

HiP-AD: Hierarchical and Multi-Granularity Planning with Deformable Attention for Autonomous Driving in a Single Decoder

Yingqi Tang* Zhuoran Xu* Zhaotie Meng Erkang Cheng[✉]

Nullmax

{tangyingqi, xuzhuoran, mengzhaotie, chengerkang}@nullmax.ai

<https://github.com/nullmax-vision/HiP-AD>

Abstract

Although end-to-end autonomous driving (E2E-AD) technologies have made significant progress in recent years, there remains an unsatisfactory performance on closed-loop evaluation. The potential of leveraging planning in query design and interaction has not yet been fully explored. In this paper, we introduce a multi-granularity planning query representation that integrates heterogeneous waypoints, including spatial, temporal, and driving-style waypoints across various sampling patterns. It provides additional supervision for trajectory prediction, enhancing precise closed-loop control for the ego vehicle. Additionally, we explicitly utilize the geometric properties of planning trajectories to effectively retrieve relevant image features based on physical locations using deformable attention. By combining these strategies, we propose a novel end-to-end autonomous driving framework, termed HiP-AD, which simultaneously performs perception, prediction, and planning within a unified decoder. HiP-AD enables comprehensive interaction by allowing planning queries to iteratively interact with perception queries in the BEV space while dynamically extracting image features from perspective views. Experiments demonstrate that HiP-AD outperforms all existing end-to-end autonomous driving methods on the closed-loop benchmark Bench2Drive and achieves competitive performance on the real-world dataset nuScenes.

1. Introduction

Recently, great progress has been achieved in end-to-end autonomous driving (E2E-AD), which directly predicts planning trajectory from raw sensor data. One of the mainstream methods is to integrate all tasks (e.g., perception, prediction, and planning) into a single model within a fully differentiable manner [6, 16, 23, 54]. Compared to the traditional standalone or multi-task paradigm [11, 40, 51], it greatly alleviates the accumulative errors and enables all task modules

* Equal contribution; [✉] Corresponding author.

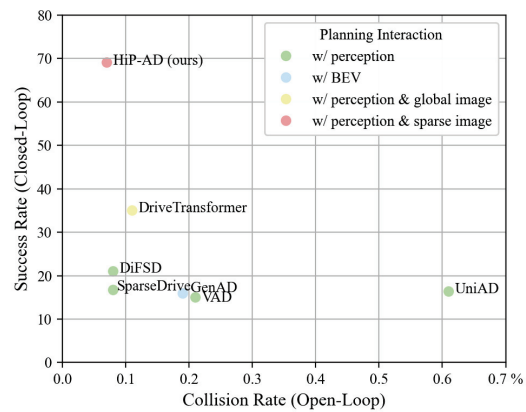


Figure 1. Comparison of existing state-of-art works on open-loop metric of Collision Rate on nuScenes dataset and closed-loop metric of Success Rate on Bench2Drive dataset, where **top left is better**. The legend indicates different planning interaction methods.

work collaboratively, which exhibits promising performance under the effect of large-scale data.

Despite these advances, a significant performance gap persists between open-loop and closed-loop evaluations, primarily attributable to differences in motivation. Open-loop methods focus on the displacement error in the planning trajectory compared to the ground truth, while closed-loop methods prioritize safe driving performance. As illustrated in Fig. 1, previous E2E-AD methods [16, 23, 48, 49, 59] demonstrate strong performance in terms of Collision Rate (lower is better) in the open-loop benchmark nuScenes [2] with some methods achieving as low as 0.1%. However, these methods show unsatisfactory performance in terms of Success Rate on the comprehensive closed-loop evaluation dataset, Bench2Drive [21], where the Success Rate remains below 35%. Even when focusing solely on emergency braking, the Success Rate is still inadequate, below 55% (as shown in Tab. 2), despite achieving an open-loop Collision Rate as low as 0.1%.

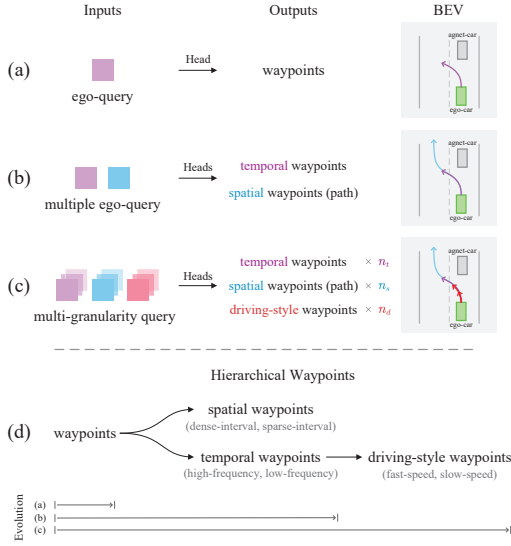


Figure 2. This diagram compares earlier methods (a-b) for predicting waypoints with our proposed multi-granularity planning design (c), where n_t , n_s , and n_d represent different number of granularity in each waypoints type in terms of frequency, interval, and speed. Part (d) illustrates the evolution of hierarchical waypoints with instantiated granularity based on different sampling strategies.

