especially in long-tail driving situations [6]. Attempts to refine E2E systems with a trajectory evaluation module often rely on open-loop evaluations (selecting trajectories without real-time feedback), making them susceptible to unforeseen failures.

[1]School of Vehicle and Mobility, Tsinghua University
[2]McGill University, Montreal, Quebec, Canada.
[3]University of Wisconsin-Madison, Madison, WI, USA.
[4]National University of SingaporeSingapore
[5]Xiaomi EV, Beijing, China.
Zilin Huang did not receive any funding for this work.
*Equal contribution. †Project Leader. ‡Corresponding authors

Recent progress in Large Language Models (LLMs) and Vision–Language Models (VLMs) has prompted the community to harness their multimodal understanding for autonomous driving, from sensor fusion [7] to commonsense reasoning [8]. Yet pure language–centric solutions face steep barriers: multi-billion–parameter models are expensive to fine-tune, require powerful GPUs at inference time, and may hallucinate spatial details [9]. A first attempt to merge reactive control with deliberative language reasoning is the fast–slow paradigm exemplified by DRIVEVLM [10], which queries a VLM alongside an end-to-end (E2E) planner. Because the slow module is polled at fixed frequency, however, such systems incur unnecessary latency and energy when the driving context is simple. Human drivers reason deeply only when the situation demands it. Most of the journey such as lane-keeping, steady car-following, relies on reflexive skills, while explicit deliberation is reserved for intersections, occlusions, or unpredictable agents. This observation raises a design question: *Can an autonomous vehicle enjoy the speed of an E2E planner yet summon VLM-level reasoning only when its own confidence degrades?*

We answer affirmatively with **FASIONAD**: an adaptive fast–slow framework that selects between an E2E *fast system* and a VLM *slow system* on demand (Fig. 1). A Laplace-based *Uncertainty Estimation* gate activates the slow module only when the fast planner's score distribution indicates risk. Crucially, the VLM never overwrites trajectories; instead, it emits compact *planning states* and *high-level plans*. An *Information Bottleneck* filters these signals, while *High-level Action* guidance injects them via cross-attention, refining the fast planner without violating real-time constraints. Visual and bird's-eye prompts further anchor the VLM's spatial reasoning, curbing hallucination. We evaluate **FASIONAD** on three public benchmarks—NUSCENES [11], TOWN05 SHORT [12], and BENCH2DRIVE [13]. Extensive experiments show that our selective fusion reduces trajectory error and collisions in both routine and long-tail scenarios, while cutting slow-system calls by more than 60 percent. Our contributions of this paper include:

- We propose **FASIONAD**, an adaptive Fast–Slow fusion framework for safe autonomous driving that calls the slow VLM system only when needed.
- We combine Uncertainty Estimation, Information Bottleneck filtering, and High-level Action guidance to deliver

