DiFSD: Ego-Centric Fully Sparse Paradigm with Uncertainty Denoising and Iterative Refinement for Efficient End-to-End Self-Driving

Authors: Haisheng Su, Wei Wu, Cewu Lu

Date: 2024-09-14T00:00:00+00:00

Abstract

Current end-to-end autonomous driving methods resort to unifying modular designs for various tasks (e.g. perception, prediction and planning). Although optimized in a planning-oriented spirit with a fully differentiable framework, existing end-to-end driving systems without ego-centric designs still suffer from unsatisfactory performance and inferior efficiency, owing to the rasterized scene representation learning and redundant information transmission. In this paper, we revisit the human driving behavior and propose an ego-centric fully sparse paradigm, named DiFSD, for end-to-end self-driving. Specifically, DiFSD mainly consists of sparse perception, hierarchical interaction and iterative motion planner. The sparse perception module performs detection, tracking and online mapping based on sparse representation of the driving scene. The hierarchical interaction module aims to select the Closest In-Path Vehicle / Stationary (CIPV / CIPS) from coarse to fine, benefiting from an additional geometric prior. As for the iterative motion planner, both selected interactive agents and ego-vehicle are considered for joint motion prediction, where the output multi-modal ego-trajectories are optimized in an iterative fashion. Besides, both position-level motion diffusion and trajectory-level planning denoising are introduced for uncertainty modeling, thus facilitating the training stability and convergence of the whole framework. Extensive experiments conducted on nuScenes dataset demonstrate the superior planning performance and great efficiency of DiFSD, which significantly reduces the average L2 error by 66% and collisionrate by 77% than UniAD while achieves 6.9x faster running efficiency.

Full Text

Preamble

DiFSD: Ego-Centric Fully Sparse Paradigm with Uncertainty Denoising and Iterative Refinement for Efficient End-to-End Self-Driving

Haisheng Su[1],[2], Wei Wu[2], Cewu Lu[1]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]SenseAuto Research
{suhaisheng, lucewu}@sjtu.edu.cn, wuwei@senseauto.com

Abstract

Current end-to-end autonomous driving methods resort to unifying modular designs for various tasks (e.g., perception, prediction, and planning). Although optimized in a planning-oriented spirit with a fully differentiable framework, existing end-to-end driving systems without ego-centric designs still suffer from unsatisfactory performance and inferior efficiency, owing to rasterized scene representation learning and redundant information transmission. In this paper, we revisit human driving behavior and propose an ego-centric fully sparse paradigm, named DiFSD, for end-to-end self-driving. Specifically, DiFSD mainly consists of sparse perception, hierarchical interaction, and iterative motion planner modules. The sparse perception module performs detection, tracking, and online mapping based on sparse representation of the driving scene. The hierarchical interaction module aims to select the Closest In-Path Vehicle/Stationary (CIPV/CIPS) from coarse to fine, benefiting from an additional geometric prior. As for the iterative motion planner, both selected interactive agents and ego-vehicle are considered for joint motion prediction, where the output multi-modal ego-trajectories are optimized in an iterative fashion. Additionally, both position-level motion diffusion and trajectory-level planning denoising are introduced for uncertainty modeling, thus facilitating training stability and convergence of the whole framework. Extensive experiments conducted on the nuScenes dataset demonstrate the superior planning performance and great efficiency of DiFSD, which significantly reduces the average L2 error by 66% and collision rate by 77% compared to UniAD while achieving 8.2× faster running efficiency.

1. Introduction

Autonomous driving has experienced notable progress in recent years. Traditional driving systems are commonly decoupled into several standalone tasks, e.g., perception, prediction, and planning. However, heavily relying on handcrafted post-processing, these well-established modular systems suffer from information loss and error accumulation across sequential modules. Recently, the end-to-end paradigm integrates all tasks into a unified model for planning-

oriented optimization, showcasing great potential in pushing the limits of autonomous driving performance.

Literally, existing end-to-end models designed for reliable trajectory planning can be classified into two mainstreams as summarized in Fig. 1(a) and (b). The dense BEV-Centric paradigm performs perception, prediction, and planning consecutively upon shared BEV (Bird's Eye View) features, which are computationally expensive and lead to inferior efficiency. The sparse Query-Centric paradigm utilizes sparse representation to achieve scene understanding and joint motion planning, thus improving overall efficiency. However, object-intensive motion prediction inevitably causes computational redundancy and violates the driving habits of human drivers, who usually only concentrate on the Closest In-Path Vehicle/Stationary (CIPV/CIPS) which are more likely to affect the driving intention and trajectory planning of the ego-vehicle. Meanwhile, excessive interaction with irrelevant agents can conversely be adverse to ego-planning. Therefore, the planning performance remains unsatisfactory in terms of safety, comfort, and personification.

