SIC-DRIVE: SPARSE INSTRUCTION CONDITIONING FOR CLOSED-LOOP END-TO-END DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving has achieved significant progress by learning a direct mapping from sensory observations to motion planning. However, future autonomous systems require not only the ability to drive autonomously, but also the ability to adapt their behaviors according to human intentions in open and dynamic environments. Natural language provides a flexible interface for specifying high-level driving objectives, such as safety-related interventions, and temporary behavioral adjustments, which are difficult to infer solely from visual observations. Despite recent advances in language-conditioned driving, incorporating language inputs does not necessarily guarantee instruction-controlled behaviors. Existing approaches often focus on integrating language representations with driving policies or generating language-conditioned trajectories, while lacking explicit mechanisms that connect human intent with planning decisions. As a result, it remains unclear whether and how language instructions actually affect closed-loop driving behaviors. In this work, we study language-conditioned controllability for autonomous driving, aiming to enable driving systems to reliably translate natural-language instructions into behavioral changes. We propose SIC-Drive, an explicit instruction grounding framework that decomposes driving instructions into lateral and longitudinal semantics and injects them into trajectory planning through path and velocity candidate scoring. Furthermore, we construct a language-augmented Bench2Drive dataset with temporally aligned driving instructions and introduce DIF-Benchmark to evaluate whether driving agents can follow instructions beyond conventional driving performance metrics. Extensive closed-loop experiments demonstrate the effectiveness of our method, highlighting the importance of explicit language grounding for controllable autonomous driving. Code will be found in our Webpage after the paper is accepted.

1 INTRODUCTION

Recent progress in end-to-end autonomous driving has brought perception, prediction, and motion planning into unified learned systems (Jiang et al., 2023; Sun et al., 2026). These systems have achieved strong closed-loop driving performance by learning to plan from sensor observations and navigation inputs (Hu et al., 2023; Sun et al., 2024). In the usual setting, expert trajectories provide demonstrated behaviors (Caesar et al., 2022; Nguyen et al., 2026), while a route specifies the navigation goal. However, the same traffic scene can admit multiple safe and feasible behaviors (Chai et al., 2019; Phan-Minh et al., 2020), and the desired one may depend on an intention supplied during driving (Cui et al., 2024; Sha et al., 2025). A driver may request a particular turn, a safe lane change, or a speed reduction near a temporary hazard. Such requests are not always fully specified by the original route. Future driving systems therefore need not only to complete the driving task, but also to select among feasible behaviors according to human intentions.

Existing language-conditioned driving methods have demonstrated that such intentions can be incorporated into closed-loop driving. Recent VLM (Renz et al., 2025) and VLA (Wang et al., 2026c) based methods have already explored in closed-loop systems that align language with actions. Nevertheless, the presence of language input alone does not establish that the resulting behavior depends on the instruction. When each training scene is associated with a single demonstrated trajectory, observations and navigation cues may already predict the target behavior, allowing imitation loss to be minimized with limited dependence on language. Thus, given the same scene, an instruction such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

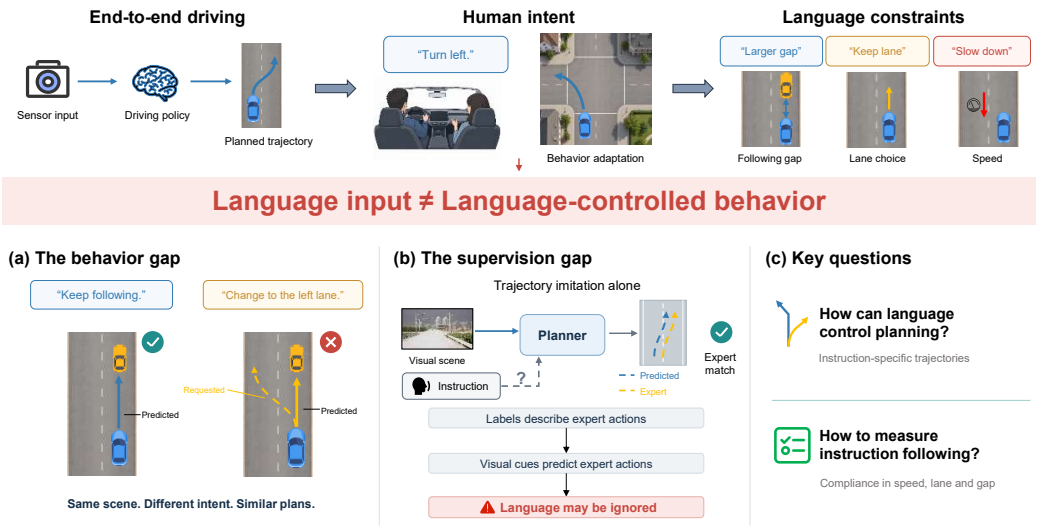


Figure 1: Motivation. Natural language provides a flexible interface for adapting driving behavior beyond predefined navigation.(a) **Behavior gap**: different instructions may express different driving intents, yet an end-to-end planner can still produce similar trajectories if language is weakly reflected in the planning process. (b) **Supervision gap**: trajectory imitation primarily encourages matching expert actions, while the effect of language on behavior may remain implicit and can be underutilized when visual cues already predict the expert trajectory. (c) These gaps motivate two key questions.

as "turn right at the intersection" may still result in the default straight-going behavior. Figure 1 illustrates the distinction between reaching a feasible driving outcome and reaching the outcome specified by an instruction.

For an existing end-to-end planner, language can intervene at different stages of planning. NudgeVAD (Yang et al., 2026), for example, preserves a pretrained planner and uses language to predict a residual correction to its output trajectory. In contrast, candidate-based planners (Li et al., 2024; Yao et al., 2026) provide another possible interface: language can modify the scores of feasible candidates before the final trajectory is selected. We investigate this as a hypothesis rather than assuming one interface is inherently superior: can instruction-conditioned candidate scoring reliably influence which closed-loop behavior is selected?

