

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

## Supplementary Material

### A. Comparison with Previous E2E-AD Methods

**Comparison.** We summarize two key differences between HiP-AD and previous E2E-AD methods, primarily in planning query design and planning interaction, as illustrated in Tab. 1. Compared to previous E2E-AD models, we introduce driving-style waypoints prediction to enhance the precision of autonomous vehicle steering and acceleration/deceleration. Furthermore, we employ a multi-granularity strategy to provide complementary long-short information, enriching the additional supervision. Moreover, we develop a comprehensive interaction mechanism in which the planning query not only interacts with perception tasks but also engages with sparse image features through planning deformable attention, facilitating the learning of unannotated scene information.

**Motivation of driving-style waypoints.** Previous methods primarily relied on temporal waypoints with different modalities (e.g., left, straight, right) to enhance the lateral diversity. However, they explicitly overlook longitudinal diversity, which poses challenges for a single lateral modality to effectively address all scenarios requiring various ego velocities, such as starting, braking, and overtaking. To address this issue, driving-style waypoints are proposed to decouple a single trajectory into several waypoints that correspond to various speed intervals, enhancing longitudinal diversity.

### B. More Implementation Details

**Supervision.** All ground truth (GT) of multi-granularity waypoints are derived from the future trajectory of the ego vehicle, differing primarily in the sampling method or strategy. For example, we gather all future locations of the ego car within a single video and apply linear fitting to obtain a trajectory function. Then, we can sample from this function at any equal distance intervals to attain the ground truth of spatial waypoints. In contrast, temporal and driving-style predictions share the same ground truth of temporal waypoints, which is sampled directly from future locations at equal time intervals.

In addition to directly supervising perception and trajectory, we also incorporate additional supervisory signals to enhance the training process. The auxiliary supervision includes a sparse depth map and ego status, which can be formulated as follows:

$$\mathcal{L}_{aux} = \mathcal{L}_{depth} + \mathcal{L}_{status}. \quad (1)$$

**Ablation Routes.** There are 220 routes in the Bench2Drive

closed-loop test routes, officially divided into 44 scenarios based on various weather conditions and scene events. It is computationally exhaustive for us to conduct all experiments on the total number of test routes. Therefore, to conserve computational resources, we utilize a small test set on Bench2Drive for our closed-loop ablation experiments. It consists of 55 routes (25% of 220) which can be divided into two parts. The first part includes 44 routes, selected one-to-one from each of the 44 scenarios, while the remaining 11 routes are chosen randomly. This small test set not only includes all scenarios but also maintains the distribution of the total test routes for comprehensive evaluation. We list route IDs in Tab. 2.

**Speed Setting.** The speed of driving-style waypoints is empirically divided into three intervals: [0, 0.4), [0.4, 3), and [3, 10). Each interval corresponds to different driving scenarios.

- The interval 0-0.4 corresponds to parking maneuvers.
- The interval 0.4-3 corresponds to low-speed actions, such as strategic lane changes and stop-and-go situations.
- The interval 3-10 corresponds to normal driving.

**NuScenes Training.** The training parameters on nuScenes dataset are similar to those used in Bench2Drive, except that the resolution is changed to  $704 \times 256$  for fair comparison. Moreover, considering the different intentions between open-loop and closed-loop evaluations, we set different training process for nuScenes. First, since nuScenes does not require precise control, we disable driving-style waypoints during training and use temporal waypoints for performance evaluation. Second, as a multi-task learning framework, we follow the training strategy of SparseDrive [6] to maximize perception performance, which entails first training the perception component, followed by motion prediction and planning.

**Closed-loop Inference.** Although we employ waypoints at varying frequencies for supervision, our model is trained and inferenced at 2Hz. Considering that the operational frequency during the Bench2Drive closed-loop simulation is 10Hz, we set up multiple memory modules (e.g., 5) to ensure the inference frequency aligns with the training phase at 2Hz. It means only one memory is available at each timestamp to output temporal task queries and keep top  $k$  updated task queries. Following [6], we set  $k$  as 600 for agent queries, 0 for map queries, and 480 for planning queries. Additionally, we list the inference time and parameters of the model in Tab. 3.