To this end, we propose DiFSD, an Ego-Centric fully sparse paradigm as shown in Fig. 1(c). Specifically, DiFSD mainly consists of sparse perception, hierarchical interaction, and iterative motion planner modules. In the sparse perception module, multi-scale image features extracted from the visual encoder are adopted for object detection, tracking, and online mapping simultaneously in a sparse manner. Then the hierarchical interaction performs ego-centric and object-centric dual interaction to select the CIPV/CIPS with the help of an additional geometric prior, allowing interactive queries to be selected gradually from coarse to fine. As for the motion planner, the mutual information between sparse interactive queries and ego-query is considered for motion prediction in a joint decoder, which is neglected in previous methods but is essential especially in scenarios like intersections. To ensure planning rationality and selection accuracy of interactive queries, iterative planning optimization is further applied to the multi-modal proposal ego-trajectories through continually updating the reference line and ego-query. Moreover, both position-level motion diffusion and trajectory-level planning denoising are introduced for stable training and fast convergence. This can not only model the uncertain positions of interactive agents for motion prediction but also enhance the quality of trajectory refinement with arbitrary offsets. With these elaborate designs, DiFSD exhibits the great potential of the fully sparse paradigm for end-to-end autonomous driving, significantly reducing the average L2 error by 66% and collision rate by 77% compared to UniAD. Notably, our DiFSD-S achieves 8.2× faster running efficiency as well.

In summary, the main contributions of our work are as follows:

- We propose an ego-centric Fully Sparse paradigm for end-to-end self-Driving, named DiFSD, without any computationally intensive dense scene representation learning or redundant environmental modeling, which is proven to be effective and efficient for ego-planning.

- We introduce a geometric prior through intention-guided attention, where the Closest In-Path Vehicle/Stationary (CIPV/CIPS) are gradually picked out through ego-centric cross attention and selection. Additionally, both position-level diffusion of interactive agents and trajectory-level denoising of ego-vehicle are adopted for uncertainty modeling of motion planning respectively.
- Extensive experiments are conducted on the famous nuScenes dataset for planning performance evaluation, which demonstrate the superiority and prominent efficiency of our DiFSD method, revealing the great potential of the proposed ego-centric fully sparse paradigm.

2. Related Work

2.1. End-to-End Perception

Recent years have witnessed remarkable progress in multi-view 3D detection, which mainly builds elaborate designs upon dense BEV (Bird's Eye View) or sparse query features. To generate BEV features, LSS lifts 2D image features to 3D space using depth estimation results, which are then splatted into the BEV plane. Follow-up works apply such operations to perform view transformation for 3D detection tasks and have made significant improvements in both detection performance and efficiency. Differently, some works project a series of predefined BEV queries in 3D space to the image domain for feature sampling. As for the sparse fashion, current methods adopt a set of sparse queries to integrate spatial-temporal aggregations from multi-view image feature sequences for iterative anchor refinement, where advanced queries contain both explicit geometric anchors and implicit semantic features.

Besides, Multi-Object Tracking (MOT) across multi-cameras is also required for downstream tasks. Traditional algorithms resort to "tracking-by-detection" paradigm, which relies on hand-crafted data association between tracked trajectories and newly perceived objects. Recent works seek to explore joint detection and tracking methods by introducing track queries to detect unique instances continuously and consistently. Sparse4Dv3 proposes an advanced 3D detector which takes full advantage of spatial-temporal information to propagate temporal instances for identity preserving, thus achieving superior end-to-end performance without additional tracking designs.

2.2. Online Mapping

Maps provide important static scenario information to ensure driving safety. Current works manage to construct online maps with on-board sensors instead of using HD-Map which is labor-intensive and expensive. HDMapNet achieves this aim through BEV semantic segmentation and heuristic post-processing to generate map instances. VectorMapNet introduces a two-stage auto-regressive transformer to refine map elements consecutively. MapTR regards map elements

as a set of points with equivalent permutations, while StreamMapNet adopts a temporal fusion strategy to enhance performance. However, all of them rely on dense BEV features for online map construction, which is computationally intensive and not extensible to the sparse manner.

2.3. End-to-End Motion Prediction

Motion prediction of surrounding agents in an end-to-end fashion can relieve accumulative error between standalone models. FaF predicts both current and future bounding boxes from images using a single convolutional network. IntentNet attempts to reason high-level behavior and long-term trajectories simultaneously. PnPNet aggregates trajectory-level features for motion prediction through an online tracking module. ViP3D takes images and HD-Map as input and adopts agent queries to conduct tracking and prediction. PIP further proposes to replace HD-Map with local vectorized map.