Answering this question also requires an instruction-sensitive evaluation protocol. A predicted trajectory that is consistent with an instruction does not necessarily imply that the vehicle will complete the requested behavior after continuous closed-loop control. MALT-Drive (Lu et al., 2026) already studies same-scene counterfactual instructions and evaluates instruction–action consistency from predicted path, speed, and executability. ICR-Drive (Hamid et al., 2026) instead studies robustness to language variations on matched CARLA routes using route-level driving metrics. Complementary to these settings, we focus on whether a feasible instructed target is actually completed in closed-loop execution. This requires checking the realized exit, lane, or speed behavior under matched decision scenes, while reporting driving safety separately.

To study these two aspects together, we propose SIC-Drive and DIF-Benchmark: SIC-Drive injects external natural-language instructions into candidate selection, while DIF evaluates whether the selected behavior is ultimately executed as instructed. Built on SparseDriveV2 (Sun et al., 2026), SIC-Drive decomposes instructions into lateral and longitudinal semantics and uses them to adjust path and speed candidate scores before trajectory selection. We further augment Bench2Drive-Base (Jia et al., 2024) with temporally aligned navigation and risk warnings to provide language supervision for instruction-conditioned planning.

DIF-Benchmark is implemented in CARLA (Dosovitskiy et al., 2017). At matched decision scenes, it assigns different feasible instructions and evaluates whether the executed closed-loop behavior

reaches the corresponding target, while measuring safety separately. On DIF-Benchmark, SIC-Drive completes 82 of 136 instruction episodes (60.29%). On standard Bench2Drive, it achieves a Driving Score of 86.96, compared with 86.03 for baseline. The modest driving-score difference and the remaining instruction failures underscore the need to measure driving performance and instruction execution separately.

Our contributions are summarized as follows: 1)We propose SIC-Drive, an explicit language grounding framework that connects natural-language instructions with lateral and longitudinal planning decisions. 2) We construct a language-augmented Bench2Drive-Base with temporally aligned navigation and risk warnings. 3) We introduce DIF-Benchmark, a matched-scene closed-loop evaluation protocol that measures whether instructed driving targets are actually completed, separately from driving safety.

2 Related Work

Planning-centric end-to-end driving. TransFuser Prakash et al. (2021) employed Transformers (Vaswani et al., 2017) to fuse features from cameras and LiDAR.; LAV (Chen & Krähenbühl, 2022) added supervision from surrounding vehicles; and TCP (Wu et al., 2022) jointly modeled planned trajectories and control. UniAD (Hu et al., 2023) connected perception, prediction, and planning, while VAD (Jiang et al., 2023) and SparseDrive (Sun et al., 2024) use more compact scene representations. SparseDriveV2 (Sun et al., 2026) factorizes trajectory candidates into path and velocity components and SIC-Drive uses this factorization as the point where lateral and longitudinal instructions affect candidate ranking. More recently, MeanFuser Wang et al. (2026b) modeled the multimodal trajectory space with continuous Gaussian mixture noise and adaptively reconstructed trajectory candidates. CLOVER Ang et al. (2026) adopted a generative scoring framework that ranked candidate trajectories according to planning-related metrics. AlignDrive Wu et al. (2026) further investigated the coordination between lateral and longitudinal planning. Despite continuous advances in scene parsing and trajectory planning, these methods generally lack an explicit interface for interacting with natural-language driving instructions.

Language-conditioned driving. Early work used natural-language advice to condition driving control (Kim et al., 2019; Roh et al., 2020). With the rapid development of LLMs (Chen et al., 2024; Liu et al., 2025), LMDrive (Shao et al., 2024) introduced language-guided closed-loop driving; SimLingo (Renz et al., 2025) studied language-action alignment; and VLA approaches (Zhou et al., 2025; 2026) model language jointly with driving outputs. NudgeVAD (Yang et al., 2026) also conditions planning through language-dependent residuals. Thus, neither language conditioning nor residual modulation alone is the novelty here. SIC-Drive's specific choice is to attach separate instruction representations to the path and speed candidate scores of a factorized planner, and to evaluate the executed instruction target at matched decision points.

3 Method

3.1 Instruction-Controlled Driving Formulation

Recent end-to-end driving models (Jiang et al., 2026; Sun et al., 2026) learn a direct mapping from environment observations to future trajectories:

$$\tau_t = f_\theta(O_t, S_t), \tag{1}$$

where O_t denotes multi-view observations, S_t represents the ego state, and τ_t is the predicted future trajectory. Although such models achieve strong autonomous driving capability, their planning objectives are mainly optimized through expert trajectory imitation (Wang et al., 2026a):

$$\mathcal{L}_{\text{imit}} = d(\tau_t, \tau_t^e), \tag{2}$$

where τ_t^e denotes the expert trajectory. However, instruction-controlled driving requires a different property: the same scene should allow different behaviors under different intentions. Specifically, given two instructions z_1 and z_2, the desired planner should satisfy:

$$\pi(O_t, S_t, z_1) \neq \pi(O_t, S_t, z_2), \tag{3}$$

Table 1: Examples of navigation and risk warnings randomly selected from the instruction templates. $[x]$ represents the float number for a specific distance.

Type	Example Instructions
Turn	Please turn left at the intersection ahead. Continue along the current road and turn right at the intersection ahead. Please go straight at the traffic-light-controlled intersection ahead.
Lane Change	Change to the left lane. Adjust your lane position and enter the left lane. When it is safe, change to the right lane.
Road Following	Continue along the current road for about $[x]$ meters. Maintain your current direction of travel for $[x]$ meters. Keep your current lane and continue for about $[x]$ meters.
Accident	An accident is blocking the lane ahead; drive carefully. Watch for the vehicles involved in the accident ahead. Leave sufficient braking space; a vehicle ahead is changing lanes into your path.
Construction	There is construction ahead; slow down while passing. Road work is ahead; reduce your speed. Construction is occupying the lane ahead; prepare to detour.

This requirement is not explicitly optimized by trajectory imitation, because multiple instructions may correspond to the same expert trajectory in the training data. Therefore, directly adding language as additional input does not guarantee that language can control planning behavior. To address this limitation, we formulate instruction-controlled driving as behavior-conditioned trajectory planning:

$$\tau_t = f_\theta(O_t, S_t, \phi(z_t)), \tag{4}$$

where $\phi(\cdot)$ converts natural language instructions into planning-level constraints rather than generic language features.