Method	Planning waypoints				Planning interaction			
	heterogeneous waypoints		driving-style	multi-granularity	perception	scenes representations		
default / temporal	spatial	BEV				global images	sparse images	
UniAD [1]	✓				✓			
VAD [3]	✓				✓			
Para-Drive [7]	✓					✓		
GenAD [8]	✓					✓		
SparseDrive [6]	✓				✓			
DiFSD [5]	✓				✓			
DriveTransformer [2]	✓				✓		✓	
CarLLaVA [4]	✓	✓			✓		✓	
HiP-AD (ours)	✓	✓	✓	✓	✓			✓

Table 1. Comparison of representative related models and our HiP-AD in terms of planning query design and planning interaction.

Scenarios	Route IDs
44 scenarios	1711, 1773, 1790, 1825, 1852, 1956, 2050, 2082, 2084, 2115, 2127, 2144, 2164, 2201, 2204, 2273, 2286, 2373, 2390, 2416, 2509, 2534, 2664, 2709, 2790, 3086, 3248, 3364, 3436, 3464, 3540, 3561, 3936, 14194, 14842, 17563, 17752, 23658, 23695, 23771, 23901, 26458, 28087, 28099
11 random routes	1792, 2086, 2129, 2283, 2539, 2668, 26406, 26956, 27494, 27532, 28154

Table 2. A small test set of Bench2Drive for ablation experiments.

Dataset	Parameters	GFLOPs	Latency	FPS
nuScenes	90.0M	202.9	109.9ms	9.1
Bench2Drive	97.4M	256.9	138.9ms	7.2

Table 3. The model is measured on a single NVIDIA 3090 GPU.

Control		Closed-loop Metric	
Lon.	Lat.	Driving Score "	Success Rate(%) "
2Hz	5m	76.35	52.72
2Hz	2m	79.03	56.36
5Hz	5m	81.85	65.45
5Hz	2m	<b>86.05</b>	<b>69.09</b>

Table 4. Ablation study on controlling ego-vehicle under different configurations.

## C. More Experiments

**Ganularity Selection.** We output spatial waypoints at dense and sparse distance intervals, along with high-frequency and low-frequency temporal waypoints, as well as driving-style waypoints. Therefore, we conduct an ablation study to explore which configuration yields better vehicle control in closed-loop system. As shown in the 4, there are four combinations: dense and sparse intervals, as well as high and low frequencies. Among these, the sparse-interval and low-frequency combination perform the worst, while the dense-interval and high-frequency combination provide the best vehicle control. Therefore, we chose the dense-interval and high-frequency combination as the final control method for our model. It is worth noting that even the worst combination

Number	Closed-loop Metric			
	DS "	SR(%) "	Eff. "	Com. "
18	86.56	62.27	197.21	17.35
48	<b>86.77</b>	<b>69.09</b>	<b>203.12</b>	<b>19.36</b>

Table 5. Comparison of the number of modalities: ‘Eff.’ and ‘Com.’ are abbreviations for Efficiency and Comfort.

our approach still outperforms the previous state-of-the-art methods. This clearly demonstrates the superiority of our method.

**Number of Modalities.** Previous methods [6] usually use 18 modalities in Bench2Drive. We provide our 18 modalities results in Tab. 5. The Driving Score, Efficiency, and Comfortness are similar on modalities 18 and 48. The 48 modalities version improves over 6% on Success Rate, which indicates increasing the number of modalities also contribute to the performance of the final completion. It is noticed that the 18 modalities version is still much better than other methods in main experiments.

**Sparse Interaction.** We compare the performance of planning queries with global and sparse interactions of image features. In the Global setting, considering memory constraints, we employ global attention, where the planning query interacts with 1/16 downsample image features. In contrast to Global setting, we utilize the proposed planning deformable attention to dynamically sample image features around trajectory points for sparse interaction. As shown in Tab. 6, the Sparse setting achieves better driving, demonstrating that sparse local interactions can effectively learn

Setting	Interaction	Closed-loop Metric	
		Driving Score "	Success Rate(%) "
Global	Cross Attention	73.4	49.1
Sparse	Deformable Attention	<b>84.2</b>	<b>65.5</b>

Table 6. Comparison of Global and Sparse Interactions.

latent scene representations, which benefits closed-loop systems.

## D. More Qualitative Analysis

We present additional visualization results of our HiP-AD to demonstrate its effectiveness in both open-loop and closed-loop evaluations.