2.4. End-to-End Planning

End-to-end planning paradigm either unites modules of perception and prediction or adopts direct optimization on planning without intermediate tasks, which lack interpretability and are hard to optimize. Recently, UniAD presents a planning-oriented model which integrates various tasks in the dense BEV-Centric paradigm, achieving convincing performance. VAD learns vectorized scene representations and improves planning safety with explicit constraints. GraphAD constructs the interaction scene graph to model both dynamic and static relations. SparseDrive introduces the sparse perception module for parallel motion planner. However, using straightforward designs and exhaustive modeling without ego-centric interaction inevitably leads to unsatisfactory planning performance and inferior efficiency.

3. Methodology

3.1. Overview

The overall framework of the proposed DiFSD is illustrated in Fig. 2, which deals with the end-to-end planning task in an ego-centric fully sparse paradigm. Specifically, DiFSD mainly consists of four parts: visual encoder, sparse perception, hierarchical interaction, and iterative motion planner.

First, the visual encoder extracts multi-scale spatial features from given multi-view images. Then the sparse perception takes the encoded features as input to perform detection, tracking, and online mapping simultaneously. In the hierarchical interaction module, the ego query equipped with a geometric prior is introduced to pick out the interactive queries through ego-centric cross attention and hierarchical selection. In the iterative motion planner, both interactive agents and ego-vehicle are considered for joint motion prediction, then the predicted multi-modal ego-trajectories are further optimized iteratively. Mean-

while, both position-level diffusion of interactive agents and trajectory-level denoising of ego-vehicle are conducted for uncertainty modeling of motion and planning tasks respectively.

3.2. Problem Formulation

The given multi-view camera image sequence can be denoted as $S = \{I_t \in \mathbb{R}^{N\times 3\times H\times W}\}_{t=T-k}^{T}$, where N is the number of camera views and k indicates the temporal length until current timestep T respectively. Annotation of input S for end-to-end planning is composed by a set of future waypoints of the ego-vehicle $\psi = \{\phi = (x_t, y_t)\}_{t=1}^{T_p}$, where $T_p = 3s$ is the planning time horizon, and (x_t, y_t) is the BEV location transformed to the ego-vehicle coordinate system at current timestep T. Meanwhile, driving command as well as ego-status is also provided. Annotation set ψ is used during training. During prediction, the planned trajectory of ego-vehicle should fit the annotation ψ with minimum L2 errors and collision rate with surrounding agents.

3.3. Sparse Perception

After extracting multi-view visual features F from sensor images using the visual encoder, the sparse perception method proposed in previous works is extended to perform detection, tracking, and online mapping simultaneously based on a group of sparse queries, removing the dependence on dense BEV representations widely used in other methods.

Detection and Tracking. Following previous sparse perception methods, surrounding agents can be represented by a group of instance features $F_a \in \mathbb{R}^{N_a\times C}$ and anchor boxes $B_a \in \mathbb{R}^{N_a\times 11}$ respectively. Each anchor box is denoted as $\{x, y, z, \ln(w), \ln(h), \ln(l), \sin(\theta), \cos(\theta), v_x, v_y, v_z\}$, which contains location, dimension, yaw angle, and velocity. Taking the visual features F, instance features F_a, and anchor boxes B_a as input, N_{dec} decoders are adopted to consecutively refine the anchor boxes and update the instance features through deformable aggregation of sample features projected from key points of the anchor box B_a. The updated instance features are adopted to predict classification scores and box offsets respectively. Temporal instance denoising is introduced to improve model stability. As for tracking, following the ID assignment process in previous work, the temporal instances across frames used for advanced detection can also serve as track queries, which remain consistent with unique IDs.

Online Mapping. Similarly, we adopt an additional detection branch of the same structure for online mapping. Differently, the geometric anchor of each static map element is denoted as N_p points. Therefore, surrounding maps can be represented by a group of map instance features $F_m \in \mathbb{R}^{N_m\times C}$ and anchor polylines $B_m \in \mathbb{R}^{N_m\times N_p\times 2}$.

3.4. Ego-Env Hierarchical Interaction

After perceiving the dynamic and static elements existing in the driving scenario in a sparse manner, we continue to perform hierarchical interaction between the ego-vehicle and surrounding agent/map instances. As shown in Fig. 2, the hierarchical interaction module mainly consists of three parts: Ego-Env Dual Interaction, Intention-Guided Geometric Attention, and Coarse-to-Fine Selection.