3.2 Language-Augmented Dataset Construction

Bench2Drive-Base (Jia et al., 2024) provides diverse closed-loop driving scenarios, with supervision primarily represented through structured annotations, including discrete waypoint-based navigation information, ego-vehicle states. However, it does not include natural-language instructions as a modality information. Building on its diversity in driving scenarios, we construct language-augmented Bench2Drive-Base. The augmentation comprises two components: navigation and risk warnings.

Navigation augmentation. The frame-level annotations contain navigation information, ego-vehicle states, and information about the traffic environment and surrounding participants. For turning and lane-changing maneuvers, we use the trigger frame of `command_far` as the advance trigger point for the instruction, and `command_near` to confirm the actual maneuver execution interval. Let t_f denote the trigger frame of a navigation instruction and $[t_s, t_e]$ the corresponding maneuver execution interval. The instruction applies over $[t_f, t_e]$, and the distance traveled by the ego vehicle over $[t_f, t_s]$ is accumulated to determine the distance specified in the instruction. This procedure yields structured semantics comprising the maneuver type, travel distance, and driving context. For driving intervals not covered by existing instructions, we add road-following instructions that specify travel distances to provide continuous navigation coverage.

Risk warnings augmentation. For different hazardous scenarios in the dataset, we determine the risk trigger time and the type of risk-warning instruction based on attributes such as the target traffic participant's type, distance from the ego vehicle, and `affects_ego`. Unlike navigation instructions, risk warnings are triggered only when hazardous scenarios are detected.

To diversify instruction wording and reduce the model's reliance on fixed expressions, we construct a library of driving instruction templates. The library covers all navigation rules required in this study

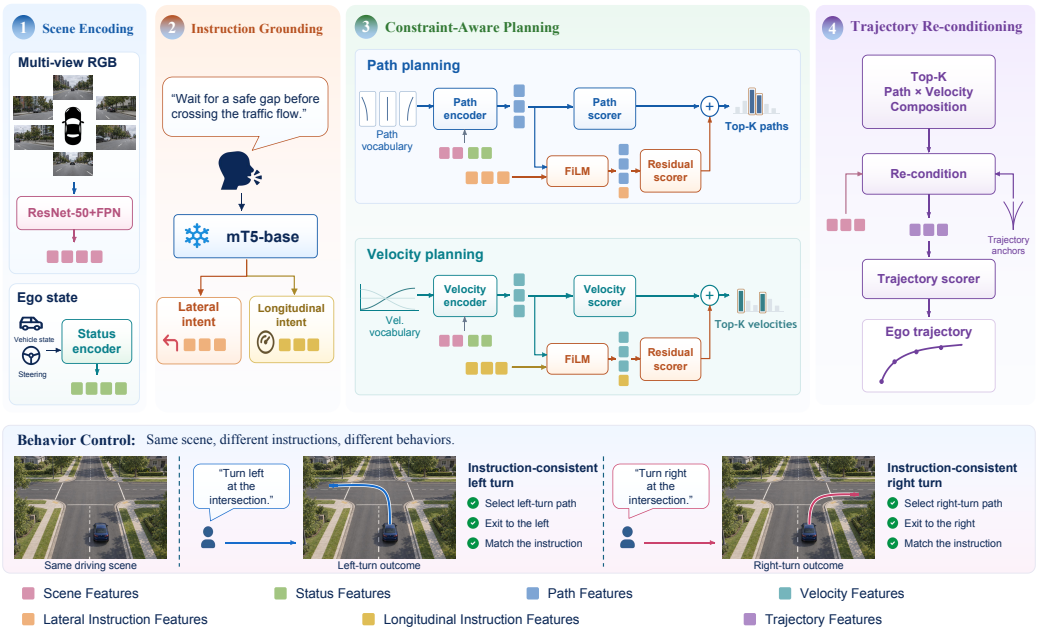


Figure 2: Structure of SIC-Drive. We augment the baseline planner with two instruction-aware components: 1) instruction grounding, which maps natural-language instructions into lateral and longitudinal representations aligned with path and velocity planning; 2) instruction-conditioned residual scoring, which uses these representations to refine path and velocity candidate scores before Top-K filtering, enabling different instructions to induce distinct instruction-consistent behaviors.

and includes risk warnings for 27 hazardous scenario types . Each instruction type has eight alternative expressions, constructing 656 instruction templates across 82 driving semantic categories, and Table 1 lists part of the typical examples,. Using this library, we construct 2,522 sparse driving instructions across 1,000 driving clips in the dataset, covering two or three instructions per clip. These instructions introduce natural-language supervision with precise temporal alignment to the driving sequences.

3.3 SIC-DRIVE

We propose SIC-Drive, a language-conditioned end-to-end driving framework built upon SparseDriveV2 (Sun et al., 2026). As illustrated in Figure 2, SIC-Drive follows the path–velocity decomposition of the baseline and introduces explicit instruction grounding into its coarse candidate filtering stage. Specifically, lateral and longitudinal representations condition the path and velocity scoring branches, respectively, to predict score residuals before Top-K selection. The selected candidates are subsequently composed and re-conditioned by the trajectory planner for fine-grained selection. This design preserves the baseline planning pipeline while allowing natural-language instructions to explicitly influence driving behavior.

Instruction-to-Behavior Grounding. Driving instructions may simultaneously specify where and how the vehicle should move (e.g., "turn left ahead and slow down"). Instead of directly injecting language representations into the planning network, SIC-Drive explicitly decomposes instructions into lateral and longitudinal behavior dimensions to align with the path–velocity structure of the planner. Given an instruction I_t, we first extract contextual language representations using a frozen mT5-base encoder (Xue et al., 2021) and project them into the planning feature space through a trainable projection layer. Two independent learnable queries then aggregate lateral and longitudinal representations:

$$H_t = \mathrm{Proj}(\mathrm{mT5}(I_t)), \qquad z_t^b = \mathrm{Attn}(q^b, H_t), \quad b \in \{\mathrm{lat}, \mathrm{lon}\}, \tag{5}$$

where q^{lat} and q^{lon} capture different driving semantics. The resulting z_t^{lat} represents path-related intentions such as turning and lane changing, while z_t^{lon} represents velocity-related intentions such as acceleration, deceleration, and stopping. This decomposition establishes an explicit correspondence between language semantics and the two planning branches.

