Echo Planning for Autonomous Driving: From Current Observations to Future Trajectories and Back

Jintao Sun, Hu Zhang, Gangyi Ding, Zhedong Zheng

***Abstract*—Modern end-to-end autonomous driving systems suffer from a critical limitation: their planners lack mechanisms to enforce temporal consistency between predicted trajectories and evolving scene dynamics. This absence of self-supervision allows early prediction errors to compound catastrophically over time. We introduce Echo Planning (EchoP), a new self-correcting framework that establishes an end-to-end Current → Future → Current (CFC) cycle to harmonize trajectory prediction with scene coherence. Our key insight is that plausible future trajectories should be bi-directionally consistent, *i.e.*, not only generated from current observations but also capable of reconstructing them. The CFC mechanism first predicts future trajectories from the Bird's-Eye-View (BEV) scene representation, then inversely maps these trajectories back to estimate the current BEV state. By enforcing consistency between the original and reconstructed BEV representations through a cycle loss, the framework intrinsically penalizes physically implausible or misaligned trajectories. Experiments on nuScenes show that the proposed method yields competitive performance, reducing L2 error (Avg) by -0.04 m and collision rate by -0.12% compared to one-shot planners. Moreover, EchoP seamlessly extends to closed-loop evaluation, *i.e.*, Bench2Drive, attaining a 26.54% success rate. Notably, EchoP requires no additional supervision: the CFC cycle acts as an inductive bias that stabilizes long-horizon planning. Overall, EchoP offers a simple, deployable pathway to improve reliability in safety-critical autonomous driving.**

***Index Terms*—Autonomous Driving, Self-Correcting Planning, BEV, End-to-End.**

I. INTRODUCTION

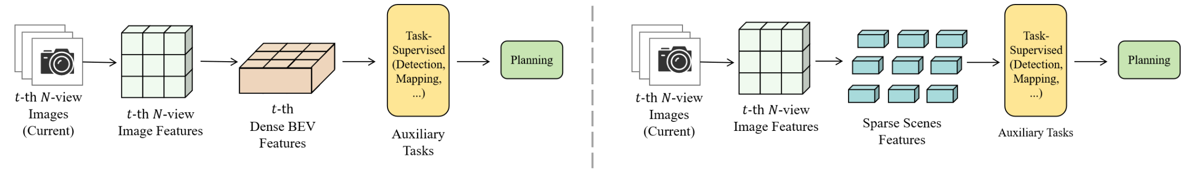
VISION-based end-to-end planning has become a leading paradigm in autonomous driving research [1]–[7]. Synchronized multi-view RGB images are usually mapped to Bird's-Eye-View (BEV) scene representation, providing a direct interface between perception and planning. Recent comprehensive surveys [8], [9] have highlighted the critical role of progressive BEV perception in safety-critical driving. BEV perception can be broadly categorized into dense and sparse paradigms based on whether they adopt a dense grid representation. For instance, dense BEV models rasterize the scene into grids and learn rich spatial features [1]–[4], [10]–[13]. Advanced methods have further extended this to precise 3D lane detection [14], bounding boxes [15], [16], and vectorized road network translation [17], providing explicit geometric constraints for downstream tasks. In contrast, sparse BEV models replace the grid with compact scene queries that reduce computation and label burden while preserving BEV geometry [5], [7], [18], [19]. Based on the BEV representation, most current approaches follow a one-shot planning paradigm, forecasting the vehicle's future path solely from a single snapshot of the present scene, shown in Fig. 1 (a, b, c). As summarized in Fig. 1(a–c), existing one-shot end-to-end planning paradigms can be grouped into three representative categories. Fig. 1(a) illustrates dense grid based BEV planning pipelines, where a rasterized BEV feature map is supervised by auxiliary tasks and then used for trajectory prediction, as exemplified by ST-P3 [1], UniAD [2], VAD [3], and GenAD [4]. Fig. 1(b) illustrates one-shot sparse query based planning that replaces dense grids with compact scene queries while still relying on auxiliary supervision, as in SparseDrive [5] and SparseAD [19]. Fig. 1(c) illustrates sparse planning without perception supervision that introduces temporal supervision via forward consistency, as represented by SSR [7]. In contrast, Fig. 1(d) summarizes EchoP with a **Current→Future→Current (CFC)** formulation, where the predicted future BEV is further used to reconstruct the current BEV state. This backward (echo) reconstruction provides an explicit reverse-consistency signal that complements the forward temporal supervision and makes the system functionality in Fig. 1(d) directly aligned with the two-loop design in Fig. 2. However, these methods lack an internal mechanism to enforce temporal consistency. For instance, while MomAD [20] has improved inter step trajectory stability by smoothing consecutive predictions, most one shot planners still lack an intrinsic self verification mechanism that checks whether the predicted future scene evolution remains consistent with the initial BEV observation. As a result, trajectory and scene self consistency verification through a bidirectional current future current check has remained underexplored. Without the temporal verification, early inaccuracies can accumulate over time and lead to unsafe driving behaviors. Hence, an open research question persists: ***how can we inherently embed self correcting scene trajectory consistency in trajectory planning without incurring additional annotation overhead or heuristic complexity?***

To address this challenge, we introduce **EchoP**, a novel self-correcting trajectory planning framework that enforces intrinsic temporal coherence through an end-to-end CFC feedback cycle (see Fig. 1 (d)). While prior work [7] models temporal variation by predicting a future BEV map from the current BEV and planned trajectory, they only enforce forward consistency and neglect the reverse constraint. In contrast, the core intuition of our EchoP is that a robust trajectory should be bidirectionally consistent: it should not only represent a plausible future evolution of the current scene but also be