We argue that the potential of planning in query design and interaction has not been fully explored in these E2E-AD methods. First, most methods [16, 22, 23, 54, 58] formulate E2E-AD as an imitation learning task via a trajectory regression (Fig. 2 (a)) with sparse supervision, focusing primarily on the trajectory fitting itself rather than closed-loop control. In contrast, closed-loop oriented methods [9, 18, 42, 45] encounter several other challenges, such as non-convex problems [6] and steering errors [43]. CarLLaVA [43] greatly alleviates these issues by decoupling standard waypoints into time-conditioned and space-conditioned waypoints for longitudinal and lateral control, as shown in Fig. 2 (b). However, it is built upon a pre-trained large language model without intermediate perceptual results, which lacks interpretability, and it has not investigated the diversification of trajectories. In this paper, we propose multi-granularity planning query representation with hierarchical waypoints predictions for E2E-AD, as shown in Fig. 2 (c-d). Specifically, we disentangle the waypoints into temporal, spatial (path), and driving-style waypoints predictions with corresponding planning queries¹. Additionally, we further diversify each type of waypoints into multiple granularities with different sampling strategies, such as frequency, distance, and speed, enriching additional supervision during training. They can be effectively aggregated to facilitate interaction between the different characteristics. As a result, sparse waypoints provide global information, and dense waypoints outputs are

¹some works refer to this as ego query

more suitable for fine-grained control. Moreover, the multi-granularity dramatically reduces the ego hesitation issue that the ego vehicle keeps waiting in some scenarios until the closed-loop simulation times out. It encourages behavior learning in complex scenarios (e.g., traffic signs, overtaking) without introducing causal clues.

Second, sequential paradigms like UniAD [16] and VAD [23] formulate interaction only between learnable ego queries and the perception transformer outputs. While the parallel approach of Para-Drive [54] applies the ego query interacting solely with BEV features. These approaches lack the comprehensive interaction for planning to effectively engage with both perception and scene features (e.g., image or BEV features). In contrast, DriveTransformer [22] allows ego query to fully interact with both perception and image features within a single Transformer. However, it remains challenging for the ego query to effectively extract valuable information from multi-view images through global attention without the prior context of the planning trajectory. To address this issue, we employ planning deformable attention with physical locations to dynamically sample relevant image features in proximity to the planning trajectory. It can be easily integrated into a unified framework alongside perception tasks. Specifically, we employ a unified decoder with hybrid task queries as inputs, which combines queries from detection tasks (including motion prediction), map understanding tasks, and planning tasks. It enables planning and perception tasks to exchange information in the BEV space and allows planning query to interact with the image space by leveraging the geometric priors of their waypoints.

With both strategies, we propose a novel end-to-end autonomous driving framework, termed HiP-AD. It is evaluated on closed-loop benchmark Bench2Drive [21] and open-loop dataset nuScenes [2], achieving outstanding results on both planning and perception tasks. Our main contributions are as follows:

- We propose a multi-granularity planning query representation that integrates various characteristics of waypoints to enhance the diversity and receptive field of the planning trajectory, enabling additional supervision and precise control in a closed-loop system.
- We propose a planning deformable attention mechanism that explicitly utilizes the geometric context of the planning trajectory, enabling dynamic retrieving image features from the physical neighborhoods of waypoints to learn sparse scene representations.
- We propose a unified decoder where a comprehensive interaction is iteratively conducted on planning-perception and planning-images, making an effective exploration of end-to-end driving in both BEV and perspective view, respectively.

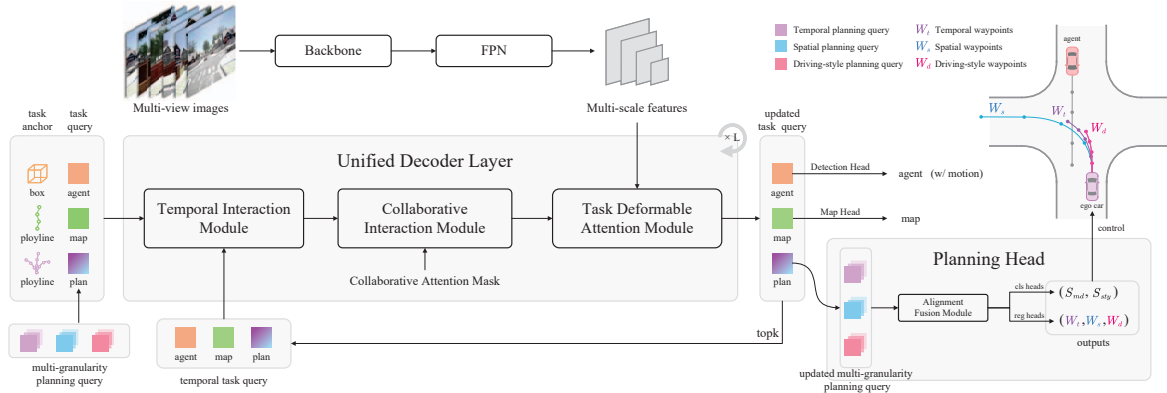


Figure 3. **The overall framework of HiP-AD.** It comprises a Backbone with FPN for image feature extraction, a Unified Decoder for iterative query refinement, and various Heads for multi-task prediction. The inputs of the Unified Decoder are task anchors and queries (agent, map, and planning), where planning query consists of multi-granularity waypoints representations. Within each decoder layer, the task queries first interact with temporal query separately, then collaboratively with each other, and finally engage the image features in an iterative manner. Last, the updated task queries are sent to the corresponding heads for perception, prediction, and planning. The planning results including various waypoints with different granularity for precise action control.