Constraint-Aware Trajectory Planning. The grounded instruction representations directly modulate candidate selection rather than being implicitly fused into scene features. Given candidate features h_i^b and their original scores s_i^b, where b denotes either lateral or longitudinal planning, the baseline planner performs candidate ranking without language conditioning. We introduce language-conditioned score refinement for the path and velocity branches, where the corresponding instruction representation generates feature-wise modulation parameters through FiLM (Perez et al., 2018):

$$(\gamma_t^b, \beta_t^b) = g_b\big(\mathrm{LN}(z_t^b)\big), \qquad \tilde{h}_i^b = \mathrm{LN}\big[(1+\gamma_t^b) \odot \mathrm{LN}(h_i^b) + \beta_t^b\big]. \tag{6}$$

The original and instruction-conditioned features are then combined to predict a score residual:

$$\Delta s_i^b = f_b([h_i^b; \tilde{h}_i^b]), \qquad \hat{s}_i^b = s_i^b + a_t \Delta s_i^b, \tag{7}$$

where a_t indicates whether an instruction is active. By predicting additive score residuals rather than replacing the baseline scores, SIC-Drive preserves the original planning prior while allowing instructions to alter candidate preferences. The refined scores are used for coarse Top-K filtering in the corresponding path and velocity branches.

Learning Language–Behavior Alignment. To ensure that the extracted instruction representations correspond to meaningful driving behaviors, we introduce semantic classification and residual planning supervision. Two auxiliary classifiers supervise z_t^{lat} and z_t^{lon} with predefined lateral and longitudinal semantic labels, while the continuous representations are directly used by the planning branches. Residual supervision further connects instruction semantics with candidate selection by directly optimizing the instruction-induced score adjustments toward candidate preferences derived from expert trajectories. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{sem}} \sum_b \mathcal{L}_{\text{sem}}^b + \lambda_{\text{res}} \sum_{\ell,b} \mathcal{L}_{\text{res}}^{\ell,b}, \tag{8}$$

where $\mathcal{L}_{\text{base}}$ denotes the original planning objective and ℓ indexes the planning decoder layers. Semantic supervision grounds the two representations in interpretable driving intentions, while residual supervision further connects these intentions to trajectory selection.

3.4 DIF-BENCHMARK

Existing closed-loop driving benchmarks primarily evaluate whether an agent can successfully and safely complete a driving task. For example, Bench2Drive (Jia et al., 2024) evaluates driving capability and safety under predefined navigation goals, while LangAuto (Shao et al., 2024) mainly measures route progress and driving infractions for language-conditioned driving. However, these route-level metrics do not isolate whether the agent's behavior changes according to the given language instruction. In particular, successful route completion does not necessarily indicate that the instruction-specified behavior has been executed.

To quantitatively evaluate instruction-conditioned behavioral changes, we introduce DIF-Benchmark (Driving Instruction Following Benchmark). Its key design principle is to vary the language instruction while keeping the underlying physical driving condition fixed. At each decision point, multiple instruction variants share the same reference route, initial ego pose, weather condition, and traffic random seed. Therefore, behavioral differences across episodes can be attributed more directly to different driving instructions rather than scene variations. Each instruction is associated with predefined lateral and longitudinal behavioral targets, enabling explicit evaluation of whether the resulting driving behavior is consistent with the requested intention. Figure 3 illustrates representative lateral and longitudinal instruction variants evaluated under this setting.

Benchmark construction. Specifically, DIF-Benchmark comprises 60 physical decision points across nine towns, including 10 three-exit intersections, 20 two-exit intersections, 12 lane-change

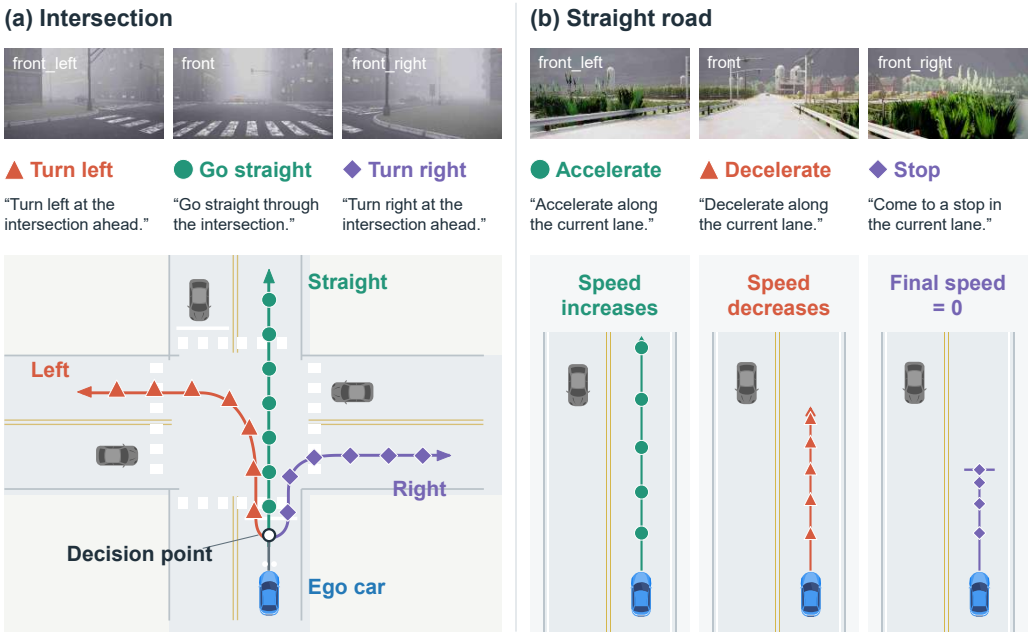


Figure 3: Illustration of language-conditioned driving decisions in two scenarios: (a) lateral route selection at an intersection, where instructions specify turning left, proceeding straight, or turning right; and (b) longitudinal control on a straight road, where instructions specify acceleration, deceleration, or stopping. Waypoints are sampled at equal time intervals, with their spacing indicating speed changes.

locations, 12 construction-obstacle locations, and six unobstructed road segments for longitudinal control, resulting in 136 instruction episodes. Each instruction is associated with predefined lateral and longitudinal completion criteria derived from its behavioral target; the detailed success conditions for each instruction type are provided in Appendix A.3. Each instruction becomes active when the vehicle enters a predefined spatial trigger region and remains active until instruction completion or episode termination. To isolate the behavioral effect of language from predefined route guidance, we neutralize the original navigation signal after instruction activation using a lane-following command and a virtual target located $8\,\mathrm{m}$ ahead of the ego vehicle.