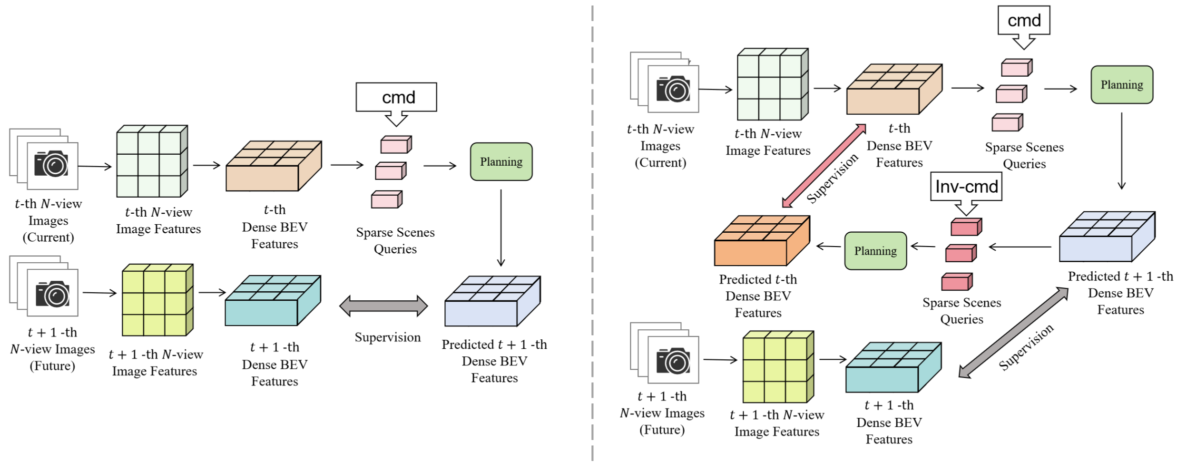
T. Sun and G. Ding are with the School of Computer Science and Technology, Beijing Institute of Technology, China 100081. E-mail: 3120215524@bit.edu.cn, dgy@bit.edu.cn

H. Zhang is with CSIRO DATA61, Australia. E-mail: hu1.zhang@csiro.au

Z. Zheng is with the Faculty of Science and Technology and Institute of Collaborative Innovation, University of Macau, Macau SAR, China 999078. E-mail: zhedongzheng@um.edu.mo


t-th N-view Images (Current)
t-th N-view Image Features
t-th Dense BEV Features
Task-Supervised (Detection, Mapping, ...)
Auxiliary Tasks
Planning
t-th N-view Images (Current)
t-th N-view Image Features
Sparse Scenes Features
Task-Supervised (Detection, Mapping, ...)
Auxiliary Tasks
Planning

(a) One-shot Dense Planning Paradigms (b) One-shot Sparse Planning Paradigms


t-th N-view Images (Current)
t-th N-view Image Features
t-th Dense BEV Features
cmd
Sparse Scenes Queries
Planning
$t+1$ -th N-view Images (Future)
$t+1$ -th N-view Image Features
$t+1$ -th Dense BEV Features
Supervision
Predicted $t+1$ -th Dense BEV Features
t-th N-view Images (Current)
t-th N-view Image Features
t-th Dense BEV Features
cmd
Sparse Scenes Queries
Planning
Supervision
Inv-cmd
Predicted t-th Dense BEV Features
Planning
Sparse Scenes Queries
Predicted $t+1$ -th Dense BEV Features
$t+1$ -th N-view Images (Future)
$t+1$ -th N-view Image Features
$t+1$ -th Dense BEV Features
Supervision

(c) One-shot Sparse Unsupervised Planning Paradigms (d) Ours Echo Planning Paradigms

Fig. 1. **Comparison between one-shot paradigms and our echo planning paradigms.** (a), (b), and (c) represent different types of one-shot approaches. Both (a) and (b) construct scene representations from image features and rely on auxiliary tasks to supervise the model, thereby enhancing environmental understanding; their main distinction lies in whether dense BEV features are used. Method (c) discards the auxiliary tasks introduced in (b) and is among the first to highlight the importance of temporal supervision, though it still considers only forward verification and thus remains within the one-shot paradigm. In contrast, ours (d) presents the echo planning approach, which employs a **Current** → **Future** → **Current** cycle. This design enforces **bidirectional self-supervision**, allowing the model to validate scene understanding without additional auxiliary tasks.

able to accurately reconstruct the current scene when projected backward. Inspired by the video generation [21], our model first generates future trajectories using a BEV representation of the current scene, and subsequently employs an inverse reconstruction step to predict back the current BEV state from these anticipated trajectories. This CFC cycle penalizes trajectories that are physically implausible or inconsistent with the initial scene configuration, thereby ensuring bidirectional consistency.

By enforcing the CFC feedback cycle, our method systematically reduces the impact of evolving scene dynamics on planning, preserves temporal consistency in the driving context, and consequently boosts trajectory accuracy. Experiments on the widely used nuScenes dataset [22] confirm that our approach achieves competitive performance, sharply lowering both L2 trajectory error and collision rates relative to existing one-shot planners. Crucially, these gains are obtained without any additional supervision or handcrafted heuristics, making the method both reliable and readily deployable. EchoP therefore provides a systematic method to intrinsically enforce trajectory–scene consistency, advancing end-to-end autonomous driving toward safer and more dependable operation. Our primary contributions are:

- **Self-Correcting Planning.** We propose a novel echo planning paradigm for autonomous driving, built upon a CFC feedback loop. Unlike one-shot prediction methods, we leverage predicted future trajectories to inversely reconstruct the current Bird's-Eye-View (BEV) scene state. This reconstruction process enables implicit self-correction by penalizing implausible trajectories without requiring any external supervision.
- **High-Fidelity Driving Performance.** Extensive experiments on open-loop driving benchmarks, specifically nuScenes, validate that EchoP arrives at a competitive average collision rate of 0.17 without additional inference overhead. Moreover, EchoP can be seamlessly extended to closed-loop trajectory prediction, achieving a 26.54 success rate on the Bench2Drive benchmark.

Furthermore, distinct from conventional planners that solely prioritize trajectory regression, EchoP introduces a critical spatio-temporal consistency regularizer into the learning process. By enforcing that predicted future trajectories should be capable of inversely reconstructing current observations, the proposed CFC mechanism serves as an intrinsic self-supervisory signal. This bidirectional validation significantly enhances the system's capability to learn robust visual representations, reducing physically implausible predictions arising from sensor noise or ambiguous scene dynamics, effectively acting as a logic coherence check within the planning loop.

Novelty, significance, and practical implications. While one-shot end-to-end planning has been extensively studied, existing temporal-supervision designs are mostly feed-forward and lack an explicit mechanism to verify whether the predicted future remains consistent with the current observation. EchoP provides a distinct insight by introducing *bidirectional* trajectory–scene consistency via a Current→Future→Current (CFC) feedback constraint, which penalizes plans that cannot coherently reconstruct the present state. This design improves practical adaptability because the CFC term serves as a *training-time* self-supervision signal and does not change the test-time planning pathway. Our findings suggest a lightweight principle for stakeholders and the research community: using bidirectional consistency as an intrinsic regularizer to enhance planning robustness, and using reconstruction-based consistency as a diagnostic signal for spatio-temporal reasoning in complex traffic.