2. Related Work

2.1. Dynamic and Static Perception

Dynamic object detection can be broadly categorized into BEV-based methods and sparse query-based methods. The BEV-based methods [17, 26, 27, 39] detect objects by constructing bird’s-eye view (BEV) features, either by flattening voxel features with estimated depth or by injecting perspective features into BEV grids. In contrast, sparse query-based methods [32, 35, 52, 53] directly compute detection results through techniques such as geometric embedding and anchor projection with deformable attention, reducing the need for extensive construction of BEV features and leading to a more efficient detection process.

Static map element detection tasks are dominated by BEV representation methods. Rasterized-based methods [25] predict static elements through techniques such as segmentation and lane detection, followed by a post-processing step. In contrast, vectorized representation methods [30, 36] directly model vectorized instances without post-processing, leading to improved efficiency and higher performance. Building on this paradigm, advancements such as long-sequence temporal fusion [56] and multi-points scatter strategies [37] further enhance the robustness of static element detection.

2.2. Motion Prediction and Planning

Motion prediction methods predict vehicle trajectories from LiDAR points or rendered HD maps. For example, IntentNet [3] develops a one-stage detector and forecaster with 3D point clouds and dynamic maps. VectorNet [14] proposes an efficient graph neural network for predicting the behaviors of multiple agents. PnPNet [29] develops a multi-object tracker that simultaneously performs perception and motion

forecasting, leveraging rich temporal context for enhanced performance. ViP3D [15] proposes a fully differentiable vision-based trajectory prediction approach.

Traditional planning has often been treated as an independent task, relying on the outputs from perception and prediction modules. Early works utilize rule-based planners [1, 11, 51] or optimization-based methods [13, 47, 60] for planning. In contrast, learning-based methods have gained prominence, offering new paradigms for planning tasks. Some works [8, 9, 41, 42] directly predict planning trajectories or control signals without intermediate results, while other methods [10, 44] optimize planning by utilizing outputs of perception and prediction for great interpretability. Additionally, reinforcement learning [4, 5, 24, 50] has shown significant promise in autonomous driving.

2.3. End-to-end Autonomous Driving

Different from the traditional standalone or multi-task paradigm [11, 40, 51], End-to-End autonomous driving methods alleviate potential cumulative errors by applying a unified pipeline that integrates perception, prediction, and planning tasks. For example, UniAD [16] pioneeringly integrates various tasks into a single model. VAD [6, 23] simplifies the scene representation to vectorized elements, enhancing both efficiency and robustness. PPAD [7] formulates a hierarchical dynamic key object attention to model the interactions in an interleaving and autoregressive manner. Instead of constructing dense BEV features, SparseAD [58] and SparseDrive [49] utilize a sparse query-based framework, achieving greater efficiency and more accurate results. Based on this manner, DiFSD [48] iteratively refines the ego trajectory from coarse-to-fine through geometric information. In contrast to the sequential scheme, Para-Drive [54] employs

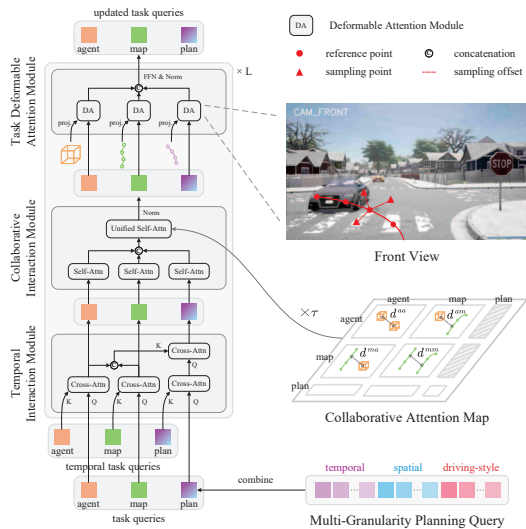


Figure 4. Illustration of the detailed architecture of three submodules within the unified decoder layer for comprehensive interaction.

a parallel approach, performing perception and planning tasks concurrently for improved performance. DriveTransformer [22] further integrates these tasks within a single transformer and achieves excellent performance in closed-loop system. Additionally, generative frameworks [31, 59] and large language models [38, 43, 46] provide innovative perspectives for decision-making and planning, and are increasingly capturing attention in the field.

3. Method

3.1. Overview

The overall network architecture of HiP-AD is illustrated in Fig. 3. It consists of a backbone followed by a feature pyramid network (FPN) module that extracts multi-scale features $\{F_i\}_{i=1}^V$ from multi-view images $\{I_i\}_{i=1}^V$, and a unified decoder with various task-specific heads. The unified decoder takes hybrid task anchors and queries as inputs, which are concatenated from agent queries $Q_a \in \mathbb{R}^{N_a \times C}$, map queries $Q_m \in \mathbb{R}^{N_m \times C}$, and planning queries $Q_p \in \mathbb{R}^{N_p \times C}$, where N represents the number of queries and C denotes the feature channel size. The agent query corresponds to object detection and motion prediction, while the map and multi-granularity planning queries (Sec. 3.3) manage online mapping and trajectory prediction. The detection and motion prediction heads, along with the map and planning heads, predict their respective tasks. The planning head outputs temporal, spatial, and driving-style waypoints for ego-vehicle control. Additionally, the top k_a, k_m, k_p updated queries are stored in memory for subsequent temporal interactions.