Metrics. To evaluate instruction execution and driving safety separately, we report Instruction Success Rate (ISR) to measure overall instruction completion, together with Lateral Success Rate (LatSR) and Longitudinal Success Rate (LonSR) to evaluate consistency with the predefined lateral and longitudinal behavioral targets. We additionally report safe-driving success rate (Success Rate), adapting the safety criteria from the CARLA LeaderBoard (Zhang et al., 2024), independently of instruction completion. Collision and traffic-signal violation counts per kilometer, together with the proportion of affected episodes, are also reported to characterize safety risks during instruction execution.

4 Experiments

4.1 Experimental Setup

Dataset and Implementation. We conduct experiments based on the language-augmented Bench2Drive-Base introduced in Section 3.2. Following the baseline data split, 950 driving clips are used for training and 50 clips for validation. To investigate the effect of driving instructions on planning behavior, SIC-Drive preserves the sparse scene representation design (Sun et al., 2024) and candidate-based scoring planner (Sun et al., 2026), while incorporating driving semantics into

Table 2: Comparison of closed-loop planning performance on Bench2Drive. DS denotes Driving Score and SR denotes Success Rate. * denotes results reported in the original papers, while ** denotes results reproduced locally under our experimental setting.

Method	w/o Instructions	DS	SR(%)	Ability					
				Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	Mean
Uni-AD* (Hu et al., 2023)	×	45.81	16.36	14.10	17.78	21.67	10.00	14.21	15.55
VAD* (Jiang et al., 2023)	×	42.35	15.00	8.11	24.44	18.64	20.00	19.15	18.07
DriveTransformer* (Jia et al., 2025)	×	63.46	35.01	17.57	35.00	48.36	40.00	52.10	38.60
HiP-AD* (Tang et al., 2025)	×	86.77	**69.09**	50.00	**84.44**	83.33	40.00	72.10	65.98
SimLingo* (Renz et al., 2025)	✓	85.07	67.27	54.01	57.04	**88.33**	53.33	**82.45**	**67.03**
SparseDriveV2**	×	86.03	65.45	61.25	66.67	75.00	50.00	52.63	61.11
SIC-Drive	✓	**86.96**	66.82	**65.00**	66.67	71.67	**70.00**	55.79	65.82

Table 3: Performance comparison under different instruction modes.

Mode	DS ↑	SR (%) ↑	Ability ↑					
			Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	Mean
correct	**86.96**	**66.82**	**65.00**	66.67	71.67	**70.00**	**55.79**	**65.82**
mute	84.04	63.64	57.50	62.22	**73.33**	40.00	51.58	56.93
shuffled	75.51	50.91	42.50	66.67	60.00	60.00	22.11	50.25

the planning stage to perform residual score refinement over candidate trajectories. We keep the image backbone, feature pyramid, and mT5-base text encoder frozen. To ensure a fair comparison with the baseline, we follow the original training protocol as closely as possible and restrict hyperparameter adjustments to those required by our experimental environment and the newly introduced instruction-related objectives. All closed-loop evaluations are conducted in the CARLA simulator of version 0.9.15 (Dosovitskiy et al., 2017). Additional implementation details, including batch size, learning-rate scaling, optimization settings, and loss-weight selection, are provided in Appendix A.1.

Evaluation Protocols. We evaluate SIC-Drive from three complementary perspectives: instruction-conditioned controllability, overall closed-loop driving capability, and the contribution of the proposed instruction-grounding components. First, we use the standard Bench2Drive closed-loop evaluation (Jia et al., 2024) to examine whether introducing language conditioning preserves the agent's overall driving capability under matched driving scenarios. For this evaluation, instruction trigger points are constructed according to the navigation information provided by each route. We further verify that all injected instruction semantics can be observed from the camera inputs, thereby avoiding information leakage. Second, we quantitatively evaluate instruction following on DIF-Benchmark. Under the same physical driving condition, different instructions are provided, and we examine whether the agent produces the expected driving behaviors in response to variations in instruction. Finally, module-level ablations are conducted to analyze the contribution of the semantic parsing and instruction-conditioned trajectory scoring components.

4.2 Quantitative Results

We first examine whether introducing language-conditioned planning compromises the original closed-loop driving capability. As shown in Table 2, under the controlled comparison with our locally reproduced SparseDriveV2 baseline, SIC-Drive improves the Driving Score from 86.03 to 86.96 and the Success Rate from 65.45% to 66.82%, corresponding to gains of 0.93 and 1.37%, respectively. The mean score across the five driving abilities further increases from 61.11% to 65.82%. In particular, SIC-Drive improves *Merging*, *Give Way*, and *Traffic Sign* by 3.75%, 20.00%, and 3.16%, respectively, while maintaining the same performance on *Overtaking*. Although *Emergency Brake* decreases by 3.33%, the aggregate results indicate that instruction conditioning does not sacrifice the original closed-loop driving capability and instead yields moderate overall improvements.

We next investigate whether the closed-loop behavior is actually conditioned on the semantic content of language instructions. We evaluate the same trained SIC-Drive model under three instruction modes: *Correct*, where instructions are consistent with the route navigation; *Mute*, where language

Table 4: Ablation study of instruction encoders and fusion strategies on the DIF-Benchmark. ISR denotes Instruction Success Rate, LatSR denotes Lateral Success Rate, and LonSR denotes Longitudinal Success Rate.