The remainder of this paper is organized as follows. Section II reviews the related literature regarding end-to-end autonomous driving and BEV representation learning. Section III elaborates on the proposed Echo Planning framework, providing a detailed formulation of the self-correcting CFC mechanism. Section IV presents extensive quantitative and qualitative experiments on standard benchmarks, along with comprehensive ablation studies to validate the effectiveness of our approach. Finally, Section V concludes the paper and outlines potential directions for future research.

II. RELATED WORK

End-to-end planning. Planning is the ultimate objective of the first phase of end-to-end autonomous driving. Early work relied on relatively simple neural networks that ignored much of the scene context and therefore offered limited interpretability [23]–[29]. With the advent of large-scale datasets and stronger BEV perception, many learning-based planners have been proposed [1]–[6], [13], [19], [30]–[33]. Most recent systems still follow a perception, prediction, and planning pipeline to maintain transparency. For instance, ST-P3 [1] chains map perception, BEV occupancy prediction, and trajectory planning to infer future ego motion from surround cameras. UniAD [2] introduces a unified query design that integrates detection, mapping, and motion forecasting. VAD [3] uses a vectorized scene representation to couple scene understanding with planning constraints, and GenAD [4] generates future trajectories for both the ego vehicle and other agents within a learned probabilistic latent space. A newer line of research skips the explicit perception and prediction stages to facilitate efficiency. BEV-Planner [33] employs ego queries to extract task-relevant cues directly from BEV features, while SSR [7] leverages navigation commands to focus on salient regions without perception supervision. We also note several very recent end-to-end planning directions that complement the above paradigms, including constraint-guided flow-based planning (GuideFlow [34]), world-model and vision-language assisted planning (MindDrive [35]), and probabilistic planning to model uncertainty (VADv2 [36]). Despite the progress, many end-to-end planners remain predominantly feed-forward, and fewer works explicitly incorporate an intrinsic self-verification signal that checks whether predicted futures stay semantically consistent with the current scene representation, beyond trajectory-sequence regularization. We address this gap with an end-to-end CFC cycle that enforces coherent scene evolution over time.

BEV perception in autonomous driving. Perception forms the bedrock of autonomous driving, as we need distill actionable information from raw sensor streams. Precise and efficient scene understanding is therefore essential. The steady evolution of BEV representations [37]–[46] has propelled perception forward, while the lower cost of RGB cameras is gradually supplanting LiDAR-based pipelines [11], [47]–[50]. LSS [51] is a landmark effort that used depth prediction to lift perspective features into BEV space. To enhance feature robustness, depth-induced multi-scale attention mechanisms [52] have been explored to better capture saliency in complex depth ranges. BEVFormer [53] introduces spatial and temporal attention within a Transformer, achieving strong camera-only detection. Furthermore, DeepInteraction [54] demonstrates that predictive interaction between modalities effectively decouples complex scene dynamics, achieving strong detection performance. More broadly, temporal modeling and spatio-temporal attention are widely adopted to strengthen BEV representations under dynamic scenes, echoing insights from spatial-temporal attention in video understanding [55] and reasoning-oriented BEV perception [43]. To mitigate the high computational cost of dense BEV feature, sparse-perception methods have begun to emerge [5], [7], [18], [19]. SparseDrive [5] presents a symmetric sparse-perception module that jointly learns detection, tracking, and online mapping to produce a fully sparse scene representation. SparseAD [19] uses a compact query set to encode the driving scene sparsely. Our work follows this sparse-representation paradigm with zero annotation overhead.

III. METHOD

Given raw sensor inputs (*i.e.*, surrounding N-view camera images $I^i, i = 1, ..., N$) and high-level navigation command C_{navi} extracted from the dataset, our Echo planning model plans the future trajectory T of the ego Vehicle (see Fig. 2). EchoP first encodes the image features with a backbone network and then uses bev queries to convert image features into BEV feature space (see Sec. III-A). Second, EchoP distills task-critical cues from the ego-centric BEV space via a sparse scene representation module [7] and condition them on the high-level navigation command, thereby mimicking the selective attention that human drivers pay to salient scene elements (Sec. III-A). Finally, at the core of our approach, EchoP introduces a CFC cycle: it first predicts the ego-vehicle's future trajectory from the navigation-conditioned scene features, then inversely infers the present ego state from this forecast, and finally reconstructs the current BEV feature map (Sec. III-B). Unlike prior end-to-end planners that propagate information only in the forward temporal direction, this bidirectional reasoning mechanism leverages temporal consistency within the ego sequence, leveraging the sequential context and yielding more reliable planning.

A. Preliminary

BEV features generation. At the t-th time step, the surrounding N-view camera images $I^i, (i = 1, ..., N)$, are transformed into a BEV representation using BEVFormer [53]. Specifically, images I^i pass through an image backbone to yield image features $F_I^i, (i = 1, ..., N)$. Retaining the BEV representation B_{t-1} from the preceding timestamp $t - 1$, a group of learnable BEV queries $Q \in \mathbb{R}^{H \times W \times K}$ simultaneously extracts temporal context from the previous BEV features B_{t-1} and spatial information from the current multi-camera features $\left[F_I^i\right]_t$ via cross-attention. This results in the