3.2. Unified Decoder

As shown in Fig. 3 and in detail in Fig. 4, the unified decoder consists of three modules: the Temporal Interac-

tion Module, the Collaborative Interaction Module, and the Task Deformable Aggregation Module. Each module is designed to facilitate temporal, cross-task, and task-image interactions, respectively. Each input task query is associated with a corresponding anchor. The agent query uses box anchors, $A_a \in \mathbb{R}^{N_a \times D_a}$, while the map query utilizes polyline anchors, $A_m \in \mathbb{R}^{N_m \times D_m}$, initialized through a clustering algorithm, where D denotes the anchor dimension. We also model the planning query as a polyline anchor $A_p \in \mathbb{R}^{N_p \times D_p}$ using its T future waypoints.

Temporal Interaction. The Temporal Interaction Module establishes communication between the features of the current task and those of historical tasks, which are retained from the previous inference frame via a top k selection mechanism. As shown in the bottom left of the Fig. 4, three distinct cross-attention mechanisms for each task’s temporal interaction, as well as an additional cross-attention mechanism that enables enhanced interaction between the planning query and temporal perception queries, focusing on historical surrounding elements.

Collaborative Interaction. The Collaborative Interaction Module enables cross-task interaction. It incorporates three separate self-attention mechanisms, each dedicated to an individual task, as well as a unified self-attention module to apply interaction across tasks. Instead of using global attention, we construct a geometric attention map for each query pair to focus on local and relative elements. Using perception query as an example, following [34], we dynamically adjust BEV receptive fields through scaling distance as attention weight.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}} - \tau D\right)V, \quad (1)$$

where τ is a learnable coefficient computed from Q by MLP functions. D represents the Euclidean distance between two object instances, denoted as $(d_{i,j})$. Similarly, we extend the computation of minimum distances to generate the attention weight by incorporating interactions between map-agent, map-map, and agent-map anchors. For planning queries, there are no distance constraints, allowing them to access information from all tasks.

Task Deformable Attention. Unlike previous works [22], which employ global attention to interact with all multi-view image features, we leverage separate deformable attention modules to sample local sparse features tailored to each task query. Specifically, we project task anchors to multi-view images through the camera parameters, as used in [33, 49]. For planning, we distribute reference waypoints across various predefined height values and then project them onto multi-view images. To sample features of neighboring points, we employ several MLPs to learn spatial offsets and associated weights based on the projected reference points. The process of planning deformable attention (PDA) can be formulated

as:

$$\text{PDA}(Q_p, F) = \sum_{i \in V} \text{DeformAttn}(Q_p, \mathcal{P}(A_p), F_i), \quad (2)$$

where \mathcal{P} indicates a project function. Therefore, it integrates features around the future trajectory to learn the sparse scene representation, avoiding a potential collision.

3.3. Hierarchical and Multi-granularity Planning

Hierarchical Waypoints. Different from previous waypoints designs [23, 43], we incorporate novel driving-style waypoints alongside traditional temporal and spatial representations Fig. 2 (d). It further integrates diverse driving styles with distinct velocities to learn ego-vehicle actions in complex environments. Additionally, a multi-sampling strategy is applied to enable rich trajectory supervision and precise control. This strategy combines dense and sparse intervals for spatial waypoints, high and low frequencies for temporal and driving-style waypoints. As a result, sparse waypoints provide a broader global context for advanced decision-making, while dense-interval waypoints enable fine-grained control for precise maneuvering. Moreover, the integration of driving-style waypoints with different speeds provides a rich understanding of complex scenarios like overtaking or emergency braking, ensuring flexible longitudinal control in close-loop evaluation.

Multi-granularity Planning Query. We construct multi-granularity planning queries to predict these heterogeneous waypoints. As illustrated in Fig. 5, there are N_g query groups, which consist of temporal, spatial, and driving-style planning queries with n_t, n_s, n_d sampling strategies. Each query group represents a specific planning granularity, encompassing N_{md} modalities such as left, straight, and right trajectories. We construct the multi-granularity planning query matrix, which has a total size $N_p = N_{md} \times N_g$.

Upon processing through the unified decoder, planning queries of varying granularities within a single modality are aligned and aggregated to create a fused query, enhancing information complementarity and overall effectiveness. The fused query is employed to predict waypoints across all granularities, leveraging additional supervision to optimize trajectory. The process is formulated as follows:

$$Q_{fuse}^i = \sum_{j \in N_g} Q_p^{ij}, \quad (3)$$

where i, j are the i -th modality and j -th granularity, respectively. Q_{fuse}^i represents the fused query of i -th modality. We use N_g MLP_{reg} layers to regress different granularity of waypoints $W^{i,j}$, while all granularity share the same modality score layer MLP_{cls-md} ,

$$\begin{aligned} W^{i,j} &= \text{MLP}_{reg}^j(Q_{fuse}^i) \\ S_{md}^i &= \text{MLP}_{cls-md}(Q_{fuse}^i), \end{aligned} \quad (4)$$

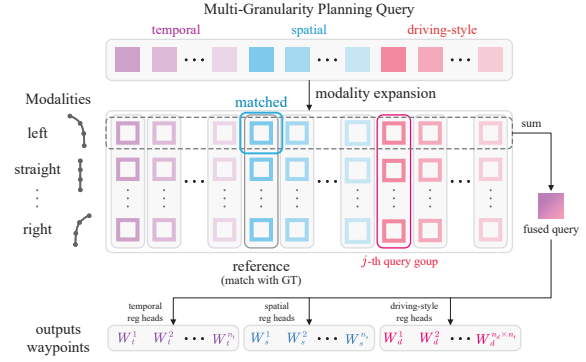


Figure 5. The illustration of multi-granularity query architecture with alignment fusion for waypoints prediction. We omit the classification head for clarity.