### D.1. Open-Loop

**Bench2Drive.** As shown in Fig. 2, we visualize three driving scenarios on the open-loop validation set of Bench2Drive in both daytime and nighttime. The three driving scenarios are giving way to pedestrians, left turn at an intersection, and detouring around obstacles. We use cyan-blue lines to represent spatial waypoints, while the purple-red lines indicate driving-style waypoints.

In the first driving scenario Fig. 2a, HiP-AD detects most vehicles at the intersection, reconstructing the map and providing suitable trajectories to ensure that the ego vehicle can successfully navigate a left turn despite missing lane markings. In the second driving scenario Fig. 2b, HiP-AD detects a pedestrian crossing the road in front of the ego vehicle and predicts waypoints for a cautious driving style, reducing speed to avoid a collision with the pedestrian. The third scenario Fig. 2c showcases HiP-AD’s ability to navigate around obstacles, even in complex situations such as starting from a stop or driving at night.

**NuScenes.** We also provide visualization results on the nuScenes dataset, as shown in Fig. 3. These driving scenarios include left turns, straight driving, and right turns under realistic conditions, such as at intersections or low illumination. Compared to ground truth, our model effectively locates obstacles, reconstructs map elements, and produces planning results that align closely with ground truth. This demonstrates the robustness of our method in perception and planning within real-world scenarios.

### D.2. Closed-loop

**Qualitative Results.** The closed-loop visualization results of Bench2Drive are shown in Fig. 4. It illustrates the decision-making process and precise control of HiP-AD across various scenarios, including urban streets, intersections, and highways, under conditions such as nighttime and fog. Thanks to the proposed strategies and the unified architecture, HiP-AD demonstrates strong adaptability and robustness in these closed-loop scenarios.

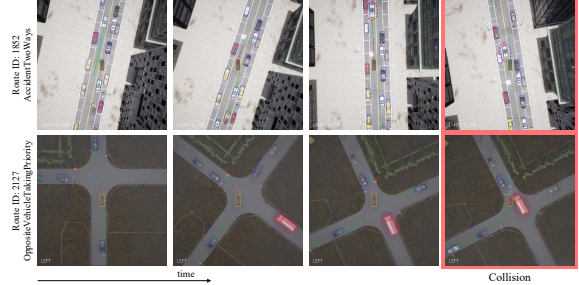
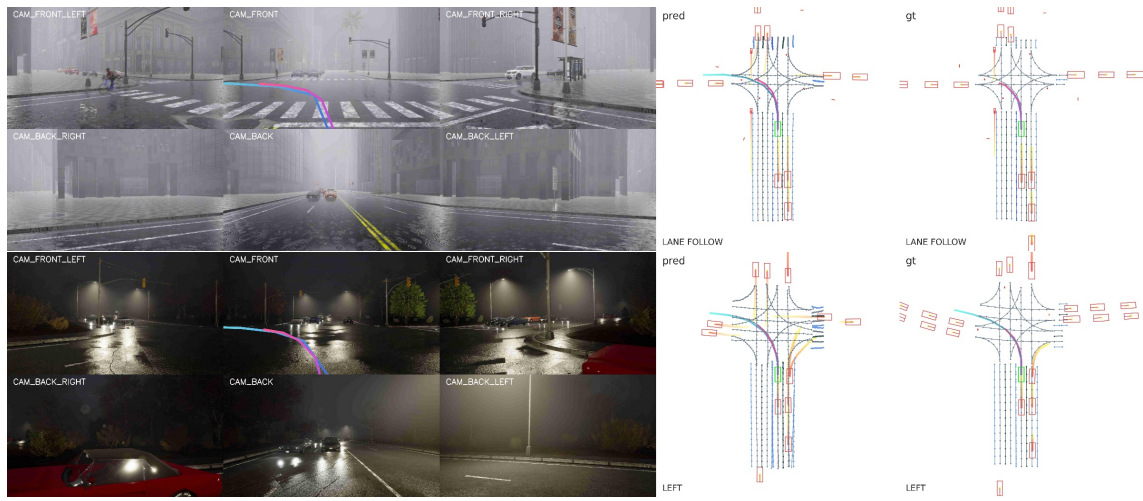


Figure 1. Illustration of failure cases on closed-loop test routes.