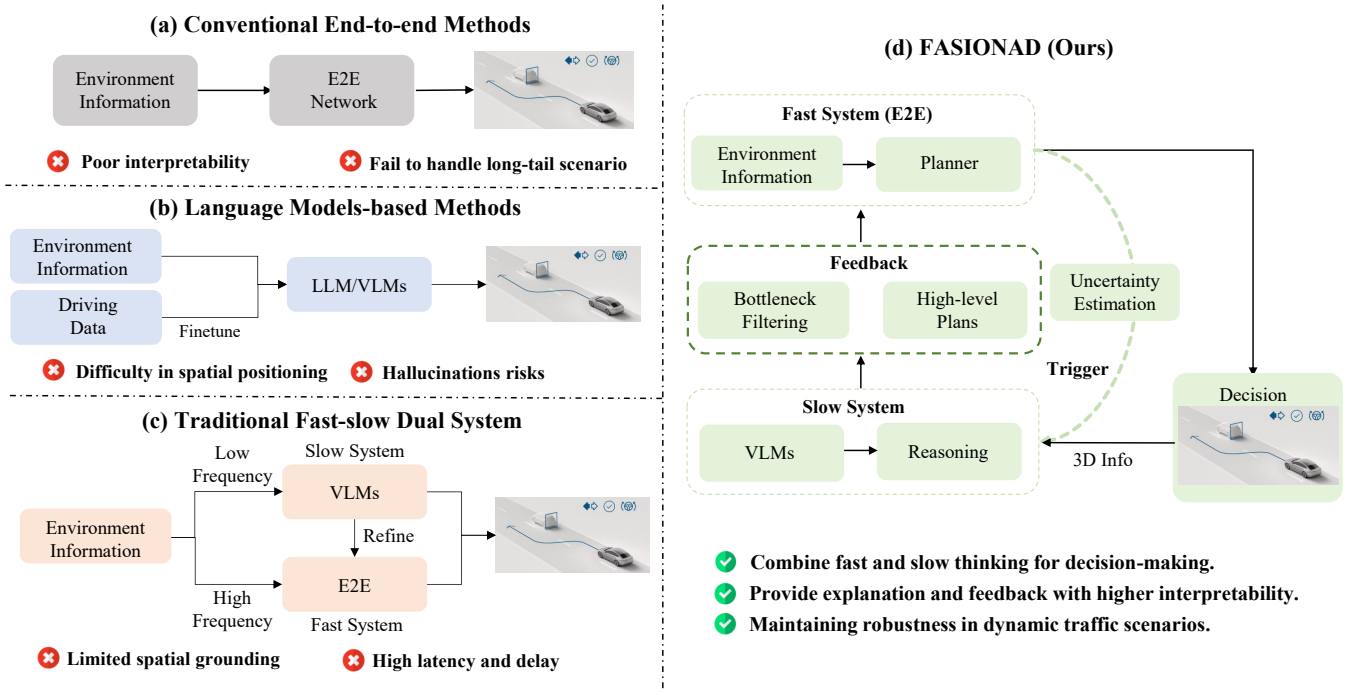

Fig. 1. The motivation of our FASIONAD. Conventional E2E methods struggle with interpretability and generalization. LLMs-based methods face slow decision-making, spatial positioning issues, and potential hallucinations. Traditional fast-slow dual system faces challenge of spatial grounding and latency.We compares different motion planning methods for autonomous driving, showcasing our method's ability to adaptive, context-aware decisions, offering better explanation and feedback.

compact, interpretable semantic feedback to the fast planner with negligible latency cost.

- We design Visual + BEV prompts plus lightweight fine-tuning to align and de-hallucinate VLM outputs for planning supervision.
- Extensive evaluations on nuScenes, Bench2Drive, and CARLA Town05 with exhaustive ablations, show that FASIONAD attains state-of-the-art performance across multiple safety metrics while delivering better real-time efficiency than conventional dual-system baselines.

II. Related Work

A. Learning-based Planning

While traditional modular autonomous driving pipelines provide interpretability [1], [4], they often struggle in complex scenarios due to restricted information flow between perception and planning [14].

End-to-end (E2E) approaches, on the other hand, learn direct mappings from sensory inputs to control signals [15], showing promising performance under routine conditions. Recent work extends E2E methods with Bird's-Eye-View (BEV) representations to handle complex urban contexts [5], [16], aiming to improve spatial awareness and decision-making. Despite these advances, purely E2E systems still suffer from limited interpretability and vulnerability to distributional shifts. Transformer-based solutions such as TransFuser [17] and InterFuser [18] have introduced multi-modal fusion and attention mechanisms, achieving more robust predictions in diverse traffic scenarios. Yet, balancing real-time performance with high-level reasoning remains an active area of research.

B. Vision-Language Models for Autonomous Driving

Vision-Language Models (VLMs) align visual and textual modalities to offer richer scene understanding [19], [20]. Foundational models demonstrate the potential for nuanced semantic representations, as showcased by Video-LLaVA [21], which support more detailed interpretations of dynamic driving scenarios. Beyond perception, VLMs have also been applied to high-level reasoning and planning in multi-agent contexts, enhancing robustness in environments with complex interactions[22].

Early efforts, including GPT-Driver [23], DriveGPT4 [24], EMMA[25] show that VLMs can process multi-view images and textual prompts to generate both trajectory plans (or low-level control commands) and human-readable rationales. While effective for commonsense reasoning and handling corner cases, integrating large foundation models into driving systems presents significant challenges: poor spatial awareness, ambiguous numerical outputs [26]. DriveVLM[10] is a pioneering model that combines VLMs with end-to-end architectures. It uses a chain-of-thought process to generate low-frequency trajectories, which are then fine-tuned by an end-to-end model to create the final planning outputs. In simpler scenarios, however, integrating VLMs can actually incur unnecessary latency and reduce model's speed. Building on these insights, our proposed **FASIONAD** framework harnesses a VLM not just for semantic extraction but also for feedback-driven decision refinement in uncertain or rare scenarios. By selectively engaging the VLM, we effectively mitigate common end-to-end pitfalls like hallucinations and poor generalization, all while preserving computational

efficiency for routine driving tasks.

III. Methodology

A. Overview

Our inspiration comes from the human driving process. In most simple scenarios, drivers don't engage in much conscious thought; they only dedicate more attention and focus to complex situations.

Hence we believe that VLMs are not required to provide continuous real-time guidance for planning like DriveVLM[10]. Their function should instead be concentrated on handling complex and difficult scenarios. As depicted in Fig. 2, FASIONAD employs a dual-system architecture: a fast system for rapid, real-time responses, and a slow system for comprehensive analysis and complex decision-making in uncertain or challenging driving scenarios. The fast system encodes image information into tokens, generating multi-modal trajectories along with a score for each trajectory (Section III-B). In contrast, the slow system(Section III-C) processes the BEV prompt and visual prompt, subsequently outputting planning states and high-level plans for the entire driving scenario.

To ensure a seamless collaboration between the fast and slow systems, we developed an innovative switching mechanism based on uncertainty estimation. This approach allows the fast system to refine its trajectory predictions by leveraging an information bottleneck and integrating high-level plans derived from the slow system (Section III-D).