where modality score S_{md} is used in the inference step to select the best modality. Additionally, a driving-style classification head is used for final waypoints selection.

$$S_{sty}^i = \text{MLP}_{cls-sty}(Q_{fuse}^i). \quad (5)$$

Align Matching. In the training process, a winner-takes-all matching approach is employed within each query group, which includes all modalities with a specific granularity, to select the optimal modality for optimization. Instead of performing independent matches for each group of waypoints, we introduce an align matching strategy that designates a single group of waypoints as the reference waypoints $W^{i,ref}$ along with its corresponding GroundTruth GT^{ref} for matching and broadcasts to the others:

$$\mathbb{1}_{ref} = \arg \min_i (L_2(W^{i,ref} - \text{GT}^{ref})), \quad (6)$$

where $\mathbb{1}_{ref}$ represents the best matching index in reference query group. All other groups then share the same matching results to align with the matched planning modalities on the remaining query groups. Consequently, the gradients across all granularities of the optimally matched modality can be effectively backpropagated.

Based on this mechanism, n_d driving-style waypoints are selected after the align-matching. Each driving-style waypoints is responsible for an area of velocity. Different from spatial or temporal waypoints that optimize across all granularities, we only select one granularity of driving-style waypoints to optimize, ensuring that each granularity of waypoints learns corresponding actions to various complex driving scenarios. This process is informed by the ground truth of the temporal waypoints.

3.4. Waypoints Selection and Action Control

Selection. During the inference step, the final waypoints are computed through a two-step selection process. First, the optimal modality is selected based on the predicted modality score S_m . Second, specific granularity waypoints are chosen

Method	Open-loop Metric	Closed-loop Metric			
	Avg. L2 ↓	Driving Score ↑	Success Rate(%) ↑	Efficiency ↑	Comfortness ↑
AD-MLP [57]	3.64	18.05	0.00	48.45	22.63
UniAD-Tiny [16]	0.80	40.73	13.18	123.92	47.04
UniAD-Base [16]	0.73	45.81	16.36	129.21	43.58
VAD [23]	0.91	42.35	15.00	157.94	46.01
SparseDrive [49]	0.87	44.54	16.71	170.21	48.63
GenAD [59]	-	44.81	15.90	-	-
DiFSD [48]	0.70	52.02	21.00	178.30	-
DriveTransformer-Large [22]	0.62	63.46	35.01	100.64	20.78
TCP-traj* [55]	1.70	59.90	30.00	76.54	18.08
ThinkTwice* [20]	0.95	62.44	31.23	69.33	16.22
DriveAdapter* [19]	1.01	64.22	33.08	70.22	16.01
HiP-AD	0.69	86.77	69.09	203.12	19.36

Table 1. Open-loop and Closed-loop results of planning in Bench2Drive. Avg. L2 is averaged over the predictions in 2 seconds under 2Hz. * denotes expert feature distillation.

Method	Ability (%) ↑					Mean
	Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	
AD-MLP [57]	0.00	0.00	0.00	0.00	4.35	0.87
UniAD-Tiny [16]	8.89	9.33	20.00	20.00	15.43	14.73
UniAD-Base [16]	14.10	17.78	21.67	10.00	14.21	15.55
VAD [23]	8.11	24.44	18.64	20.00	19.15	18.07
DriveTransformer-Large [22]	17.57	35.00	48.36	40.00	52.10	38.60
TCP-traj* [55]	12.50	22.73	52.72	40.00	46.63	34.92
ThinkTwice* [20]	13.72	22.93	52.99	50.00	47.78	37.48
DriveAdapter* [19]	14.55	22.61	54.04	50.00	50.45	38.33
HiP-AD	50.00	84.44	83.33	40.00	72.10	65.98

Table 2. Multi-Ability Results of E2E-AD Methods in Bench2Drive. * denotes expert feature distillation.

according to predefined rules: dense intervals are selected for spatial waypoints, and high-frequency waypoints are preferred for temporal granularity. For driving-style waypoints, the selection is based on the highest score from the predicted style classifications.

Control. We employ spatial waypoints for lateral control and driving-style waypoints for longitudinal control. An exception occurs when the current speed deviates from the chosen driving-style waypoints, in which case the driving-style waypoints will revert to default temporal waypoints for longitudinal control.

3.5. Loss Functions

HiP-AD can be end-to-end trained and optimized in a fully differentiable manner. The overall optimization function includes four primary tasks (detection, motion prediction, mapping, and planning) and an auxiliary task. Each primary task can be optimized using both classification and regression losses with corresponding weight. The overall loss function can be formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{det} + \mathcal{L}_{motion} + \mathcal{L}_{map} + \mathcal{L}_{plan} + \mathcal{L}_{aux}. \quad (7)$$

Specifically, the planning loss integrates a multi-granularity waypoint regression loss and classification loss for modality and driving-style:

$$\mathcal{L}_{plan} = \sum_{j=1}^{N_g} \mathcal{L}_{reg}^j + \mathcal{L}_{cls-md} + \mathcal{L}_{cls-sty}. \quad (8)$$

4. Experiments

4.1. Dataset and Metrics

Dataset. In order to comprehensively evaluate the performance of the end-to-end autonomous driving algorithm, our experiments are mainly conducted on challenging closed-loop benchmark Bench2Drive [21] dataset, which collects 1000 short clips uniformly distributed under 44 interactive scenarios in CARLA v2 [12]. There are 950 clips for training and 50 clips for open-loop validation. Furthermore, it provides closed-loop evaluation protocols under 220 test routes for fair comparison. We also evaluate the open-loop performance on the realistic dataset nuScenes [2], which comprises 1,000 videos divided into training, validation, and testing sets with 700, 150, and 150 videos, respectively.