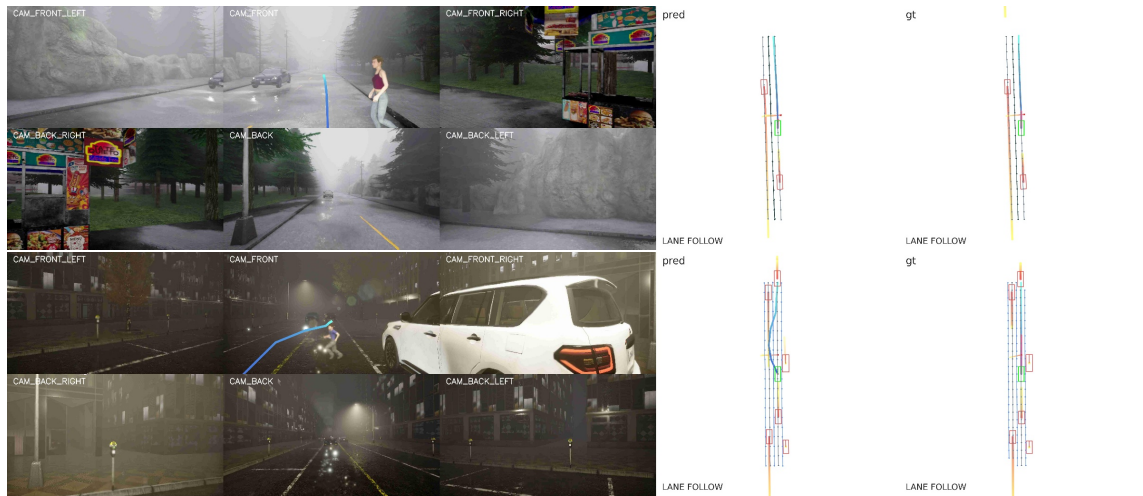
**Limitation.** We illustrate some typical limitation cases as illustrated in Fig. 1. HiP-AD fails to yield to oncoming vehicles when overtaking and give way to emergency vehicles when turning. A potential explanation is that HiP-AD lacks the capability for long-range perception and active avoidance, which will be focused on our future work.

## References

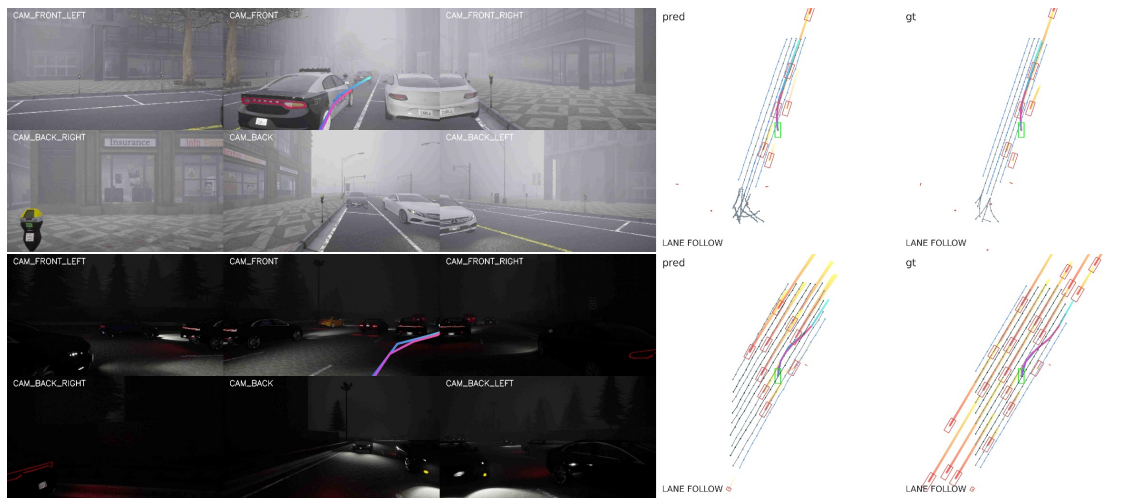
- [1] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 2
- [2] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *ICLR*, 2025. 2
- [3] 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. 2
- [4] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünermann, 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
- [5] Haisheng Su, Wei Wu, and Junchi Yan. Difs-d: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2409.09777*, 2024. 2
- [6] 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, 2
- [7] 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. 2
- [8] 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. 2



(a) Left turn at an intersection.