Instruction Model	Fusion Mode	ISR (%)	LatSR (%)	LonSR (%)	Success Rate (%)
mT5-base	FiLM	60.29	67.65	92.65	66.18
mT5-base	Concat+MLP	58.08	64.71	93.38	61.76
RoBERTa-base	FiLM	58.62	63.24	94.85	66.18

input is empty instruction; and *Shuffled*, where conflicting instructions are provided. Only the instruction content is changed, while the triggering time and duration remain identical across the three settings, isolating the effect of instruction semantics. As shown in Table 3, *Correct* instructions consistently provide the strongest aggregate performance. Compared with the *Mute* setting, correct instructions improve DS by 2.92 points, SR by 3.18%, and the mean ability score by 8.89 percentage points. In contrast, replacing them with conflicting instructions decreases DS from 86.96 to 75.51, SR from 66.82% to 50.91%, and the mean ability score from 65.82% to 50.25%. Importantly, the substantial degradation under *Shuffled* instructions is obtained without changing their temporal activation, indicating that the performance difference is primarily associated with instruction semantics rather than instruction timing. These results show that SIC-Drive does not merely accept language as an auxiliary input; its closed-loop decisions are sensitive to the semantic content of the provided instruction.

The evaluation result of SIC-Drive on DIF-Benchmark, which quantitatively measures instruction-conditioned behavioral changes under controlled physical driving conditions. SIC-Drive successfully satisfies 82 instruction targets, resulting in an Instruction Success Rate (ISR) of 60.29%. The corresponding lateral and longitudinal success rates are 67.65% and 92.65%, respectively. The substantial gap between LatSR and LonSR indicates that SIC-Drive follows longitudinal instructions more reliably than lateral instructions. Moreover, ISR is lower than either marginal success rate because an instruction involving both semantic dimensions requires the corresponding lateral and longitudinal objectives to be satisfied jointly. This reveals that the main remaining challenge lies not only in recognizing individual instruction semantics, but also in consistently coordinating multiple instruction-conditioned behaviors during closed-loop interaction. Together with the instruction-mode experiment, these results provide direct evidence that SIC-Drive can translate language variations into measurable behavioral changes, while also showing that its controllability remains uneven across different semantic dimensions. The contributions of individual instruction-grounding components are further analyzed through ablation studies in Appendix A.4.

5 Conclusion

We propose SIC-Drive, an instruction-grounded end-to-end driving framework that connects natural-language instructions to closed-loop behavior through candidate selection. SIC-Drive decomposes instructions into lateral and longitudinal semantics and uses them to refine path and velocity candidate scores before trajectory selection, providing an explicit mechanism for language to alter planning preferences while preserving the original planning prior. To support instruction-conditioned learning and evaluation, we construct a language-augmented Bench2Drive-Base with temporally aligned navigation and risk warnings, and introduce DIF-Benchmark to measure whether the executed closed-loop behavior reaches the instruction-specified target under matched physical driving conditions. Extensive experiments show that changing instruction semantics leads to systematic changes in the selected driving behavior, confirming that instruction-conditioned candidate scoring provides an effective control signal for planning decisions while preserving competitive overall driving performance. However, the remaining failures on DIF-Benchmark indicate that this control is not yet fully reliable and remains challenging. These findings distinguish instruction sensitivity from reliable instruction following and highlight instruction execution as a capability that should be explicitly modeled and evaluated beyond general closed-loop driving performance.

AI USE STATEMENT

Generative AI tools, such as GPT 5.6 Sol, assisted with part of experimental code, manuscript organization and language editing, and critical feedback on the interpretation of the reported experiments. All authors have reviewed the final manuscript and all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

The experiments use CARLA simulation and template-based driving instructions. Success in these settings does not establish safe real-world operation or equal performance across road users, environments, or linguistic expressions. We report instruction completion separately from driving safety to avoid treating completion of a requested maneuver as a safety guarantee.

REPRODUCIBILITY STATEMENT

Section 3 describes the model, annotation process, and benchmark design. Section 4 reports the evaluation settings and results; Appendix A.1 records the training configuration and Appendix A.3 describes the available episode-completion criteria. Exact benchmark episode definitions, evaluation thresholds, and anonymized code should accompany the submission to permit independent reproduction.

REFERENCES

Sining Ang, Yuguang Yang, Canyu Chen, and Yan Wang. Clover: Closed-loop value estimation and ranking for end-to-end autonomous driving planning, 2026. URL https://arxiv.org/abs/2605.15120.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles, 2022. URL https://arxiv.org/abs/2106.11810.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction, 2019. URL https://arxiv.org/abs/1910.05449.

Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *CVPR*, 2022.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14093–14100, 2024. doi: 10.1109/ICRA57147.2024.10611018.

Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, Lingxi Li, Yaobin Chen, Jitesh Panchal, and Ziran Wang. Personalized autonomous driving with large language models: Field experiments, 2024. URL https://arxiv.org/abs/2312.09397.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Kaiser Hamid, Can Cui, and Nade Liang. Icr-drive: Instruction counterfactual robustness for end-to-end language-driven autonomous driving, 2026. URL https://arxiv.org/abs/2604.05378.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, June 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 819–844. Curran Associates, Inc., 2024. doi: 10.52202/079017-0025. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/017761f94a1cd66d01c041aff85492c4-Paper-Datasets_and_Benchmarks_Track.pdf.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 67227–67243, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a7afc9957f1190223763b6ea93218f98-Paper-Conference.pdf.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8350, October 2023.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 68886–68900, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/70925fe2336ae05917c817d5e4660ab5-Paper-Conference.pdf.

Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10583–10591, 2019. doi: 10.1109/CVPR.2019.01084.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, Yu-Gang Jiang, and Jose M. Alvarez. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation, 2024. URL https://arxiv.org/abs/2406.06978.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Changxing Liu, Genjia Liu, Zijun Wang, Jinchang Yang, and Siheng Chen. Colmdriver: Llm-based negotiation benefits cooperative autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25951–25960, 2025. doi: 10.1109/ICCV51701.2025.02407.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Ziheng Lu, Yingfeng Cai, Wei Dong, Hai Wang, and Long Chen. MALT-Drive: Moment-Aligned Language-to-Trajectory Planning for Autonomous Driving. *Computer Modeling in Engineering & Sciences*, 148(2):40, 2026. doi: 10.32604/cmes.2026.086003. URL https://www.techscience.com/CMES/v148n2/68588.