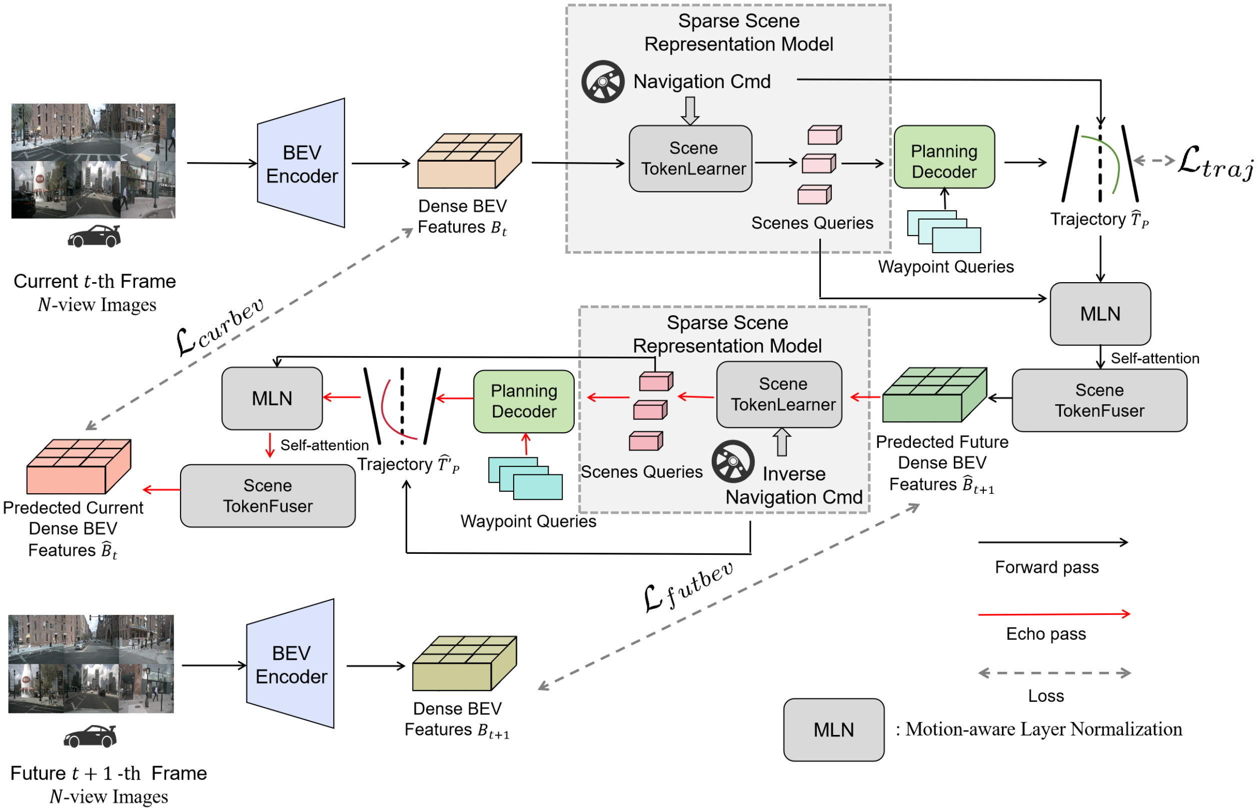

Sparse Scene Representation Model
Navigation Cmd
Scene TokenLearner
Scenes Queries
Planning Decoder
Waypoint Queries
Trajectory $\widehat{T}_P$
$\mathcal{L}_{traj}$
BEV Encoder
Dense BEV Features B_t
Current t-th Frame
N-view Images
$\mathcal{L}_{curbev}$
MLN
Self-attention
Scene TokenFuser
Predected Future Dense BEV Features $\widehat{B}_{t+1}$
Sparse Scene Representation Model
Scene TokenLearner
Scenes Queries
Inverse Navigation Cmd
Planning Decoder
Waypoint Queries
Trajectory $\widehat{T}'_P$
MLN
Self-attention
Scene TokenFuser
Predected Current Dense BEV Features $\widehat{B}_t$
$\mathcal{L}_{futbev}$
BEV Encoder
Dense BEV Features B_{t+1}
Future $t+1$-th Frame
N-view Images
Forward pass
Echo pass
Loss
MLN
: Motion-aware Layer Normalization

Fig. 2. **The overview of our EchoP framework.** EchoP is trained through two complementary loops. The forward loop, indicated by black arrows in the figure, predicts future BEV features through the sparse scene representation model [7] and applies self-supervision against the ground-truth BEV of future frames. The echo loop, shown in red, takes those predicted future features via the Motion-aware Layer Normalization (MLN) module [7], [56] and TokenFuser [7], [57], reconstructs the current BEV, and self-supervises against the ground-truth BEV of the current frame. By validating perception in both directions along the CFC cycle, the model cross-checks its understanding of the surrounding scene and thus produces reliable trajectory plans. **Notably, modules like Scene TokenLearner, Scene TokenFuser, Planning Decoder, and MLN, share weights during training. The CFC cycle does not impact inference speed, as only a forward pass from Current → Future is required during testing, consistent with existing methods.**

updated BEV representation $B_t \in \mathbb{R}^{H\times W\times K}$, where H, W define the spatial resolution of the BEV plane, and K denotes the feature channel dimensionality.

Sparse scene representation module. Given the BEV features and navigation command, the sparse scene representation module, such as SSR [7], adopts the scene token learner to extract scene queries $S \in \mathbb{R}^{N_s\times K}$ from dense BEV features B, where N_s is the number of scene queries. Firstly, the navigation command C_{navi} is encoded into the dense BEV feature [58]. Then, the navigation-aware BEV feature goes through the BEV TokenLearner module [57] and multi-layer self-attention [59] to get the current frame sparse scene representation S_t. After obtaining the current-frame scene representation S_t that encodes the salient BEV context, we introduce a set of waypoint queries $W_t \in \mathbb{R}^{N_t\times N_c\times K}$. These queries attend to S_t to capture the prospective motion of vehicle, where N_t is the prediction horizon (in time steps), N_c is the number of discrete navigation commands, and K is the feature dimensionality. A lightweight multi-layer perceptron subsequently regresses the waypoint embeddings into planar coordinates, producing the predicted ego trajectory $T_P \in \mathbb{R}^{N_t\times 2}$ over the next N_t time steps. For supervision, the high-level navigation command C_{navi} selects the mode-consistent branch $\hat{T}_P$, which is compared with the ground-truth trajectory T_{GT} using an L1 loss:

$$\mathcal{L}_{traj} = \left\| \hat{T}_P - T_{GT} \right\|_1 . \tag{1}$$

B. Echo planning

Motivation. In end-to-end autonomous driving planning, real-time scene context should strictly inform the prediction of the vehicle's trajectory. However, relying solely on forward prediction is often an ill-posed problem, as the model may generate trajectories that are statistically probable but semantically inconsistent with the current fine-grained scene details. Such feed-forward approaches are prone to hallucinations when facing complex occlusions or open-set sensor noise, lacking a mechanism to verify the causality of the generated plan. To address this, we argue that a robust planner should enforce bi-directional consistency: a plausible trajectory should not only be inferred from the current observation but should also contain sufficient information to inversely reconstruct that observation. This reversibility serves as a hard geometric and semantic constraint to filter out inconsistent predictions that deviate from the actual scene evolution. Motivated by this need for intrinsic spatio-temporal verification, we propose EchoP, a cycle-loop self-correcting framework that leverages self-supervised training across time to substantially enhance planning fidelity without requiring external supervision. As illustrated in Fig. 2, we introduce a recurrent planner that forms a CFC BEV cycle loop, continuously validating trajectory predictions by inversely reconstructing the present BEV state. Fig. 2 visualizes the full training graph of EchoP, where the *forward loop* (black arrows) and the *echo loop* (red arrows) share the same core modules. From current N-view images, a BEV encoder produces dense BEV features B_t, which are compressed into sparse scene tokens S_t by the Scene TokenLearner. Conditioned on the navigation command, the planning decoder regresses the ego trajectory and is supervised by the trajectory loss in Eq. 1. In the forward loop, the predicted future scene tokens are fused back to dense BEV $\hat{B}_{t+1}$ (TokenFuser) and are self-supervised by the future-BEV loss in Eq. 2. In the echo loop, $\hat{B}_{t+1}$ is processed again to reconstruct the current BEV $\hat{B}_t$ and is self-supervised by the current-BEV loss in Eq. 3. The dashed lines in Fig. 2 indicate these three supervision signals.

Forward pass (Current → Future). The navigation-consistent trajectory prediction $\hat{T}_P$ and the current scene embedding S_t are first processed by a Motion-aware Layer Normalization (MLN) module [7], [56], producing preliminary prediction queries. A self-attention block [59] then refines these queries into the future scene representation $\hat{S}_{t+1}$. Because the regions of interest in autonomous-driving scenes shift over time, we provide an additional source of supervision. The predicted future scene embedding $\hat{S}_{t+1}$ is fed to TokenFuser [7], [57] to hallucinate a dense BEV feature map $\hat{B}_{t+1}$. We then enforce an L2 penalty between $\hat{B}_{t+1}$ and B_t the ground-truth BEV, yielding the future-BEV reconstruction loss $\mathcal{L}_{futbev}$:

$$\mathcal{L}_{futbev} = \left\| \hat{B}_{t+1} - B_{t+1} \right\|_2 . \tag{2}$$

Echo pass (Future → Current). Our goal in this stage is to start from the predicted dense BEV at $T+1$, $\hat{B}_{t+1}$, and run the pipeline in reverse to reconstruct the current BEV $\hat{B}_t$ for self-supervision. The Echo pass comprises four steps. Firstly, to mimic backward driving, we derive a reversed navigation command C'_{navi} from the original C_{navi} and feed $\left[\hat{B}_{t+1}, C'_{navi}\right]$ into the same sparse-scene-representation module. The reversed navigation command C'_{navi} is derived by inverting the original discrete command C_{navi}: specifically, the "left" and "right" directions are swapped, while "go straight" remains unchanged. Secondly, the scene token learner [7], equipped with an attention module, extracts predicted future scene queries $\hat{S}'_{t+1}$ from the input BEV features $\hat{B}_{t+1}$. Thirdly, a new set of waypoint queries W'_{t+1} attends to $\hat{S}'_{t+1}$, generating a reversed trajectory proposal $\hat{T}'_p$, after which the inverse command C'_{navi} selects the mode-consistent branch of $\hat{T}'_p$. Finally, analogous to the forward pass, $\hat{T}'_p$ and $\hat{S}'_{t+1}$ are processed by an MLN block followed by stacked self-attention layers, producing the current-frame scene queries $\hat{S}'_t$. TokenFuser [7], [57] then reconstructs the dense BEV map $\hat{B}_t$ from $\hat{S}'_t$. The discrepancy between the reconstructed and ground-truth BEV features derives the current BEV reconstruction loss as follow:

$$\mathcal{L}_{curbev} = \left\| \hat{B}_t - B_t \right\|_2 . \tag{3}$$

Total loss. In summary, EchoP is trained end-to-end with a composite objective comprising the trajectory loss $\mathcal{L}_{traj}$, the future-BEV reconstruction loss $\mathcal{L}_{futbev}$, and the current-BEV reconstruction loss $\mathcal{L}_{curbev}$, combined as: $\mathcal{L}_{total} = \mathcal{L}_{traj} + \lambda_{futbev}\mathcal{L}_{futbev} + \lambda_{curbev}\mathcal{L}_{curbev}$.

C. Method Roadmap and Practical Interpretation

To make the construction of EchoP easier to follow, we summarize the method as a two-pass training procedure over a **Current→Future→Current (CFC)** cycle. **Forward pass (Current→Future).** Given the current BEV feature B_t and the navigation command C_{navi}, EchoP predicts a navigation-consistent trajectory $\hat{T}_P$ and a future scene representation $\hat{S}_{t+1}$, which is further decoded into a dense future BEV $\hat{B}_{t+1}$. This stage is supervised by the future-BEV reconstruction loss L_{futbev} (Eq. 2). **Echo pass (Future→Current).** Starting from the predicted future BEV $\hat{B}_{t+1}$, EchoP runs a mirrored pipeline (with shared weights) to reconstruct the current BEV $\hat{B}_t$, supervised by the current-BEV reconstruction loss L_{curbev} (Eq. 3). This cycle provides an intrinsic spatio-temporal consistency check: a plausible plan should be reversible enough to reconstruct the observation it is conditioned on. **Deployment note.** The echo pass is used *only during training*. At inference time, EchoP performs a single forward pass (Current→Future) to output $\hat{T}_P$, incurring no additional runtime overhead compared with one-shot planners.

IV. EXPERIMENT

In this section, we provide a comprehensive evaluation of the proposed EchoP framework, organized into three main parts: experimental setup, comparative analysis, and in-depth discussion. We begin by introducing the standard open-loop benchmark nuScenes [22] and the closed-loop benchmark Bench2Drive [31], followed by the definitions of key evaluation metrics (e.g., L2 error, collision rate, and driving score) and the implementation specifics. The extensive comparison with state-of-the-art end-to-end planners is presented in Section IV-A. Finally, Sections IV-B and IV-C discuss the ablation studies validating the spatio-temporal consistency of the CFC cycle and provide qualitative visualizations of planning behaviors, respectively.