B. Fast System

Waypoints Prediction. Given a set of N multi-view images $\mathbf{I}_t = \{I_t^1, I_t^2, \ldots, I_t^N\}$ and high-level navigation commands $\mathbf{C}_t$, the model generates N_K candidate trajectories $\mathbf{T}_t = \{T_i\}_{i=1}^{N_T}$, where each trajectory $T_i \in \mathbb{R}^{\text{bs} \times T_s \times 2}$ represents a sequence of waypoints over a time horizon T_s. Here, N_K is the top-K sampled multi-modal trajectories. This system can be formulated as:

$$\text{FASIONAD (fast system):} \quad (\mathbf{I}_t, \mathbf{C}_t) \rightarrow \mathbf{T}_t. \tag{1}$$

Score Evaluation. To ensure an efficient transition between our fast and slow systems, the fast system employs a scoring model $\mathscr{F}_{\text{Score}}$ to assign a scalar score s_i to each trajectory T_i. This score serves as a critical metric for estimating uncertainty, thereby informing the switching mechanism.

Specifically, inspired by HE-Drive[27], the scoring model $\mathscr{F}_{\text{Score}}$ is formulated as a composite function that integrates several key factors, including safety, comfort considerations, and we further take efficiency, and economic factors into consideration:

$$\begin{aligned} \mathscr{F}\text{Score} = {} & \alpha_{\text{safety}} S_{\text{safety}} + \alpha_{\text{comfort}} S_{\text{comfort}} \\ & + \alpha_{\text{efficiency}} S_{\text{efficiency}} + \alpha_{\text{economic}} S_{\text{economic}} \end{aligned} \tag{2}$$

where the coefficients α_{safety}, α_{comfort}, $\alpha_{\text{efficiency}}$, and α_{economic} are weights that balance the auxiliary losses.

The comfort score function takes into account the lateral, longitudinal, and centripetal accelerations of the ego vehicle. The safety score is an aggregation of four key metrics: minimum distance to obstacles, distance between end and target positions, cumulative deviation from the target heading, and deviation from the target speed.

C. Slow System

In complex scenarios, accurate interpretation of environmental factors is vital for safe decision-making. The slow system emulates human-like reasoning to infer context and predict future actions, similar to human drivers. This section discusses how VLMs can support such reasoning, with a focus on QA design in Section III-C.1, which formats the output of VLM models, and VLM tuning in Section III-C.2.

1) Planning-oriented QA: Building on existing QA frameworks for autonomous driving [10], we propose a structured approach aimed at human-like reasoning. As illustrated in Fig. 3, our design centers on five key aspects critical to robust driving policies:

(i) **Scene analysis:** Evaluates environmental conditions (e.g., weather, lighting, traffic density) to guide overall decision-making.
(ii) **Traffic sign recognition:** Detects and interprets traffic signs for regulatory compliance.
(iii) **Key object recognition:** Identifies and predicts the behavior of nearby objects, aiding hazard anticipation.
(iv) **Planning state:** Encodes driving context as K-dimensional binary vectors $\mathbf{Y}t$, derived through *Yes*/*No* queries. This representation helps prioritize actions and optimize routing.
(v) **High-level planning and justification:** Decomposes driving decisions into meta-actions, which are mapped by a learnable encoder EA into features $\mathbf{A}_t$. This modular design supports flexible, constraint-aware planning.

Each QA task refines the vision-language model's understanding of the scene, ensuring both lower-level perception and higher-level planning remain adaptive and interpretable.

We feed the planning state and meta-action features into the fast system, creating a human-like decision-making loop. Additionally, we introduce two prompts to enhance QA: (i) a visual prompt for human-like interpretation of scene elements, and (ii) a BEV prompt for a top-down perspective of spatial relationships.

Visual Prompt: In typical autonomous driving systems, waypoints generated by high-level planners are numerical outputs [5], [28]. However, VLMs are not inherently designed to process numerical data in this context. Human decision-making in complex driving scenarios relies more on intuitive reasoning and visual cues than on direct numerical computation. To bridge this gap, we integrate trajectory visual prompts into our slow system planning. Specifically, we project the waypoints generated by the fast system planner onto the front-view camera, creating a visual representation of the trajectory, $\mathbf{V}_t^f$. This visual approximation of the planned path facilitates human-like reasoning processes, enabling more intuitive evaluation and modification of decisions, which leads to more reliable and effective high-level plans.