Long Nguyen, Micha Fauth, Bernhard Jaeger, Daniel Dauner, Maximilian Igl, Andreas Geiger, and Kashyap Chitta. Lead: Minimizing learner-expert asymmetry in end-to-end driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39775–39785, June 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Tung Phan-Minh, Elena Corina Grigore, Freddy A. Boulton, Oscar Beijbom, and Eric M. Wolff. Covernet: Multimodal behavior prediction using trajectory sets, 2020. URL https://arxiv.org/abs/1911.10298.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7073–7083, 2021. doi: 10.1109/CVPR46437.2021.00700.

Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11993–12003, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Renz_SimLingo_Vision-Only_Closed-Loop_Autonomous_Driving_with_Language-Action_Alignment_CVPR_2025_paper.html.

Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, and Dieter Fox. Conditional driving from natural language instructions. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 540–551. PMLR, 2020.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving, 2025. URL https://arxiv.org/abs/2310.03026.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *Conference on Computer Vision and Pattern Recognition*, pp. 15120–15130, 2024. doi: 10.1109/CVPR52733.2024.01432. URL https://mlanthology.org/cvpr/2024/shao2024cvpr-lmdrive/.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation, 2024. URL https://arxiv.org/abs/2405.19620.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving, 2026. URL https://arxiv.org/abs/2603.29163.

Yingqi Tang, Zhuoran Xu, Zhaotie Meng, and Erkang Cheng. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25605–25615, 2025. doi: 10.1109/ICCV51701.2025.02375.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Junli Wang, Zhihua Hua, Xueyi Liu, Zebin Xing, Haochen Tian, Kun Ma, Hangjun Ye, Guang Chen, Long Chen, and Qichao Zhang. Beyond imitation: Learning safe end-to-end autonomous driving from hard negatives, 2026a. URL https://arxiv.org/abs/2605.19771.

Junli Wang, Yinan Zheng, Xueyi Liu, Zebin Xing, Pengfei Li, Kun Ma, Hangjun Ye, Guang Chen, Guang Li, Long Chen, Zhongpu Xia, and Qichao Zhang. Meanfuser: Fast one-step multi-modal trajectory generation and adaptive reconstruction via meanflow for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17884–17893, June 2026b.

Xinyang Wang, Qian Liu, Wenjie Ding, Zhao Yang, Wei Li, Chang Liu, Bailin Li, Kun Zhan, Xianpeng Lang, and Wei Chen. Unifying language-action understanding and generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25193–25203, June 2026c.

Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6119–6132. Curran Associates, Inc., 2022. doi: 10.52202/068431-0443. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/286a371d8a0a559281f682f8fbf89834-Paper-Conference.pdf.

Yanhao Wu, Haoyang Zhang, Fei He, Rui Wu, Yanhu Shan, Congpei Qiu, Liang Gao, Wei Ke, and Tong Zhang. AlignDrive: Aligned lateral-longitudinal planning for end-to-end autonomous driving, 2026. URL https://arxiv.org/abs/2601.01762.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Chieh-Chi Yang, Yu-Hsiang Chen, and Yi-Ting Chen. Nudgevad: Language-nudged end-to-end driving via film residuals, 2026. URL https://arxiv.org/abs/2605.24531.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11910–11918, 2026.

Weize Zhang, Mohammed Elmahgiubi, Kasra Rezaee, Behzad Khamidehi, Hamidreza Mirkhani, Fazel Arasteh, Chunlin Li, Muhammad Ahsan Kaleem, Eduardo R. Corral-Soto, Dhruv Sharma, and Tongtong Cao. Analysis of a modular autonomous driving architecture: The top submission to carla leaderboard 2.0 challenge, 2024. URL https://arxiv.org/abs/2405.01394.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13782–13790, 2026. doi: 10.1609/aaai.v40i16.38386.

Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2506.13757.

A Appendix

The appendix provides additional details on the training configuration, validation performance under three language input settings, and ablation analysis of the language-related modules.

A.1 Additional Experimental Details

The reported experiments use eight NVIDIA RTX 4090 GPUs and the 950/50 clip train/validation split of the language-augmented Bench2Drive-Base data. The original 16-GPU configuration uses a per-GPU batch size of six, a global batch size of 96, and an AdamW learning rate of 3×10^{-4}. With eight GPUs (Sun et al., 2026), the reported setup keeps six samples per GPU, giving a global batch size of 48, and scales the learning rate to 1.5×10^{-4}. It retains ten training epochs; the recorded iteration count changes from 24,450 to 48,910. No gradient accumulation is reported.

Table 5: Reported open-loop trajectory errors under different language inputs. Lower is better.

Instruction input	L2 (m)	ADE (m)	FDE (m)
Correct	**0.8150**	**1.2805**	**2.7475**
Mute	0.8732	1.3854	2.9926
Shuffled	0.8849	1.4093	3.0806

The mT5-base text encoder (Xue et al., 2021) is frozen, uses at most 64 tokens per instruction, and feeds a 256-dimensional projection. The instruction pooling heads use eight attention heads. Dropout is 0.1 in the text and modulation branches. The image backbone (He et al., 2016) and feature pyramid (Lin et al., 2017) are also frozen. The other trainable planning and language components are optimized in the planning stage. The exact baseline training configuration and evaluation seeds should be released with the code so that the local comparison can be reproduced.

In equation **??**, we empirically set $\lambda_{\text{sem}} = 0.1$ and $\lambda_{\text{res}} = 0.01$ through iterative tuning, with training stability and gradient balance as the main criteria. Applying the same weight to both auxiliary objectives can result in excessively large gradients from one loss and unbalanced optimization. The adopted weights keep their gradient contributions at comparable scales, preventing any auxiliary objective from dominating the baseline planning loss.