Metrics. For closed-loop evaluation, we employ the official evaluation metrics recommended in Bench2Drive: Driving Score (DS), Success Rate (SR), Efficiency (Eff.), and Comfortness (Com.). For open-loop evaluation in both dataset, we follow previous works [16, 23], using L2 Displacement Error (L2) and Collision Rate (CR) to measure the planning trajectory. The results of 3D object detection, tracking, and online mapping on nuScenes dataset are also reported with commonly used metrics [2, 27, 30].

4.2. Implementation Details

We adopt ResNet50 as the backbone with 6 unified decoder layers, utilizing an input resolution of 640×352 in

Method	L2 (m) ↓				Collision (%) ↓				Latency ↓	FPS ↑
	1s	2s	3s	Avg.	1s	2s	3s	Avg.		
VAD-Base [23]	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21	224.3	4.5
GenAD [59]	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19	149.2	6.7
SparseDrive-S [49]	0.29	0.58	0.96	0.61	0.01	0.05	0.18	0.08	101.0	9.9
DiFSD-S [48]	0.15	0.31	0.56	0.33	0.00	0.06	0.19	0.08	93.7	10.7
DriveTransformer-Large [22]	0.16	0.30	0.55	0.33	0.01	0.06	0.15	0.07	221.7	4.5
HiP-AD	0.28	0.53	0.87	0.56	0.01	0.05	0.15	0.07	109.9	9.1

Table 3. Open-loop planning evaluation results on the nuScenes validation dataset with the evaluation protocol [28]. Latency and FPS of DriveTransformer [22] are measured on NVIDIA H800 GPU while the others are measured on NVIDIA 3090 GPU.

Method	detection		map	track	motion
	mAP↑	NDS↑	mAP↑	AMOTA↑	
UniAD [16]	0.380	0.359	-	-	-
VAD [23]	0.276	0.397	0.476	-	-
SparseDrive-S [49]	0.418	0.525	0.551	0.386	0.62
DiFSD [48]	0.410	0.528	0.560	-	-
HiP-AD	0.424	0.535	0.571	0.406	0.61

Table 4. Comparison of perception, mapping, tracking, and motion prediction performance on the nuScenes validation dataset.

Variants	PDA	MG	Closed-loop	
			Driving Score↑	Success Rate(%)↑
Sequential	✓	✓	73.2	45.5
Unified	-	-	41.3	16.4
	✓	-	49.7	25.5
	-	✓	76.4	47.2
	✓	✓	88.3	72.7

Table 5. Ablation study of the architecture and proposed modules. PDA: Planning Deformable Attention; MG: Multi-Granularity.

Bench2Drive, which serves as the default dataset for our experiments. We establish fixed quantities for hybrid task queries, consisting of 900 agents, 100 maps, and 480 planning queries. The total planning query comprises 10 granularities, with each serving as a planning group encompassing 48 modalities. These granularities include spatial waypoints sampled at uniform intervals of 2m and 5m, temporal waypoints sampled at frequencies of 2Hz and 5Hz, and manual divisions of driving styles across three speed ranges: $[0, 0.4)$, $[0.4, 3)$, and $[3, 10)$ m/s, each with two frequency settings. Target points and high-level commands are embedded into the planning query through MLPs before the regression head, while the ego status is excluded from inputs which is only used for supervision in training.

The training process consists of two phases. Initially, we disable the driving-style head for 12 epochs, followed by 6 epochs of fine-tuning with the driving-style head enabled. We train the model on 8 NVIDIA 4090 GPUs with a total batch size of 32. The AdamW optimizer and Cosine Annealing scheduler are utilized with an initial learning rate of 2×10^{-4} and a weight decay of 0.01. Training on the nuScenes dataset using a similar process. Additional implementation details are provided in the appendix.

4.3. Main Results

Bench2Drive. We evaluate HiP-AD against state-of-the-art end-to-end autonomous driving methods on the Bench2Drive dataset, with results presented in Tab. 1. HiP-AD achieves the best closed-loop performance, demonstrating superior Driving Score and Success Rate. Compared to the second-place method [22], it significantly improves by over 20% in Driving Score and 30% in Success Rate. Furthermore, HiP-AD attains a comparable L2 error score when compared to other leading methods. The primary limitation of our approach is in Comfortness. However, we emphasize that comparing Comfortness is meaningful only among methods with similar Success Rate scores. Behaviors such as sudden braking or turning, while potentially reducing Comfortness, are often necessary to ensure the successful completion of the evaluation.

Additionally, we present the multi-ability scores in Tab. 2, highlighting HiP-AD’s exceptional performance across diverse driving scenarios. HiP-AD significantly enhances capabilities in scenarios such as Merging, Overtaking, Emergency Brake, and Traffic Sign, leading to an overall score improvement of over 25%. Give Way illustrates some rare driving scenarios that require ego vehicle to yield to an emergency vehicle coming from behind. Considering the imbalance datasets, making quick decisions is still a challenge for E2E-AD methods.

NuScenes. We further evaluate HiP-AD’s performance on the open-loop dataset nuScenes, focusing on perception and motion prediction tasks. As shown in Tab. 3, HiP-AD achieves the lowest Collision Rate among all compared methods while maintaining a competitive L2 error. The perception and prediction results, presented in Tab. 4, demonstrate that HiP-AD delivers strong performance in both tasks, highlighting the robustness and effectiveness of the proposed unified framework.