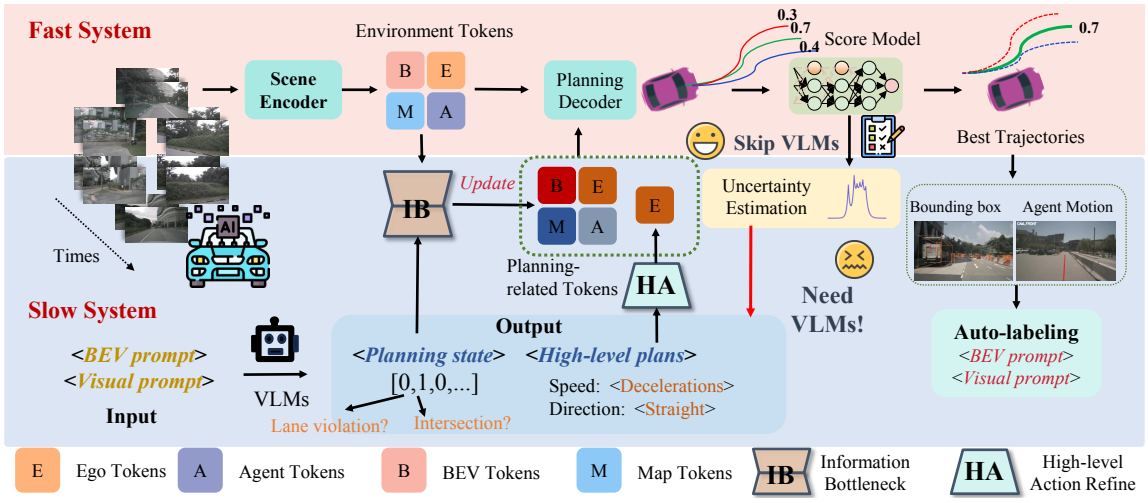

Fig. 2. The framework operates through dual-system: fast and slow. The fast ststem encodes image information into instance tokens(E, B, M, A relatively denotes ego tokens, BEV tokens, map tokens and agent tokens), generating multi-modal trajectories via a planning head. A score model selects the optimal trajectory, while uncertainty estimation determines slow system activation. When engaged, the slow system utilizes VLM feedback, which is integrated both as High-level Action(HA) and as scene-derived planning state vectors by Information Bottleneck(IB), enabling trajectory refinement through the planning head.

BEV Prompt: To further enhance the system's spatial understanding, we introduce a BEV prompt. Based on the vehicle's BEV coordinate system, this prompt provides a clear depiction of spatial relationships between the ego vehicle and surrounding agents, represented as $\mathbf{B}_t$.

In summary, The slow system pipeline can be formulated as follows:

$$\mathbf{P}_t, \mathbf{A}_t = \Phi(E(\mathbf{V}_t^f), E(\mathbf{B}_t)) \tag{3}$$

2) VLM Tuning: To adapt the VLM for QA-based planning, we combine an auto-labeled dataset with a reward-guided training scheme:

QA Dataset Generation: We automatically annotate 3D detection boxes and tracked trajectories from the fast system, then leverage VLMs such as QwenVL [29] to produce descriptive QA pairs that align with the observed scene.

Reward-Guided VLM Tuning: Unlike standard LLM approaches reliant on pure auto-regressive learning, we incorporate both Maximum Likelihood Estimation (MLE) loss and a reward-guided regression loss. Inspired by, but distinct from, GPT Driver [23], our method uses automatically generated guidance to replicate the planning state and high-level plans. Additionally, we integrate Proximal Policy Optimization (PPO) [30] with masking to apply supervision at the token level, while treating the entire sequence as meaningful for regression. Concretely, we compute:

$$\mathscr{L}_{\text{rvlm}} = \mathscr{F}_{\text{Reward}}\left(\mathbf{s}^{1:T_i}\right) \cdot \Phi\left(\mathbf{s}^{T_i} \,\middle|\, \mathbf{s}^{1:T_i-1}\right), \tag{4}$$

where $\mathbf{s}^{T_i}$ is the predicted token, $\mathscr{F}_{\text{Reward}}(\cdot)$ evaluates trajectories, and $\Phi(\cdot|\cdot)$ represents the policy. The final training objective combines the standard language loss and the reward-guided term:

$$\mathscr{L}_{\text{slow}} = \lambda_{\text{MLE}}\mathscr{L}_{\text{MLE}} + \lambda_{\text{rvlm}}\mathscr{L}_{\text{rvlm}}. \tag{5}$$

D. FAst-Slow Collaborative Framework

Our collaborative Fast-Slow framework is designed to enhance final trajectory accuracy by integrating information at both the feature and trajectory levels. This dual-level approach leverages the complementary strengths of a VLM and an efficient E2E model: the E2E model is responsible for handling routine scenarios, while the VLM is brought in to provide high-level reasoning for more complex situations.

The collaboration between these systems is seamlessly managed by a switching mechanism based on uncertainty estimation. This mechanism enables the fast system to enhance its predictions at two key levels: at the feature level via an Information Bottleneck, and at the semantic action level through High-level Action Guidance from the slow system.

Uncertainty Estimation: To effectively navigate dynamic and unpredictable environments, estimating uncertainty in waypoint predictions is essential, as it allows the system to adapt its decision-making based on prediction reliability. To handle outliers and model uncertainty in waypoint predictions, we employ a Laplace distribution:

$$p(\mathrm{S} \mid \Theta) = \prod_{t=1}^{T} \frac{1}{2b} \exp\left(-\frac{\|\mathbf{s}_t - \hat{\mu}_t\|_1}{b}\right) \tag{6}$$

Where $\hat{\mu}_t$ denotes the expectation of predicted score at time t, b is the scale parameter, S is the score, and Θ represents the model parameters. The Laplace distribution's heavy tails and sharp peak make it robust to outliers and effective for uncertainty estimation in dynamic driving environments. The system uses the fast mode for planning when score S surpasses a set threshold with low uncertainty, and switches to the slow mode for detailed analysis in all other cases.