A.2 Open-loop Evaluation under Different Language Inputs

Before conducting closed-loop evaluation, we first perform open-loop experiments on the validation set to investigate whether the model can actively adjust its planning outputs according to different natural-language inputs while keeping the environmental observations unchanged. For the same frame, we fix the visual observations, scene features, ego state, and underlying navigation information, and vary only the language instruction. According to the relationship between the driving instruction and the original navigation objective, we consider the following three input settings and examine whether the resulting plans exhibit semantically consistent changes.

Navigation Only(mute). The model receives only the original navigation information, without any additional natural-language instruction that explicitly affects driving behavior. This setting corresponds to the standard input configuration of the baseline model and serves as the reference condition for evaluating the effect of language conditioning.

Navigation-Aligned Instruction(correct). In addition to the original navigation goal, the model receives a natural-language instruction that is consistent with both the current navigation objective and the expert driving behavior. For example, when the navigation command requires the vehicle to proceed straight, the instruction may further specify maintaining the current lane, smoothly passing through the intersection, or paying attention to a vehicle ahead. Such instructions do not alter the overall navigation objective, but provide finer-grained constraints on path selection or speed adjustment. This setting is used to evaluate whether SIC-Drive can exploit fine-grained driving semantics to produce planning behaviors that are better aligned with the intended driving strategy.

Navigation-Conflicting Instruction(shuffled). We further construct instructions that are inconsistent with the original navigation tendency to examine whether the model truly utilizes natural language as a high-level driving condition rather than relying solely on the underlying navigation signal. For example, when the original navigation input favors maintaining the current driving strategy and proceeding straight at a steady speed, the shuffled instruction may instead request the vehicle to turn while slowing down or to adopt a different local driving behavior. This setting mainly evaluates the sensitivity of the model to changes in language input and the ability of natural-language instructions to modulate the original planning decisions.

For each validation frame, the reported open-loop test holds the visual observation, scene features, ego state, and underlying navigation input fixed while changing only the instruction. *Correct* instructions align with the navigation objective; *mute* omits language; and *shuffled* supplies an instruction with mismatched semantics. Table 5 records the trajectory errors.

Correct input reduces L2, ADE, and FDE by 6.66%, 7.57%, and 8.19% relative to mute input. These are trajectory-matching metrics; they do not independently establish that a conflicting instruction is followed in closed loop.

A.3 BENCHMARK DETAILED DESIGN

For each test scenario in DIF-Benchmark, a driving instruction is activated at a predefined spatial location according to the current environment state. During execution, we continuously monitor the ego state, road topology, and the motion states of surrounding traffic participants to determine whether the corresponding instruction is successfully completed within its valid time window.

Each language instruction at a decision point is associated with predefined lateral and longitudinal scoring targets. The lateral semantic space consists of *turn left*, *turn right*, *keep*, *change lane left*, and *change lane right*, while the longitudinal semantic space consists of *keep*, *decelerate*, *accelerate*, and *stop*. If an instruction does not explicitly specify a requirement along one dimension, the corresponding semantic label is set to *keep*. In particular, when no explicit speed requirement is given, the *keep* label does not impose an additional constant-speed constraint during evaluation.

An instruction is activated once the ego vehicle reaches its predefined spatial trigger region and remains active until it is successfully completed, the episode terminates, or the instruction times out. DIF-Benchmark is designed to evaluate whether an agent can safely execute driving behaviors specified by natural-language instructions. To reduce interference from the predefined route with language-conditioned decision making, the navigation cue is replaced with a neutral condition after instruction activation, indicating only that the vehicle should proceed straight within the next 8 meters. This formulation is also consistent with SIC-Drive, which independently parses lateral and longitudinal driving semantics and injects them into path scoring and speed scoring, respectively.

Instruction completion is determined from the actual ego trajectory and speed profile. For turning instructions, the vehicle is required to enter the target outgoing road branch and travel for at least 20 meters after entering it, where the target branch is independently determined from the road geometry. A lane-change instruction is considered successful when the vehicle enters the target lane, maintains a heading error below 15°, remains in the target lane continuously for at least 2 seconds, and travels at least 10 meters within it. For construction-zone avoidance, the vehicle is required to perform the necessary lateral maneuver and remain clear of the obstacle until it has completely passed the construction region.

Acceleration and deceleration instructions are evaluated according to the reference speed before instruction activation and the duration of the resulting speed change. For explicit speed-maintaining instructions, we measure the proportion of valid frames whose speed deviation remains within the predefined tolerance. A stopping instruction follows a separate completion criterion: the vehicle must rapidly reduce its speed to below $1\,\mathrm{m/s}$ during the valid instruction interval and remain below this threshold for at least 1 second. Once this condition is satisfied, the current episode is considered successfully completed, without imposing additional constraints on the exact stopping location or subsequent vehicle motion.

Instructions are not active throughout the entire clip. Instead, they are triggered when the vehicle approaches or reaches a scenario-relevant location. For example, a turning instruction is issued before the vehicle enters the corresponding intersection, while an avoidance instruction for an obstacle ahead is activated before the ego vehicle reaches the associated risk region. This design better resembles how a human driver or passenger would provide driving instructions in real-world interactions.

For each test instruction, we first determine whether it is completed according to the semantic-specific criteria described above. Driving safety is then evaluated using closed-loop events such as collisions, road departures, traffic-rule violations, and task completion. Therefore, an instruction-following behavior is considered effective only when the agent both executes the specified instruction correctly and maintains safe driving behavior.

This design also avoids the limitations of evaluating language-conditioned driving solely through trajectory errors. DIF-Benchmark focuses on whether the vehicle realizes the driving intent specified by the instruction, rather than whether it reproduces a unique expert trajectory, thereby allowing multiple behaviorally distinct but semantically correct and safe driving strategies.

A.4 LANGUAGE-MODULE ABLATIONS

The reported ablations replace FiLM with feature concatenation and an MLP, or replace mT5-base with RoBERTa-base (Liu et al., 2019); see Table 4. mT5-base with FiLM has the highest reported ISR, whereas RoBERTa-base with FiLM has the highest LonSR. The available results do not isolate the semantic-head supervision or candidate-residual objective, and they include no uncertainty estimates. Attribution to a specific component therefore remains tentative.