Ego-Env Dual Interaction. As shown in Fig. 3, a learnable embedding $F_e \in \mathbb{R}^{1\times C}$ is randomly initialized to serve as ego query, along with an ego anchor box $B_e \in \mathbb{R}^{1\times 11}$ to represent the ego-vehicle. Both ego-centric cross attention with surrounding objects $F_o \in \mathbb{R}^{N_o\times C}$ and object-centric self attention are conducted consecutively to capture mutual information comprehensively. During the attention calculation process, we adopt the decouple attention mechanism proposed in previous work to combine positional embedding and query feature in a concatenated manner instead of an additive approach, which can effectively retain both semantic and geometric clues for interaction modeling.

Intention-Guided Geometric Attention. To enhance the accuracy and explainability of query ranking to facilitate selection, we introduce an ego-centric geometric prior additionally. As shown in Fig. 2, the intention-guided attention module is adopted to assess the importance of surrounding agent and map queries, which mainly consists of three steps: Response Map Learning, Reference Line Generation, and Interactive Score Fusion.

Specifically, we use four MLPs to encode the ego-intention respectively, including velocity, acceleration, angular velocity, and driving command. We then concatenate these embeddings to obtain ego-intention features $I_e \in \mathbb{R}^{1\times C}$, which are further concatenated with the position embeddings $F_p \in \mathbb{R}^{H\times W\times C}$ of a group of pre-defined locations $P \in \mathbb{R}^{H\times W\times 2}$ to cover densely distributed grid cells in the BEV plane. The position of each grid cell is represented as $p = (x, y)$. Finally, the concatenated geometric features are fed to a single SE block to learn response map $M_r \in \mathbb{R}^{H\times W\times 1}$, which is supervised by the normalized minimum distance from p to the ego future waypoints. The motivation is that the Closest In-Path Vehicle/Stationary are prone to affect the ego-intention, and vice versa.

With the predicted response map M_r, we first generate the reference line through row-wise thresholding, which is further used to generate the normalized distance map M_d (see Fig. 2). We can then obtain the geometric score S_{geo} for each surrounding query by referring to M_d. The reason we don't obtain the geometric score from M_r directly is that the imbalanced distribution of ego-intention and future waypoints may lead to inferior quality of M_r.

Finally, as shown in Fig. 4, we perform interactive score fusion through multiplying the attention, geometric, and classification scores during the ego-centric cross attention:

$$S_{inter} = \text{Softmax}\left(\frac{F_e \odot F_o^T}{\sqrt{d_k}}\right) \cdot S_{geo} \cdot S_{cls}$$

where the distance-prior is weighted with the attention score S_{attn} for both interaction and selection. $\odot$ denotes inner product, $\cdot$ is dot product, and d_k is the channel dimension.

Coarse-to-Fine Selection. To capture interaction information from coarse to fine, we stack M dual-interaction layers in a cascaded manner, where a top-K operation is appended between each two consecutive layers, thus interactive objects can be gradually selected for later prediction and planning usages. We claim that only a few interactive objects need to be considered for motion prediction, which are sufficient yet efficient for ego-centric path planning, instead of all detected agents existing in the driving scene.

3.5. Iterative Motion Planner

As shown in Fig. 2, the iterative motion planner is designed to conduct motion prediction for both interactive agents and ego-vehicle, and then optimize the proposal ego-trajectory with safety and kinematic constraints iteratively.

Joint Motion Prediction. With regard to trajectory prediction, both surrounding agents and ego-vehicle are adopted for motion modeling in a joint decoder, unlike previous works which neglect the crucial interactions between near agents and ego-vehicle when making motion predictions, especially in common scenarios like intersections. Another difference is that only the interactive objects F_{io} (CIPV) sparsely selected in the former module are considered, instead of all detected agents in the driving scene which may be irrelevant to ego-vehicle planning.

As for the joint motion decoder, we prepare three copies of ego query F_e to indicate different driving intentions (i.e., turn left, turn right, and keep forward), which are combined with F_{io} to conduct agent-level self attention and agent-map cross attention respectively. We then concatenate these output attended features to predict multi-modal trajectories $\tau_a \in \mathbb{R}^{N_a \times K_a \times T_a \times 2}$, $\tau_e \in \mathbb{R}^{N_e \times K_e \times T_e \times 2}$ and classification scores $S_a \in \mathbb{R}^{N_a \times K_a}$, $S_e \in \mathbb{R}^{N_e \times K_e}$ for both agents and ego-vehicle, where $N_e = 3$ is the number of driving commands for planning, $K_a = K_e = 6$ are the mode numbers, and $T_a = T_e = 6$ are the future timestamps.