Information Bottleneck: Driving environments often contain irrelevant or noisy information that does not contribute to planning. To address this, we apply the Information Bottleneck(IB) principle to distill the information relevant

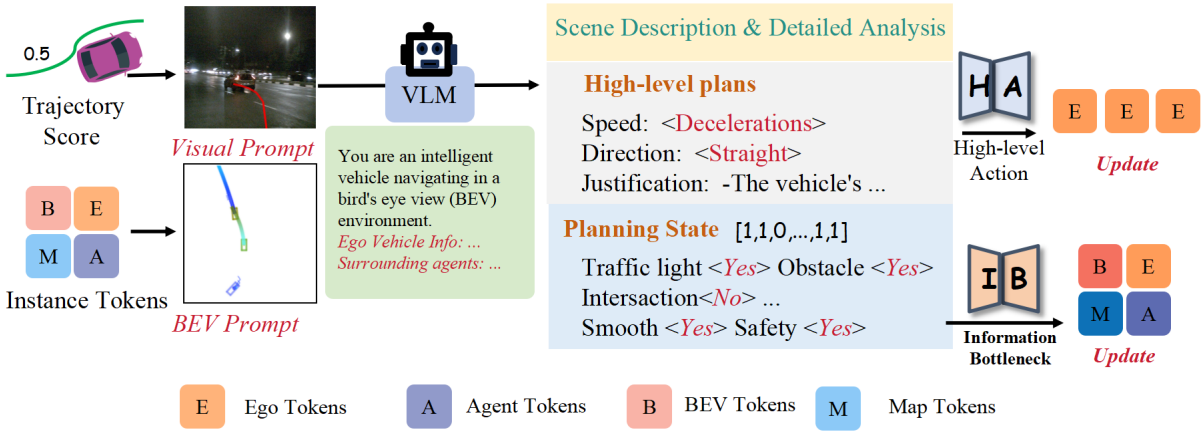

Fig. 3. The adaptive feedback mechanism integrates dual inputs - visual prompts and BEV prompts - into a VLM. This VLM produces three outputs: scene descriptions, detailed analyses, and high-level plans, along with planning state vectors that encapsulate scene conditions. High-level plans are embedded into ego tokens, whereas planning state vectors pass through an IB to refine environment information in query tokens.

to decision-making. Through interaction with environmental information, we derive the features of ego vehicle, denoted as z. To align the learned planning-relevant representation z with the planning state y_t, we employ the MLP layers that maps z to a one-dimensional vector y_i. The knowledge distillation process minimizes the following objective:

$$\mathscr{L}_{\text{KD}} = \sum \log q_d(y_t|y_i) - \beta \, \text{KL}\left(q_e(y_i|z_{\text{current}}) \| p(z)\right) \quad (7)$$

where $q_d(y_t|y_i)$ is the probability distribution over the VLM-derived vector y_t given y_i, and $q_e(y_i|z_{\text{current}})$ encodes query features from the current state. Here, $p(z)$ is a prior distribution on z, and β is a regularization parameter.

High-level Action Guidance: To enhance the fast system's trajectory predictions, we utilize a cross-attention mechanism to integrate high-level action guidance (HA) from the slow system. Specifically, we use the ego token $e_{\text{ego}} \in \mathbb{R}^{d_A}$ queries the learnable embeddings $E_A \in \mathbb{R}^{N_A \times d_A}$, which serve as key-value pairs.

IV. Experiments

In this section, we conduct experiments to address the following questions: (1) *Does our feedback mechanism improve the planning performance of the fast E2E model?* (2) *How does the uncertainty estimation meets the needs of handling complex driving scenarios?* (3) *Do our information bottleneck and high-level plan instructions enhance the planning process?* (4) *Does the VLM equipped with "visual and BEV prompts" provide a reasonable planning process?*

A. Experimental Setup

Benchmarks: We assessed FASIONAD's capabilities through a series of rigorous benchmarks, evaluating both its open-loop and closed-loop performance across diverse real-world and simulated environments.

Open-loop evaluation utilized the nuScenes [11] and the more recent Bench2Drive (B2D) benchmark [13]. We measure trajectory prediction accuracy against expert demonstrations using *L2* distance and collision rate metrics. Closed-loop performance is evaluated on CARLA Town05 Short Benchmark. The benchmrak measures Driving Score (DS).

Implementation Details: We constructed our fast system by leveraging VAD [5] and GenAD[16], with its training configuration aligning with that of VAD. For the VLMs, we incorporated Qwen-VL-7B[29], InternVL1.5-7B[38], and Video-LLaVA[21]. All experiments were carried out on a server equipped with 8 NVIDIA A100 GPUs.

For the Supervised Fine-Tuning (SFT) of the VLMs, we applied a learning rate of 1×10^{-4} and trained for 20 epochs. To efficiently adapt the models, we utilized a LoRA adapter, setting its rank to 8 and alpha to 32.

B. Main Results

1) Open-loop Evaluation: **nuScenes Performance:**

We evaluate FASIONAD against representative fast , slow , and dual-system baselines. As shown in Tab. I, FASIONAD consistently achieves state-of-the-art performance across all horizons, with a mean L2 error of 0.28m and a 0.09% collision rate. Compared to DriveVLM , our framework reduces average L2 error by 9.6% and collision rate by 10%. Notably, integrating our adaptive mechanism into GenAD and VAD-Base yields substantial gains, reducing L2 errors by 24.2% and 18.8%, and collision rates by 58.1% and 49.1%, respectively. These results validate that adaptive switching significantly enhances planning accuracy and safety across diverse base planners.