(b) Give way to pedestrians.



(c) Detour around obstacles.

Figure 2. Illustration of open-loop results on Bench2Drive validation dataset.

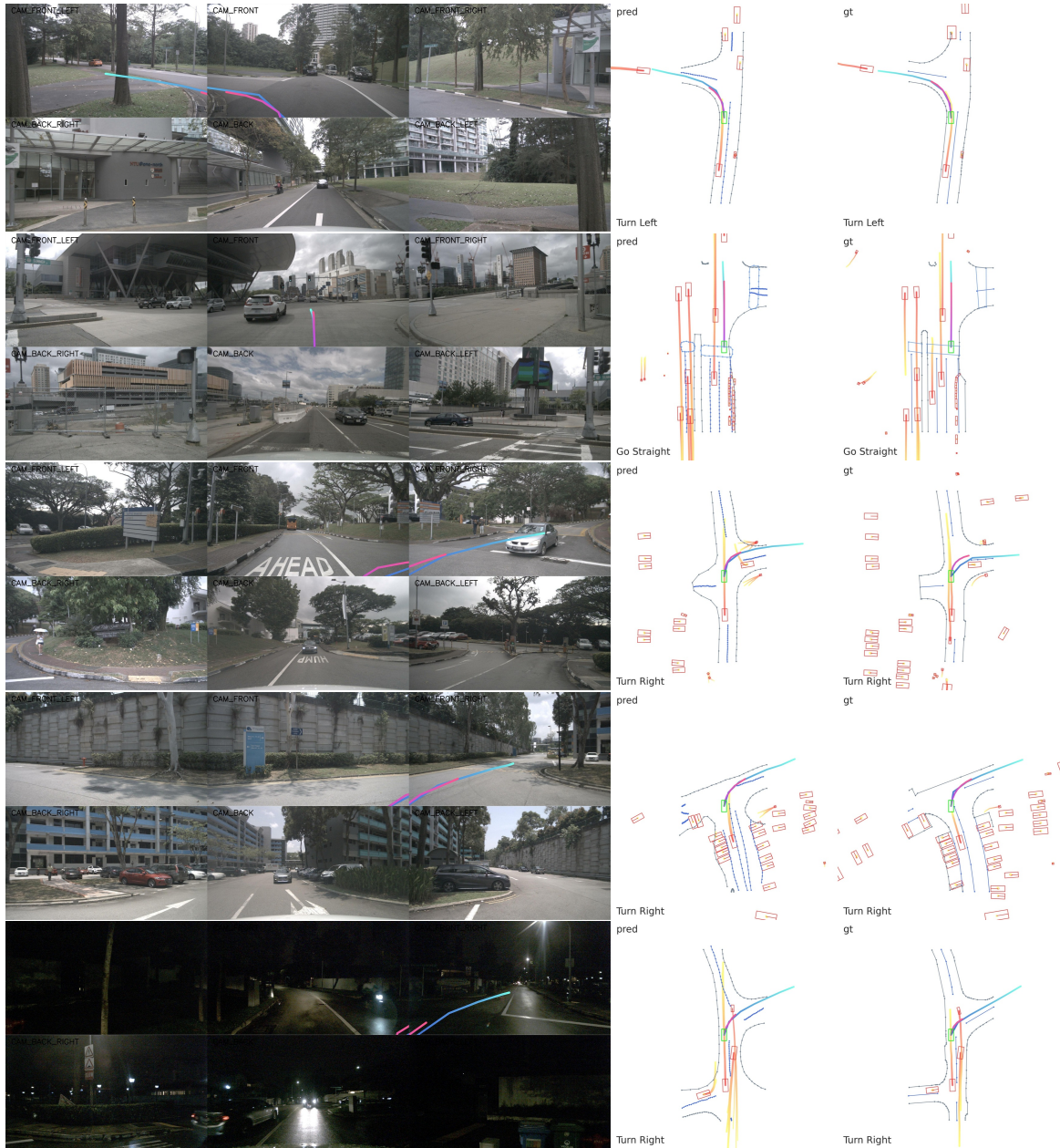


Figure 3. Illustration of open-loop results on nuScenes validation dataset.