Planning Optimization. With the predicted multi-intention and multi-modal trajectories of ego-vehicle, we can select the most probable proposal trajectory with the input driving command and classification score S_e. As shown in Fig. 3(b), ego-agent, ego-map, and ego-navigator cross attentions are conducted consecutively for planning optimization, where offsets for each future waypoint are predicted upon the proposal trajectory respectively with several planning constraints to ensure safety.

Iterative Refinement. To further promote the stability and performance of the whole end-to-end system, an additional iterative refinement strategy is proposed to continuously update the reference line and distance map M_d with refined ego trajectory as illustrated in Fig. 2, thus ensuring interaction quality and selection accuracy of interactive queries.

3.6. Uncertainty Denoising

Due to the planning-oriented modular design, output uncertainty from each individual module will inevitably be introduced and passed through to downstream tasks, leading to inferior and fragile system performance. Under this circumstance, we propose a two-level uncertainty modeling strategy to further stabilize the whole framework. On one hand, position-level diffusion process is performed on ground-truth boxes of interactive agents $B_i \in \mathbb{R}^{K\times 11}$ for additional trajectory prediction of noisy agents $B_n = B_i + \Delta B_{pos} \in \mathbb{R}^{G\times K\times 11}$ equipped with G groups of random noises following uniform distributions. ΔB_{pos} locates within two different ranges of $\{-s, s\}$ and $\{-2s, -s\} \cup \{s, 2s\}$ to indicate positives and negatives respectively, where s indicates the noise scale. This process aims to promote the stability of motion forecasting for interactive agents with uncertain detected positions, scales, and velocities. On the other hand, trajectory-level denoising process is also introduced for robust offset prediction of proposal trajectory of ego-vehicle in the planning optimization stage. Different from the position diffusion of detection or motion query described above, we apply random noise to trajectory offsets of ego-vehicle $\Delta B_{traj} \in \mathbb{R}^{G\times T_e\times 2}$, where s depends on the Final Displacement (FD) of ground-truth ego future trajectory.

3.7. End-to-End Learning

Multi-stage Training. To facilitate model convergence and training performance, we divide the training process into two stages. In stage-1, the sparse perception, hierarchical interaction, and joint motion prediction tasks are trained from scratch to learn sparse scene representation, interaction, and motion capability respectively. Note that no selection operation is adopted in stage-1, namely all detected agents are considered for motion forecasting to make full use of annotations. In stage-2, the geometric attention module and iterative planning optimizer are added to train jointly for overall optimization with uncertainty modeling.

Loss Functions. The overall optimization function mainly includes five tasks, where each task can be optimized with both classification and regression losses. The overall loss function for end-to-end training can be formulated as:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{map} + \mathcal{L}_{interact} + \sum_{i=1}^{N}(\mathcal{L}_{motion} + \mathcal{L}_{plan})$$

where $\mathcal{L}_{interact}$ is a combination of binary classification loss and L2 regression loss to learn geometric score, where positive (interactive) samples are denoted as grid cells with geometric score $S_{geo} \geq 0.9$ (within 3m for each future waypoint). An additional regression loss is included in $\mathcal{L}_{plan}$ for ego status prediction, instead of directly using it as input to the planner as in previous works, which would lead to information leakage as proven in recent studies. Meanwhile, vectorized planning constraints identified in previous work such as collision, overstepping, and direction are also included in $\mathcal{L}_{plan}$ for regularization. N is the number of motion planning stages.

4. Experiments

4.1. Datasets and Setup

Our experiments are conducted on the challenging public nuScenes dataset, which contains 1000 driving scenes lasting 20 seconds each. Over 1.4M 3D bounding boxes of 23 categories are provided in total, annotated at 2Hz. Following conventions, Collision Rate (%) and L2 Displacement Error (DE) (m) are adopted to measure open-loop planning performance. Additionally, to study the effect of various perception encoders, we evaluate 3D object detection and online mapping results using mAP and NDS metrics respectively.

4.2. Implementation Details

DiFSD plans a 3s future trajectory of ego-vehicle with 2s history information as input. Our DiFSD has two variants: DiFSD-B and DiFSD-S. For DiFSD-S, both sparse perception version DiFSD-S (Sparse) and dense BEV perception version DiFSD-S (Dense) are implemented for comparison. ResNet50 is adopted as the default backbone network for visual encoding. The perception range is set to 60m$\times$30$mlongitudinallyandlaterally.Inputimagesizeof DiFSD - Sisresizedto$640×360.$ForDiFSD-S(Dense), thedefaultnumberofBEVquery, mapquery, andagentqueryis1$0 and 300, respectively. For DiFSD-S (Sparse), N_{dec} is 6, N_a is 900, and N_m is 100 respectively. Each map element contains 20 map points. The feature dimension C is 256. The noise scale s is set to 2.0 and 0.2×FD for motion and planning respectively. G is set to 3. DiFSD-B has larger input image resolution (1280$×$720) and backbone network (ResNet101).