To better illustrate the effectiveness of the switching mechanism, Fig. 4 gives examples where FASIONAD successfully adapts to complex planning tasks. When approaching an intersection, the system dynamically adjusts its trajectory. During a lane change, it activates the slow system, considering the prevailing road conditions to select a more efficient and optimal route. On highways, it assesses traffic conditions and selectively activates the slow system for safe and efficient lane changes. By integrating a structured planning state and

TABLE I

OPEN-LOOP PLANNING PERFORMANCE ON THE NUSCENES VALIDATION DATASET

Method	Input	System	L2 (m) ↓				Collision Rate (%) ↓				FPS
			1s	*2s*	*3s*	*Avg*	*1s*	*2s*	*3s*	*Avg*	
IL [31]	LiDAR	-	0.44	1.15	2.47	1.35	0.08	0.27	1.95	0.77	-
NMP [32]	LiDAR	Fast.	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34	-
FF [33]	LiDAR	Fast.	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43	-
EO [34]	LiDAR	Fast.	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33	-
ST-P3 [35]	Camera	Fast.	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	1.6
OccNet [36]	Camera	Fast.	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72	2.6
UniAD [28]	Camera	Fast.	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31	1.8
Agent-Driver [37]	Camera	Slow.	0.22	0.65	1.34	0.74	0.02	0.13	0.48	0.21	0.8†
EMMA[25]	Camera	Slow.	0.14	0.29	0.54	0.32	-	-	-	-	-
DriveVLM* [10]	Camera	Dual.	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10	5.1†
FASIONAD* w/ GenAD	Camera	Dual.	**0.13**	**0.26**	**0.45**	**0.28**	**0.05**	**0.08**	**0.15**	**0.09**	**5.8**†
VAD-Tiny [5]	Camera	Fast.	0.60	1.23	2.06	1.30	0.31	0.53	1.33	0.72	6.9
VAD-Base [5]	Camera	Fast.	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53	3.6
FASIONAD w/ VAD-Base	Camera	Dual.	0.41	0.95	1.62	0.99	0.03	0.15	0.62	0.27	3.2†
GenAD [16]	Camera	Fast.	0.25	0.68	1.30	0.74	0.05	0.17	0.53	0.25	6.7
FASIONAD w/ GenAD	Camera	Dual.	**0.19**	**0.62**	**1.25**	**0.69**	**0.02**	**0.09**	**0.44**	**0.18**	**5.8**†

Note: * denotes using ego status features as input. † represents that the metrics are computed with an average of all the predicted frames. † denotes FPS measured in the same environment on our machine with a single RTX 4090 GPU.

high-level plans with adaptive feedback, the system improves decision-making interpretability and planning safety.

Fig. 4. Example scenarios demonstrating FASIONAD's adaptive feedback framework in various driving environments. Each scene shows different navigation challenges, including obstacles, lane adjustments, and turns. The proposed system provides suggested driving operations and ensures safe, smooth trajectories with minimal abrupt maneuvers, enhancing safety in complex situations.

Bench2Drive Performance: Tab.II shows comparison results with several well-established E2E methods. FASIONAD achieves an L2 error of 0.82m and a collision rate of 0.12%, demonstrating notable improvements over VAD (0.91m, 0.19%). While UniAD-Base achieves a slightly lower L2 error (0.73m), its reliance on deterministic trajectory generation without explicit uncertainty modeling may lead to increased safety risks in real-world deployment. Compared to AD-MLP, which has a much higher *L2* error of 3.64m, our method benefits from its adaptive feedback mechanism, improving both accuracy and safety. These results highlight the effectiveness of FASIONAD's feedback-driven adaptation, where high-level vision-language reasoning complements fast trajectory generation, leading to more precise and safer predictions across diverse traffic scenarios.

2) Closed-loop evaluation: To validate FASIONAD's driving skills in closed-loop evaluations, we compare our proposed FASIONAD with a variety of published algorithms.

TABLE II

COMPARISON OF METHODS BASED ON BENCH2DRIVE BENCHMARK OPEN-LOOP EVALUATION

Method	Avg. L2 ↓	Avg. C.R. ↓
AD-MLP [39]	3.64	-
UniAD-Tiny [28]	0.80	-
UniAD-Base [28]	**0.73**	-
VAD [5]	0.91	0.19
FASIONAD w/ VAD-Base	0.82	**0.12**

Note: Avg. L2 is calculated similarly to the UniAD.

TABLE III

CLOSED-LOOP EVALUATION ON TOWN05 SHORT BENCHMARK.

Methods	Modality	DS (%) ↑	RC (%) ↑
CILRS [40]	C	7.47	13.40
LBC [41]	C	30.97	55.01
Transfuser [17]	C+L	54.52	78.41
ST-P3 [35]	C	55.14	86.74
VAD [5]	C	64.29	87.26
Agent-Driver [37]	C	64.31	87.31
FASIONAD w/ GenAD	C	**64.83**	**89.04**

Tab. III presents a comparative analysis against state-of-the-art E2E autonomous driving models such as multi-modal based Transfuser [17], query-based VAD [5], and LLM-based methods Agent-Driver [37]. FASIONAD achieves the highest DS (64.83%) and RC (89.04%), surpassing prior methods in both driving stability and route-following accuracy. These results demonstrate that our approach not only improves planning accuracy in open-loop settings but also enhances overall driving performance in interactive scenarios.