4.4. Ablation Study

In the ablation experiments, we use a small test set of Bench2Drive to save computational resources. This subset consists of 55 routes (25% of the total 220 routes), with 44 routes selected one-to-one from 44 unique scenarios and the remaining 11 routes randomly chosen from the rest routes.

Index	Fusion	Align-Matching	Multi-Granularity Details						Control		Closed-loop		
			Temporal 2Hz	5Hz	Spatial 5m 2m		Driving Style sty-2Hz sty-5Hz		Lon.	Lat.	Driving Score \uparrow	Success Rate(%) \uparrow	Time Out(%) \downarrow
1			✓						2Hz	2Hz	49.7	25.5	41.8
2			✓			✓			2Hz	5m	53.6	29.0	27.3
3	✓		✓			✓			2Hz	5m	62.4	29.0	20.0
4	✓	✓	✓			✓			2Hz	5m	79.5	56.3	5.5
5	✓	✓	✓	✓		✓	✓		5Hz	2m	82.1	60.0	3.6
6	✓	✓	✓	✓	✓	✓	✓		5Hz	2m	84.2	65.5	1.8
7	✓	✓	✓	✓	✓	✓	✓	✓	sty-5Hz	2m	88.3	72.7	1.8

Table 6. Ablation study of multi-granularity planning. Lon. and Lat. refer to longitudinal and lateral control.

Effect of architecture and modules. As shown in Tab. 5, we evaluate the contributions of planning deformable attention and multi-granularity representation. The results indicate that both components play a critical role in enhancing overall performance, with multi-granularity particularly noteworthy for its provision of improved control. Additionally, we compare the proposed unified framework with its sequential variant. In the sequential version, the perception components of HiP-AD are executed first, followed by the planning components. In contrast, the unified version runs perception and planning iteratively in parallel. Experimental results reveal that the unified version significantly outperforms the sequential variant, demonstrating the superiority of the unified framework.

Effect of multi-granularity planning. Tab. 6 presents the ablation study on multi-granularity planning query design. The 1st setting employs only 2Hz temporal waypoints, consistent with VAD [23] or UniAD [16]. The 2nd setting incorporates both temporal and spatial waypoints, similar to CarLLaVA [43]. The 3rd and 4th settings use the same waypoints as the 2nd but introduces granularity fusion and align-matching, demonstrating significant performance improvements. The 5th setting utilizes 5Hz and 2m waypoints, highlighting that higher-frequency waypoints enhance fine-grained control. The 6th setting combines dense and sparse sampling intervals, showing that granularity fusion of sampling intervals boosts performance. The 7th setting integrates driving style, achieving a 7% improvement in Success Rate and delivering the best overall performance. The waypoints used for control correspond to the granularity settings, as detailed in Tab. 6. Additionally, the hesitation phenomenon is measured by “Agent timed out” status, which shows the multi-granularity planning not only maintain safe driving, but also encourages behavior learning.

4.5. Qualitative Results

We visualize the closed-loop results on the Bench2Drive test routes, as shown in Fig. 6. Two typical scenarios are illustrated: the first involves an unprotected left turn, and the second features a hazard in the roadside lane. In the first scenario, guided by spatial waypoints, the ego vehicle successfully executes a left turn, adhering to the target trajectory and changing lanes to avoid the bicycle. In the second

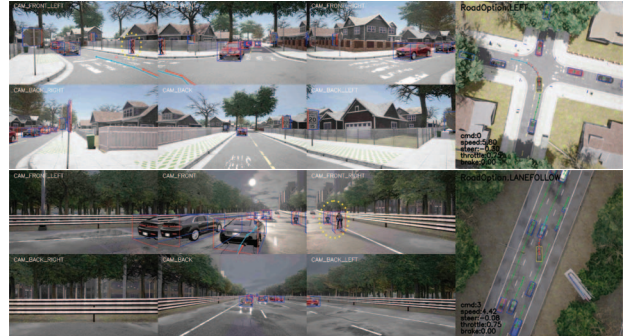


Figure 6. Qualitative results of HiP-AD on closed-loop routes, including perception, motion, and planning trajectories. Spatial waypoints are colored in skyblue, while driving-style waypoints are colored in red. Important objects are highlight in yellow circle.

scenario, the ego vehicle decelerates in accordance with the proposed driving-style waypoints upon detecting obstacles, such as a pedestrian on the side and a car ahead. These results demonstrate the effectiveness of HiP-AD in handling complex driving scenarios. Additional qualitative results are provided in the supplementary materials.

5. Conclusion and Limitation

In this paper, we introduce HiP-AD, a novel end-to-end autonomous driving framework. HiP-AD features a unified decoder capable of simultaneously executing perception, prediction, and planning tasks. The planning query iteratively interacts with perception features in BEV space and multi-view image features in the perspective view, enabling comprehensive interaction. Additionally, we propose a multi-granularity planning strategy that integrates diverse planning trajectories with rich supervision to enhance ego-vehicle control. Extensive experiments on both closed-loop and open-loop datasets demonstrate outstanding planning performance, verifying the effectiveness of HiP-AD in complex driving scenarios.

Limitations: Despite achieving excellent performance in both closed-loop and open-loop evaluations, extensive real-world testing is still necessary. Furthermore, avoiding collisions with vehicles rapidly approaching from behind remains a challenge. These issues will be a focus of our future research.