Dataset. We conduct extensive experiments on the widely adopted nuScenes dataset [22] to assess the open-loop planning capability of our EchoP framework. nuScenes comprises 1,000 driving logs, partitioned into 700, 150, and 150 scenes for training, validation, and testing, respectively. Each scene provides 20 seconds (s) of synchronized RGB and LiDAR data captured at 12 Hz, supplemented by key-frame annotations at 2 Hz. The sensor suite includes six surround-view cameras covering a full 360° field of view and a 32-beam LiDAR. The dataset supplies dense semantic maps alongside 1.4M 3D bounding boxes spanning 23 object classes, providing a rich testbed for perception and prediction. **Real-world factors in nuScenes.** We emphasize that our main open-loop benchmark is *not* confined to simulation: nuScenes is collected from real driving logs with multi-view RGB streams, and contains diverse real-world conditions (e.g., rainy scenes and low-light nighttime driving). Therefore, our evaluation already covers practical visual challenges beyond idealized settings, as also illustrated by the qualitative cases in Fig. 3(b, d).

To rigorously evaluate the planning robustness in an interactive environment, we further employ Bench2Drive [31], a state-of-the-art closed-loop benchmark built upon the CARLA simulator. Unlike open-loop evaluation, Bench2Drive requires

TABLE I
Equation-to-module mapping and how each term relates to evaluation metrics.

Component	Output	Supervision	Main effect on metrics
Trajectory regression L_{traj} (Eq. (1))	$\hat{T}_P$	T_{GT}	Reduces L2 displacement error
Future BEV reconstruction L_{futbev} (Eq. (2))	$\hat{B}_{t+1}$	B_{t+1}	Improves future scene grounding
Current BEV reconstruction L_{curbev} (Eq. (3))	$\hat{B}_t$	B_t	Enforces temporal consistency; reduces collisions

TABLE II
Open-Loop Planning on nuScenes. The **top block** with ^{o} means the LiDAR-based methods. The **middle block** follows the UniAD protocol (final/max aggregation). The **lower block** with ‡ follows VAD protocol (average over all predicted frames). *: The backbone is ResNet-101 [60], while other methods without * adopt ResNet-50 or similar variants. §: The result that we re-implement SSR with official weights [7]. The ↓ indicates that lower is better.

Method	Auxiliary Task	**L2** (m) ↓				**Collision Rate** (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
NMP^{o} [61]	Det & Motion	0.53	1.25	2.67	1.48	0.04	0.12	0.87	0.34
FF^{o} [62]	FreeSpace	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO^{o} [63]	FreeSpace	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST–P3 [1]	Det & Map & Depth	1.72	3.26	4.86	3.28	0.44	1.08	3.01	1.51
UniAD* [2]	Det&Track&Map&Motion&Occ	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
OccNeXt* [64]	Det & Map & Occ	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
VAD–Base [3]	Det & Map & Motion	0.54	1.15	1.98	1.22	0.04	0.39	1.17	0.53
PARA–Drive [6]	Det&Track&Map&Motion&Occ	0.40	0.77	1.31	0.83	0.07	0.25	0.60	0.30
GenAD [4]	Det & Map & Motion	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
SparseDrive [5]	Det&Track&Map&Motion	0.44	0.92	1.69	1.01	0.07	0.19	0.71	0.32
UAD–Tiny [65]	Det	0.47	0.99	1.71	1.06	0.08	0.39	0.90	0.46
UAD* [65]	Det	0.39	0.81	1.50	0.90	**0.01**	0.12	0.43	0.19
SSR§ [7]	None	0.25	0.64	1.33	0.74	0.16	0.21	0.51	0.29
MomAD [20]	None	0.43	0.88	1.62	0.98	0.06	0.16	0.68	0.30
EchoP (Ours)	None	**0.23**	**0.60**	**1.27**	**0.70** (-0.04)	0.02	**0.12**	**0.39**	**0.17** (-0.12)
ST–P3‡ [1]	Det & Map & Depth	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD*‡ [2]	Det&Track&Map&Motion&Occ	0.44	0.67	0.96	0.69	0.04	0.08	0.23	0.12
VAD–Tiny‡ [3]	Det & Map & Motion	0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38
VAD–Base‡ [3]	Det & Map & Motion	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
HE–Drive‡ [66]	Det&Track&Map	0.31	0.58	0.93	0.60	0.01	0.05	0.16	0.07
BEV–Planner‡ [33]	None	0.28	0.42	0.68	0.46	0.04	0.37	1.07	0.49
PARA–Drive‡ [6]	Det&Track&Map&Motion&Occ	0.25	0.46	0.74	0.48	0.14	0.23	0.39	0.25
LAW‡ [67]	None	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
GenAD‡ [4]	Det & Map & Motion	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19
SparseDrive‡ [5]	Det&Track&Map&Motion	0.29	0.58	0.96	0.61	0.01	0.05	0.18	0.08
UAD*‡ [65]	Det	0.28	0.41	0.65	0.45	0.01	**0.03**	0.14	**0.06**
DiffDrive ‡ [68]	Det &Map	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
SSR§‡ [7]	None	0.19	0.36	0.62	0.39	0.10	0.13	0.22	0.15
MomAD ‡ [20]	None	0.31	0.57	0.91	0.60	0.01	0.05	0.22	0.09
BridgeAD ‡ [69]	None	0.28	0.55	0.92	0.58	0.00	0.04	0.20	0.08
FocalAD ‡ [70]	Det & Map & Motion	0.27	0.57	0.96	0.60	**0.00**	0.04	0.24	0.09
EchoP‡ (Ours)	None	**0.17**	**0.33**	**0.58**	**0.36** (-0.03)	0.03	0.05	**0.14**	0.07 (-0.08)

the agent to actively control the vehicle across 220 official routes, navigating through dynamic traffic and complex topologies under diverse weather conditions. This benchmark features a multi-ability evaluation protocol that dissects driving performance into distinct competencies, such as lane changing, intersection traversal, and yielding, thereby providing a granular and comprehensive assessment of the planner's generalization capability and safety in long-horizon execution. We stress that Bench2Drive complements open-loop evaluation by enforcing *closed-loop* execution, where the agent must repeatedly re-plan under dynamic traffic and long-horizon interactions, thereby partially bridging the gap between offline metrics and real-world operational constraints.

Evaluation metrics. Following prior work, we evaluate open-loop performance on nuScenes using two primary metrics: L2 displacement error (the Euclidean distance between the planned and ground-truth trajectories) and Collision Rate (the proportion of planned trajectories that collide with any traffic participant). Unless specified otherwise, the model is fed with 2 seconds of history corresponding to five frames, and planning quality is reported for 1s, 2s, and 3s horizons. Because VAD [3] and UniAD [2] adopt different aggregation rules, we report results under both protocols: VAD averages over all predicted frames, whereas UniAD uses the final value (maximum collision). Table II presents the outcomes under both protocols.