We use AdamW optimizer and Cosine Annealing scheduler to train DiFSD with weight decay 0.01 and initial learning rate 2$×10^{-4}$. DiFSD is trained for 48 epochs in stage-1 and 20 epochs in stage-2, running on 8 NVIDIA Tesla A100 GPUs with batch size 1 per GPU.

4.3. Main Results

As shown in Tab. 1, DiFSD shows great advantages in both performance and efficiency compared with previous works, including visual-based or multi-modality based methods. On one hand, DiFSD-S achieves the minimum L2 error even

with lightweight visual backbone and inferior dense perception encoder. Specifically, compared with BEVFormer-based end-to-end methods, DiFSD-S (Dense) reduces the average L2 error by a great margin (0.68m and 0.37m, separately), while significantly reducing the average collision rates by 77% and 68% respectively. Equipped with deeper visual backbone and advanced sparse perception, the average L2 error and collision rates can be further reduced to 0.32m and 0.06% respectively. Notably, we are the first to achieve 0% collision rate on 1s. On the other hand, benefiting from ego-centric hierarchical interaction, only sparse interactive agents (2%) are considered for motion planning. Hence, DiFSD-S can achieve great efficiency with 14.8 FPS, 8.2× and 3.3× faster than UniAD and VAD respectively.

4.4. Ablation Study

We conduct extensive experiments to study the effectiveness and necessity of each design choice proposed in DiFSD. We use DiFSD-S as the default model for ablation.

Effect of Sparse Perception. In addition to BEV-perception based end-to-end methods, recent end-to-end planning methods resort to the sparse perception fashion to provide advanced 3D detection and online mapping results with high efficiency. To study the significance of advanced perception encoders for ego-planning, we compare the perception performance of various end-to-end methods as shown in Tab. 2. With sparse perception encoder, the performance of 3D object detection and online mapping can be greatly improved (10.6 NDS and 7.5 mAP, respectively) compared with dense BEV-based perception paradigm. And the end-to-end planner equipped with the advanced perception encoder can consistently boost planning performance as shown in Tab. 1. Therefore, perception performance is essential for the end-to-end planner, which decides the planning upper-bound and provides rich clues of surrounding environment including both dynamic and static elements.

Necessity of Geometric Prior. We claim that only interactive agent and map queries are significant for ego-vehicle planning, where the Closest In-Path Vehicle as well as Stationary (CIPV/CIPS) are more likely to interact with the ego-vehicle. To verify the necessity of such geometric prior, we conduct exhaustive ablations of the ego-centric query selector as shown in Tab. 3. Without ego-centric selection, fewer objects randomly selected can result in worse planning results. While using ego-centric cross attention, only 2% of surrounding queries are enough for achieving convincing planning performance, instead of considering all existing dynamic/static elements. Besides, introducing the geometric prior through attention can dramatically reduce the L2 error and collision rate by 8% and 42% respectively. Meanwhile, when utilizing ground-truth geometric score for upper-limit evaluation, we can obtain extremely low average L2 error and collision rate (0.23m and 0.07% respectively). Undoubtedly, the proposed ego-centric selector equipped with geometric attention is non-trivial for efficient interaction and motion planning.

Effect of Designs in Hierarchical Interaction. Tab. 4 shows the effectiveness of our elaborate designs in the hierarchical interaction module, which contains three main components: Dual Interaction (DI), Geometric Attention (GA), and Coarse-to-Fine Selection (CFS). DI models both ego-centric and object-centric interactions respectively, which improves planning performance greatly as expected. GA facilitates the query selection process as discussed in Tab. 3, which reduces the collision rate by a great margin (42%). CFS contributes to interaction modeling quality through hierarchical receptive fields from global to local. All three designs combined together can achieve overall convincing planning performance.

Necessity and Order of Object Selection. Tab. 5 studies the necessity of agent and map selection during ego-centric hierarchical interaction. We observe that agent selection contributes more than map selection, especially for driving safety. Conducting both agent and map interactions in a cascaded order is inferior to the parallel manner, where the updated ego query from parallel outputs is concatenated for joint motion prediction.

Effect of Interactive Score Fusion. During ego-centric query selection, both geometric and classification scores are considered to ensure that selected closest in-path queries are true positive agents or maps for motion planning. Tab. 6 shows the effect of three types of scores used for query ranking: attention, geometry, and confidence scores. As in Eq. 1, interactive score S_{inter} obtained by multiplying these three scores achieves the best selection quality and planning performance. S_{inter} without confidence score fails to distinguish between background and foreground queries, resulting in inferior performance.