3) Explainability and reliability in planning states and high-level plans: Since FASIONAD integrates VLMs to enhance trajectory planning, we conduct experiments following the RAG-Driver [14] setup to quantitatively analyze different models' performance on planning state recognition, high-level

TABLE IV

COMPARISON OF VLMS IN PLANNING-ORIENTED TASKS.

Method	Plan. S. Acc↑	High. A. Acc↑	BLEU.↑	CID.↑	MET.↑
QwenVL [29]	15.13	9.75	9.45	6.22	32.11
InternVL [38]	**17.92**	**10.18**	**12.62**	**7.32**	**35.68**
Video-LLaVA [21]	8.46	8.14	8.85	4.19	31.92
QwenVL [29]†	52.33	61.87	24.77	20.13	48.25
InternVL [38]†	**56.81**	62.45	23.41	**20.84**	48.19
Video-LLaVA [21]†	55.74	**62.85**	**25.34**	19.53	**50.48**

+ **Note:** The †denotes configurations equipped with task-specific prompts using our proposed planning-oriented QAs. Plan. S. and High. A. relatively denote planning states and high-level actions.

TABLE V

ABLATION STUDY OF IB AND HA.

Setting		L2 (m) ↓				Collision Rate (%) ↓			
IB	HA	1s	2s	3s	Avg.	1s	2s	3s	Avg.
✓	✗	0.21	0.63	1.27	0.71	0.03	0.12	0.47	0.21
✗	✓	0.24	0.68	1.37	0.77	**0.02**	0.10	0.45	0.19
✓	✓	**0.19**	**0.62**	**1.25**	**0.69**	**0.02**	**0.09**	**0.44**	**0.18**

action prediction, and explanation quality (BLEU-4, CIDEr, METEOR), as shown in Table IV. Results show that task-specific prompts significantly improve all models, with Video-LLaVA achieving the highest accuracy (55.74% planning state, 62.85% high-level action) and best explainability (25.34 BLEU-4, 50.48 METEOR). While InternVL and QwenVL also perform well, their improvements are less pronounced. The substantial performance gap between standard and task-specific prompts highlights the importance of structured input, aligning with FASIONAD's approach of integrating VLM-guided reasoning to enhance planning accuracy and interpretability.

TABLE VI

ABLATION STUDY OF UNCERTAINTY MODULE.

Setting		L2 (m) ↓				Collision Rate(%) ↓				VLM Trigger Rate	
Asyn.	Uncer.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	Fast	Slow↓
✓	✗	**0.17**	**0.60**	**1.22**	0.66	0.02	0.08	0.45	0.18	76.92%	23.08%
✗	✓	0.19	0.62	1.25	0.69	**0.02**	0.09	**0.44**	**0.18**	91.26%	**8.74%(↓62.13%)**

C. Ablation Study

In this section, we implement FASIONAD without ego-state to purely evaluate the components. The fast E2E model used is GenAD [16].

1) Modular designs: Our ablation study demonstrates the complementary benefits of the Information Bottleneck(IB) and High-level Action(HA) components (Tab. V). The full model incorporating both components achieved the best performance (L2: 0.69m, collision rate: 0.18%). Using either component alone led to decreased performance - IB-only (L2: 0.74m, collision rate: 0.21%) and HA-only (L2: 0.77m, collision rate: 0.19%) - highlighting their synergistic relationship in improving prediction accuracy through effective information filtering and high-level planning. To assess the impact of the uncertainty estimation mechanism, we conduct an ablation study comparing two setups in Tab.VI: (1) triggering the fast-slow systems asynchronously, and (2) incorporating uncertainty estimation. With the uncertainty switch, planning performance remains stable while reducing computational load. Specifically, the VLM trigger rate decreases by 62.13% compared to asynchronous methods (e.g., DriveVLM[10]).

TABLE VII

VALIDATION OF VLM PROMPT STRATEGIES

Setting	L2 (m) ↓				Collision Rate (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Simple.P	0.31	0.71	1.38	0.80	0.05	0.16	0.74	0.32
BEV.P	0.29	0.70	1.36	0.79	0.04	0.14	0.65	0.24
Visual.P	0.24	0.67	1.30	0.74	**0.02**	0.11	0.48	0.20
Full.P	**0.19**	**0.62**	**1.25**	**0.69**	**0.02**	**0.09**	**0.44**	**0.18**

Note: In the setting, "P" denotes a prompt (e.g., BEV.P indicates a BEV prompt).).

2) VLM prompt strategy: Our ablation study on VLM prompt strategies revealed the significant impact of prompt design (Tab. VII). The Full.P configuration, featuring comprehensive prompt instructions, achieved the best results with an L2 distance of 0.69 meters and 0.18% collision rate. Performance gradually declined with simpler prompting approaches: Visual.P (0.74m, 0.20%), BEV.P (0.79m, 0.24%), and Simple.P (0.80m, 0.32%). These results demonstrate that detailed, well-structured prompts are crucial for maximizing VLM's predictive capabilities.