TABLE III
OPEN-LOOP AND CLOSED-LOOP PLANNING ON BENCH2DRIVE. THE ↑ INDICATES THAT HIGHER IS BETTER. THE ↓ INDICATES THAT LOWER IS BETTER. §: THE RESULT THAT WE RETRAIN THE SSR [7].

Method	Open-loop	Closed-loop		
	Avg. L2 ↓	Driving score ↑	Success Rate (%) ↑	Efficiency ↑
AD-MLP [72]	3.64	18.05	0.00	48.45
UniAD-Tiny [2]	0.80	40.73	13.18	123.92
UniAD-Base [2]	**0.73**	45.81	16.36	129.21
VAD [3]	0.91	42.35	15.00	157.94
GenAD [4]	-	44.81	15.90	-
SSR§ [7]	0.90	32.34	11.85	157.62
MomAD [20]	0.85	45.35	17.44	162.09
FocalAD [70]	0.85	45.77	17.30	174.01
DIVER [73]	1.05	49.21	21.56	177.00
EchoP (ours)	0.79	**50.35**	**26.54**	**185.18**

For Bench2Drive closed-loop evaluation, we report: Success Rate (SR), the percentage of failure-free route completions; Driving Score (DS), the primary metric weighting completion by infraction penalties; and Efficiency, which evaluates trajectory smoothness and travel time relative to the standard.

Implementation details. EchoP is trained for 12 epochs on 4 RTX-A6000 GPUs with a batch size of 1 per GPU. The image encoder is a ResNet-50 [60] fed with 640 × 360 images, yielding a 100 × 100 BEV grid and 16 × 256 sparse tokens. The navigation-command set remains 3, matching prior work. For training, we deploy AdamW [71] with a learning rate of 5×10^{-5}, and all other hyperparameters follow SSR [7]. Unless stated, loss weights for future-BEV, and current-BEV are $\lambda_{futbev} = 0.5$, and $\lambda_{curbev} = 0.1$ by default. **Runtime and throughput.** We report end-to-end inference speed on an NVIDIA A100-PCIE-40GB using the standard speed-test protocol (world size = 1, batch size = 1). Our implementation achieves 46.27 ms/iter, corresponding to 21.61 FPS. The overhead is limited because the CFC branch reuses the BEV latent features and shares backbone computation, while the inverse reconstruction is implemented as a lightweight head on compact BEV representations rather than an additional full perception stack.

Scalability and feasibility in large SDV networks. EchoP is designed as an on-board planner that consumes only the ego vehicle's local multi-view observations and a high-level navigation command, and outputs short-horizon waypoints. From a systems perspective, this implies favorable scalability in large dynamic SDV fleets: the computational cost is *per-vehicle* and does not require centralized coordination for inference. Moreover, our planning head operates on compact scene tokens (e.g., 16×256 by default) distilled from the BEV representation, which helps keep the planning computation lightweight. As a result, EchoP can operate in real time under the measured single-GPU throughput, while retaining the benefit of the CFC self-supervision during training.

A. Comparison with the State-of-the-art Methods

Open-Loop Evaluation. To assess the effectiveness of EchoP, we train and evaluate the model on the widely used nuScenes benchmark. As summarised in Table II, EchoP achieves the best results among contemporary end-to-end planners on both the L2 error and the collision rate, attaining an average $L2_{MAX}$ error of only 0.70 m. Compared to strong one-shot planners, our approach delivers consistent benefits. Against UniAD the average $L2_{MAX}$ error drops by 0.33 m (32% relatively) and the average CR_{MAX} falls by 0.14% (45% relatively). Compared with VAD-Base, the gains are even larger, with reductions of 0.52 m in average $L2_{MAX}$ (43% relatively) and 0.36% in average CR_{MAX} (68% relatively). In addition, we compare our approach to the SSR baseline. Due to recent updates in the SSR codebase, our reproduced results on a 4-GPU setup differ from those reported in the original paper, a point also acknowledged by the authors on GitHub, as their modifications affected collision rate performance. For a fair comparison, we benchmark EchoP against our reproduced SSR results and observe substantial improvements in both L2 error and collision rate. EchoP still lowers the average $L2_{MAX}$ by 0.04 m and the average CR_{MAX} by 0.12%, while also improving the $L2_{AVG}$ and CR_{AVG} metrics by 0.03 m and 0.08% respectively. The consistent improvements across all indicators confirm that the Current → Future → Current cycle provides a reliable mechanism for accurate and safe trajectory planning.

TABLE IV
ABLATION FOR OUR CFC CYCLE. THE UPPER BLOCK FOLLOWS THE UNIAD PROTOCOL (FINAL/MAX AGGREGATION). THE LOWER BLOCK WITH ‡ FOLLOWS VAD PROTOCOL (AVERAGE OVER ALL PREDICTED FRAMES).§: THE RESULT THAT WE RETRAIN THE SSR [7] ON 4 GPUS. THE ↓ INDICATES THAT LOWER IS BETTER.

Method	CFC cycle	**L2** (m) ↓				**Collision Rate** (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
SSR§ [7]		0.25	0.64	1.33	0.74	0.16	0.21	0.51	0.29
EchoP (Ours)	✓	**0.23**	**0.60**	**1.27**	**0.70**	**0.02**	**0.12**	**0.39**	**0.17**
SSR§‡ [7]		0.19	0.36	0.62	0.39	0.10	0.13	0.22	0.15
EchoP‡ (Ours)	✓	**0.17**	**0.33**	**0.58**	**0.36**	**0.03**	**0.05**	**0.14**	**0.07**

TABLE V
ABLATION FOR THE CFC CYCLE EFFECTIVENESS IN DIFFERENT SCENE REPRESENTATION NUMBERS. THE BLOCK ALL FOLLOWS THE UNIAD PROTOCOL (FINAL/MAX AGGREGATION). THE ROW WITH A GRAY BACKGROUND IS THE RE-IMPLEMENTATION OF OUR BASELINE. THE ↓ INDICATES THAT LOWER IS BETTER.