Effect of Designs in Motion Planner. As for the motion planner in DiFSD, Joint Motion Prediction (JMP), Planning Optimization (PO), and Iterative Refinement (IR) make up the planning pipeline of ego-vehicle. Additionally, Uncertainty Denoising (UD) contributes to system stability and training convergence. Tab. 7 explores the effect of each design exhaustively. ID-1 indicates evaluating the proposal trajectory of ego-vehicle predicted together with interactive agents, which achieves competitive L2 error but is more prone to collision with surrounding agents. ID-2 improves the collision rate greatly by 38.9% with the help of PO and planning constraints during training. ID-4 emphasizes the importance of IR in improving the quality of ego-planning trajectory (average 5.4% L2 error and 36.3% collision rate reduction respectively). ID-3 reflects the benefit of UD used for end-to-end training compared with ID-4.

Effect of Iterative Refinement Stages. We continue to study the number of refinement stages in Tab. 8. We observe that DiFSD can obtain superior planning performance with one additional refinement stage (36.3% collision rate reduction), which becomes saturated when introducing more stages. Hence, a two-stage interacted motion planner is sufficient for achieving convincing results.

Effect of Uncertainty Denoising. We also validate the effectiveness of uncertainty denoising strategy including position-level motion diffusion and

trajectory-level planning denoising. As shown in Tab. 9, motion diffusion can improve prediction stability with uncertain agent positions, while planning denoising can also strengthen trajectory regression precision of ego-vehicle.

Runtime of Each Module. We evaluate the runtime of each module of DiFSD-S as shown in Tab. 10. Visual backbone and sparse perception occupy most of the runtime (70.4%) for feature extraction and scene understanding. Hierarchical interaction also takes a significant portion (21.2%) for interaction modeling and interactive query selection. Thanks to sparse representation and ego-centric interaction, the motion planner only consumes 7.9ms to plan the future ego-trajectory (8.4% in total).

4.5. Qualitative Results

We visualize the motion trajectories of sparse interactive agents as well as planning results of DiFSD as illustrated in Fig. 5. Both surrounding camera images and prediction results on BEV are provided accordingly. We also project the planning trajectories to the front-view camera image. Only the top-3 trajectories of selected agents interacting with ego-vehicle are visualized for better understanding of DiFSD's motivation. DiFSD outputs planning results based on fully sparse representation in an end-to-end manner, not requiring any dense interaction or redundant motion modeling, let alone hand-crafted post-processing.

5. Conclusion and Future Work

Conclusion. In this paper, we propose a fully sparse paradigm for end-to-end self-driving in an ego-centric manner, termed DiFSD. DiFSD revisits human driving behavior and conducts hierarchical interaction based on sparse representation and perception results. Only interactive agents are considered for joint motion prediction with the ego-vehicle. Iterative planning optimization strategy contributes to driving safety with high efficiency. Additionally, uncertainty modeling is conducted to improve the stability of the end-to-end system. Extensive ablations and comparisons reveal the superiority and great potential of our ego-centric fully sparse paradigm for future research.

Future Work. There remain some limitations in our work. First, DiFSD performs ego-centric query selection based on geometric score, which is obtained through referring to the distance map using static query positions. How to generate better geometric score considering agent dynamics is worthy of future discussion. Additionally, how to incorporate additional traffic signals and vision-language models into the current driving system for autonomous decision making also deserves further exploration.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Bei-

jbom. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11621–11631, 2020.

[2] 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] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In 2018 IEEE international conference on robotics and automation (ICRA), pages 4693–4700. IEEE, 2018.

[4] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9329–9338, 2019.

[5] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5496–5506, 2023.

[6] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. IEEE Robotics and Automation Letters, 2024.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018.

[9] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. European Conference on Computer Vision, pages 533–549. Springer, 2022.

[10] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17853–17862, 2023.

[11] Bin Huang, Yangguang Li, Enze Xie, Feng Liang, Luya Wang, Mingzhu Shen, Fenggang Liu, Tianqi Wang, Ping Luo, and Jing Shao. Fast-bev: Towards real-time on-vehicle bird's-eye view perception. arXiv preprint arXiv:2301.07870, 2023.

[12] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054, 2022.

[13] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. arXiv preprint arXiv:2211.17111, 2022.

[14] 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.

[15] Bo Jiang, Shaoyu Chen, Xinggang Wang, Bencheng Liao, Tianheng Cheng, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, and Chang Huang. Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction. arXiv preprint arXiv:2212.02181, 2022.

[16] 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, pages 8340–8350, 2023.

[17] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In 2022 International Conference on Robotics and Automation (ICRA), pages 4628–4634. IEEE, 2022.