D. Real-time performance analysis

As shown in Table I, on an RTX 4090, DriveVLM runs at 5.10 FPS (≈196.08 ms/frame) whereas FASIONAD attains 5.80 FPS (172.41 ms/frame), a 12.0% mean latency reduction (+13.7% FPS). Thus, uncertainty-gated sparse Slow calls amortize multimodal reasoning over only difficult frames, improving throughput without sacrificing quality.

V. CONCLUSION

We proposed **FASIONAD**, an adaptive fast–slow fusion autonomous driving framework where information bottleneck representations and high-level plans guide the fast planner, and visual/BEV prompts plus lightweight fine-tuning stabilize the slow module. Experiments on nuScenes, Bench2Drive, and CARLA closed-loop driving show consistent trajectory improvements and collision reduction at lower computational cost than prior dual-system baselines.

VI. ACKNOWLEDGEMENT

Work done during RIMBOT's internship. This work was supported in part by the National Natural Science Foundation of China under Grants 52372414, 52394264, 52402499, and 52472449; the Beijing Natural Science Foundation under Grants L231008 and L243008. The authors also would like to acknowledge the support from the Tsinghua University-Toyota Joint Research Center and the Tsinghua University-Didi Joint Research Center.

REFERENCES

[1] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2021.

[2] W. Zhou, Z. Cao, N. Deng, X. Liu, K. Jiang, and D. Yang, "Dynamically conservative self-driving planner for long-tail cases," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3476–3488, 2022.

[3] T. Shi, P. Wang, X. Cheng, C.-Y. Chan, and D. Huang, "Driving decision and control for automated lane change behavior based on deep reinforcement learning," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 2895–2900.

[4] S. Jiang, S. Choi, and L. Sun, "Communication-aware reinforcement learning for cooperative adaptive cruise control," *arXiv preprint arXiv:2407.08964*, 2024.

[5] 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 (ICCV)*, 2023, pp. 8340–8350.

[6] H. X. Liu and S. Feng, "Curse of rarity for autonomous vehicles," *nature communications*, vol. 15, no. 1, p. 4808, 2024.

[7] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.

[8] J. Wang, G. He, and Y. Kantaros, "Safe task planning for language-instructed multi-robot systems using conformal prediction," *arXiv preprint arXiv:2402.15368*, 2024.

[9] R. Tan, S. Lou, Y. Zhou, and C. Lv, "Multi-modal llm-enabled long-horizon skill learning for robotic manipulation," in *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, 2024, pp. 14–19.

[10] 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.

[11] 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 (CVPR)*, 2020, pp. 11 621–11 631.

[12] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*. PMLR, 2017, pp. 1–16.

[13] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," *arXiv preprint arXiv:2406.03877*, 2024.

[14] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv:2402.10828*, 2024.

[15] F. Codevilla, A. M. Lopez, and et al., "End-to-end driving via conditional imitation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1–10.

[16] W. Zheng, R. Song, X. Guo, C. Zhang, and L. Chen, "Genad: Generative end-to-end autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–104.

[17] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *arXiv preprint arXiv:2205.15997v1*, 2022.

[18] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," *arXiv preprint arXiv:2207.14024*, 2023. [Online]. Available: https://arxiv.org/abs/2207.14024

[19] S. Jiang, Z. Huang, K. Qian, Z. Luo, T. Zhu, Y. Zhong, Y. Tang, M. Kong, Y. Wang, S. Jiao *et al.*, "A survey on vision-language-action models for autonomous driving," *arXiv preprint arXiv:2506.24044*, 2025.

[20] K. Qian, S. Jiang, Y. Zhong, Z. Luo, Z. Huang, T. Zhu, K. Jiang, M. Yang, Z. Fu, J. Miao *et al.*, "Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving," *arXiv preprint arXiv:2505.15298*, 2025.

[21] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122v2*, 2023.

[22] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2023.

[23] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv preprint arXiv:2310.01415*, 2023.

[24] 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.

[25] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[26] 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.

[27] J. Wang, X. Zhang, Z. Xing, S. Gu, X. Guo, Y. Hu, Z. Song, Q. Zhang, X. Long, and W. Yin, "He-drive: Human-like end-to-end driving with vision language models," *arXiv preprint arXiv:2410.05051*, 2024.

[28] 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 (CVPR)*, 2023.

[29] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[31] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 729–736.

[32] 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*, 2019, pp. 8660–8669.

[33] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan, "Safe local motion planning with self-supervised freespace forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 732–12 741.

[34] T. Khurana, P. Hu, A. Dave, J. Ziglar, D. Held, and D. Ramanan, "Differentiable raycasting for self-supervised occupancy forecasting," in *European Conference on Computer Vision*. Springer, 2022, pp. 353–369.

[35] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[36] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8406–8415.

[37] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2024. [Online]. Available: https://arxiv.org/abs/2311.10813

[38] 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*, 2024, pp. 24 185–24 198.

[39] 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.

[40] F. Codevilla, E. Santana, A. M. Lopez, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[41] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.