References

- [1] Frédéric Bouchard, Sean Sedwards, and Krzysztof Czarnecki. A rule-based behaviour planner for autonomous driving. In *International Joint Conference on Rules and Reasoning*, pages 263–279. Springer, 2022. 3
- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11618–11628, 2020. 1, 2, 6
- [3] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018. 3
- [4] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. Gri: General reinforced imitation and its application to vision-based autonomous driving. *Robotics*, 12(5):127, 2023. 3
- [5] Dian Chen, Vladlen Koltun, and Philipp Krähenbühl. Learning to drive from a world on rails. In *ICCV*, pages 15590–15599, 2021. 3
- [6] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 1, 2, 3
- [7] Zhili Chen, Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In *ECCV*, pages 239–256. Springer, 2024. 3
- [8] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, pages 4693–4700. IEEE, 2018. 3
- [9] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *ICCV*, pages 9329–9338, 2019. 2, 3
- [10] Alexander Cui, Sergio Casas, Abbas Sadat, Renjie Liao, and Raquel Urtasun. Lookout: Diverse multi-future prediction and planning for self-driving. In *ICCV*, pages 16107–16116, 2021. 3
- [11] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning*, pages 1268–1281. PMLR, 2023. 1, 3
- [12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio López, and Vladlen Koltun. Carla: An open urban driving simulator. *Conference on Robot Learning*, 2017. 6
- [13] Haoyang Fan, Fan Zhu, Changchun Liu, Liangliang Zhang, Li Zhuang, Dong Li, Weicheng Zhu, Jiangtao Hu, Hongye Li, and Qi Kong. Baidu apollo em motion planner. *arXiv preprint arXiv:1807.08048*, 2018. 3
- [14] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *CVPR*, pages 11525–11533, 2020. 3
- [15] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *CVPR*, pages 5496–5506, 2023. 3
- [16] 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 *CVPR*, pages 17853–17862, 2023. 1, 2, 3, 6, 7, 8
- [17] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3
- [18] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. *ICCV*, pages 8206–8215, 2023. 2
- [19] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *ICCV*, 2023. 6
- [20] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *CVPR*, 2023. 6
- [21] 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 *NeurIPS*, 2024. 1, 2, 6
- [22] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *ICLR*, 2025. 2, 4, 6, 7
- [23] 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 *ICCV*, 2023. 1, 2, 3, 5, 6, 7, 8
- [24] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *ICRA*, pages 8248–8254. IEEE, 2019. 3
- [25] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *IEEE International Conference on robotics and automation (ICRA)*, pages 4628–4634, 2022. 3
- [26] Yin hao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, pages 1477–1485, 2023. 3
- [27] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18, 2022. 3, 6
- [28] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, pages 14864–14873, 2024. 7

- [29] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *CVPR*, 2020. 3
- [30] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022. 3, 6
- [31] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xingang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025. 4
- [32] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 3
- [33] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023. 4
- [34] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, pages 18580–18590, 2023. 4
- [35] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548, 2022. 3
- [36] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In *ICML*, pages 22352–22369. PMLR, 2023. 3
- [37] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *ECCV*, pages 461–477. Springer, 2025. 3
- [38] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *CVPR*, pages 14760–14769, 2024. 4
- [39] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 3
- [40] Stefano Pini, Christian S Perone, Aayush Ahuja, Ana Sofia Rufino Ferreira, Moritz Niendorf, and Sergey Zagoruyko. Safe real-world autonomous driving by learning to predict and plan with a mixture of experts. In *ICRA*, pages 10069–10075. IEEE, 2023. 1, 3
- [41] Dean A Pomerleau. Alvin: An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1, 1988. 3
- [42] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, pages 7077–7087, 2021. 2, 3
- [43] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünemann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. Carllava: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 2, 4, 5, 8
- [44] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *ECCV*, pages 414–430. Springer, 2020. 3
- [45] Hao Shao, Letian Wang, RuoBing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. *arXiv preprint arXiv:2207.14024*, 2022. 2
- [46] 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 *CVPR*, pages 15120–15130, 2024. 4
- [47] Kangjing Shi, Li Huang, Du Jiang, Ying Sun, Xiliang Tong, Yuanming Xie, and Zifan Fang. Path planning optimization of intelligent vehicle based on improved genetic and ant colony hybrid algorithm. *Frontiers in Bioengineering and Biotechnology*, 10:905983, 2022. 3
- [48] Haisheng Su, Wei Wu, and Junchi Yan. Difs: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2409.09777*, 2024. 1, 3, 6, 7
- [49] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *ICRA*, 2025. 1, 3, 4, 6, 7
- [50] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *CVPR*, pages 7153–7162, 2020. 3
- [51] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000. 1, 3
- [52] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 3621–3631, 2023. 3
- [53] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *COLR*, pages 180–191, 2021. 3
- [54] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *CVPR*, pages 15449–15458, 2024. 1, 2, 3
- [55] 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 *NeurIPS*, 2022. 6
- [56] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *WACV*, pages 7356–7365, 2024. 3
- [57] Jiang-Tian Zhai, Ze Feng, Jihao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 6

- [58] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. [2](#), [3](#)
- [59] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *ECCV*, pages 87–104. Springer, 2024. [1](#), [4](#), [6](#), [7](#)
- [60] Julius Ziegler, Philipp Bender, Thao Dang, and Christoph Stiller. Trajectory planning for berth—a local, continuous method. In *IEEE Intelligent Vehicles Symposium Proceedings*, pages 450–457. IEEE, 2014. [3](#)