[18] Yinhao 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 Proceedings of the AAAI Conference on Artificial Intelligence, pages 1477–1485, 2023.

[19] 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 European conference on computer vision, pages 1–18. Springer, 2022.

[20] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14864–14873, 2024.

[21] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11553–11562, 2020.

[22] 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.

[23] Xuewu Lin, Zixiang Pei, Tianwei Lin, Lichao Huang, and Zhizhong Su.

Sparse4d v3: Advancing end-to-end 3d detection and tracking. arXiv preprint arXiv:2311.11722, 2023.

[24] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 18580–18590, 2023.

[25] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In European Conference on Computer Vision, pages 531–548. Springer, 2022.

[26] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3262–3272, 2023.

[27] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Vectormapnet: End-to-end vectorized hd map learning. In International Conference on Machine Learning, pages 22352–22369. PMLR, 2023.

[28] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In 2023 IEEE international conference on robotics and automation (ICRA), pages 2774–2781. IEEE, 2023.

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[31] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 3569–3577, 2018.

[32] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8844–8854, 2022.

[33] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 194–210. Springer, 2020.

[34] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7077–7087, 2021.

[35] Peize Sun, J Cao, Y Jiang, R Zhang, E Xie, Z Yuan, C Wang, and P Luo. Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460, 2020.

[36] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. arXiv preprint arXiv:2405.19620, 2024.

[37] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3621–3631, 2023.

[38] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In Conference on Robot Learning, pages 180–191. PMLR, 2022.

[39] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10359–10366. IEEE, 2020.

[40] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018.

[41] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17830–17839, 2023.

[42] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. arXiv preprint arXiv:2308.01006, 2023.

[43] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11784–11793, 2021.

[44] En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. arXiv preprint arXiv:2305.14298, 2023.

[45] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 7356–7365, 2024.

[46] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and

Yichen Wei. Motr: End-to-end multiple object tracking with transformer. In European Conference on Computer Vision, pages 659–675. Springer, 2022.

[47] Jiang-Tian Zhai, Ze Feng, Jinhao 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.

[48] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22056–22065, 2023.

[49] Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. arXiv preprint arXiv:2403.19098, 2024.

Supplementary Material

A. Evaluation Metrics

Perception. The evaluation for detection and tracking follows standard evaluation protocols. For detection, we use mean Average Precision (mAP), mean Average Error of Translation (mATE), Scale (mASE), Orientation (mAOE), Velocity (mAVE), Attribute (mAAE), and nuScenes Detection Score (NDS) to evaluate model performance. For online mapping, we calculate the Average Precision (AP) of three map classes: lane divider, pedestrian crossing, and road boundary, then average across all classes to get mean Average Precision (mAP).

Planning. We adopt commonly used L2 error and collision rate to evaluate planning performance. The evaluation of L2 error is aligned with VAD. For collision rate, there are two drawbacks in previous implementations, resulting in inaccurate evaluation of planning performance. On one hand, previous benchmarks convert obstacle bounding boxes into occupancy maps with a grid size of 0.5m, resulting in false collisions in certain cases, e.g., when the ego vehicle approaches obstacles smaller than a single occupancy map pixel. On the other hand, the heading of ego vehicle is not considered and assumed to remain unchanged. To accurately evaluate planning performance, we account for changes in ego heading by estimating the yaw angle through trajectory points, and assess the presence of collision by examining overlap between the bounding boxes of ego vehicle and obstacles. We reproduce planning results on our benchmark with official checkpoints for fair comparison.

B. Analysis & Discussion

The GroundTruth future state distribution of ego-vehicle on nuScenes validation set is illustrated in Fig. A1, calculated with fixed time interval (1s) between consecutive predicted waypoints. We also compare the output ego-state distribution of different popular end-to-end methods based on planned trajectories respectively, as shown in Fig. A2. We observe that without ego-centric design, the optimized end-to-end model is unable to handle various emergencies appearing in driving scenarios, where the absolute values of Δv and Δa are larger than normal situations. Under this circumstance, the output planned trajectories cannot conform to expert routes as expected. However, our DiFSD performs consistently better in planning future ego states with variable speed and acceleration, owing to ego-centric hierarchical interaction and selection mechanism, thus the iterative motion planner can focus on interactive agents rather than irrelevant objects.

C. Visualization

As shown in Fig. A3, A4, and A5, we provide additional visualization results to illustrate the generalizability of DiFSD on various driving scenarios under different commands. DiFSD outputs planning results based on hierarchical interaction and joint motion of sparse interactive agents without considering other irrelevant objects. We omit map selection results for clarity of road structure details.

Note: Figure translations are in progress. See original paper for figures.