N_s	CFC cycle	**L2** (m) ↓				**Collision Rate** (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
8		0.27	0.69	1.43	0.79	0.20	0.35	0.68	0.41
8	✓	**0.21**	**0.57**	**1.22**	**0.67**	**0.00**	**0.06**	**0.45**	**0.17**
16		0.25	0.64	1.33	0.74	0.16	0.21	0.51	0.29
16	✓	**0.23**	**0.60**	**1.27**	**0.70**	**0.02**	**0.12**	**0.39**	**0.17**

Closed-Loop Evaluation. To evaluate the effectiveness of our approach under closed-loop control, we follow the Bench2Drive benchmark [31]. As shown in Table III, EchoP markedly surpasses the SSR baseline [7] in Driving Score. Relative to widely used methods, EchoP attains +4.54 and +8.00 points over UniAD [2] and VAD [3] , respectively. With the proposed CFC cycle enabled, EchoP further lifts the closed-loop Success Rate by +10.18% compared to UniAD [2]. Moreover, because EchoP does not rely on auxiliary tasks, it achieves superior runtime efficiency while maintaining high success rates.

TABLE VI

TEMPORAL CONSISTENCY ON THE TEST SET. WE MEASURE THE $\mathcal{L}_{curbev} = \left\| \hat{B}_t - B_t \right\|_2$. LOWER IS BETTER. WE COULD OBSERVE THAT OUR METHOD ENSURE THE BIDIRECTIONAL CONSISTENCY ON THE UNSEEN TEST SET.

Method	$\mathcal{L}_{curbev} \downarrow$
Baseline	0.2442
EchoP (CFC cycle)	**0.0962**

TABLE VII

ABLATION FOR THE LOSS WEIGHT. VARYING $\mathcal{L}_{curbev}$ AND $\mathcal{L}_{futbev}$.

λ_{curbev}	λ_{futbev}	**L2** (m) $\downarrow$				**Collision Rate** (%) $\downarrow$			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
0.1	0.5	**0.23**	**0.60**	**1.27**	**0.70**	0.02	0.12	0.39	0.17
0.5	0.5	0.24	0.63	1.34	0.73	**0.00**	**0.10**	**0.35**	**0.15**
0.1	0.8	0.24	0.63	1.31	0.72	0.04	0.14	0.45	0.21
0.1	0.5	**0.17**	**0.33**	**0.58**	**0.36**	**0.03**	**0.05**	**0.14**	**0.07**
0.5	0.5	0.19	0.36	0.64	0.39	0.08	0.06	0.13	0.09
0.1	0.8	0.18	0.35	0.61	0.38	0.00	0.05	0.21	0.09

B. Ablation Studies and Further Discussion

Connecting methodology to evaluation (open-loop and closed-loop). The three-term objective in total loss is designed to improve not only trajectory accuracy but also scene-consistent planning. L_{traj} directly optimizes waypoint regression (L2 displacement), while L_{futbev} encourages future scene grounding. Crucially, the CFC echo term L_{curbev} regularizes temporal consistency by requiring that the predicted future state can reconstruct the current observation, which reduces physically implausible plans and thus lowers collision rates. In Tab. I, this mechanism is reflected by improved temporal consistency, open-loop L2/collision reductions, and stronger closed-loop robustness on Bench2Drive.

Effect of the CFC cycle. Table IV details the ablation of the CFC BEV cycle loop. Incorporating the CFC cycle into the one-shot planner reduces the average $L2_{MAX}$ error by 0.04 m and CR_{MAX} by 0.12%, while simultaneously lowering the average $L2_{AVG}$ and CR_{AVG} by 0.03 m and 0.08%, respectively. These results underscore the significant efficacy of bidirectional validation via inverse reconstruction, which markedly improves the understanding of dynamic scenes. Consequently, EchoP facilitates safer trajectory generation and delivers substantial planning enhancements. To further quantify the temporal consistency, we compute the Mean Squared Error (MSE) between the Current $\rightarrow$ Future $\rightarrow$ Current predicted BEV and the ground-truth Current BEV on the test set. As shown in Table VI, we benchmark the baseline against our EchoP equipped with the CFC mechanism.

Effect of different scene representation numbers. In Table V, we assess the robustness of the CFC cycle when the scene representation is compressed to different capacities. We compare two settings: 8 and 16 scene tokens, while keeping all other factors fixed. In both cases, the CFC-augmented model secures clear improvements over its non-cycle counterpart. Remarkably, even with only 8 tokens, our method can still learn the surrounding environment of the vehicle very well, confirming that the cycle constraint remains effective even in highly compact representations.

Effect of the loss of weight. As shown in Table VII, we show an ablation study on the weights of our three loss terms, with emphasis on the two components tied to the CFC BEV cycle. Keeping the trajectory loss $\mathcal{L}_{traj}$ fixed, we systematically vary the forward-pass loss $\mathcal{L}_{futbev}$ and the echo-pass loss $\mathcal{L}_{curbev}$. The results show that planning quality remains consistently high across a broad range of weight settings, underscoring the stability imparted by the CFC cycle and confirming that EchoP reliably strengthens planning performance.

Why is the end-to-end CFC BEV cycle loop effective? Accurate trajectory planning requires the model to form a faithful understanding of the driving scene. Most recent planners follow a one-shot paradigm: the network observes a single sensor frame, predicts the future trajectory, and is trained either with auxiliary labels such as detection or mapping or with a forward self-supervised loss that compares the predicted future scene to ground truth. This design leaves two critical gaps. First, heavy auxiliary heads greatly increase model complexity while offering no direct feedback on how well the planner actually understands its surroundings. Second, relying only on forward supervision ignores the reverse constraint: because the driving scene evolves unpredictably, a plan generated from the present frame alone may remain unverified and can drift from reality. EchoP addresses these issues with a CFC cycle. Without any extra supervision, the framework not only learns from the forward prediction of the future scene but also reconstructs the current BEV from that prediction, thereby checking whether the imagined future is consistent with the observed present. This bidirectional loop provides an intrinsic error detector that continuously validates trajectories and significantly lowers the probability of collisions, yielding more reliable planning.

C. Qualitative Results

Visualization protocol. In Fig. 3and Fig. 4, map context is rendered from dataset annotations. The green icon denotes the ego vehicle and red icons denote surrounding agents. We overlay the predicted ego trajectory of EchoP with a representative one-shot baseline and the ground-truth trajectory to highlight behavioral differences under diverse conditions.

Visualization of planning trajectories. In Figure 3, we qualitatively compare trajectories planned by EchoP with ground truth (GT) and GenAD [4] across four representative scenarios. Figure 3 (a) depicts an intersection requiring a left turn. EchoP generates a smoother and more anticipatory turning arc than the GT trajectory, aligning better with the lane geometry. Figure 3 (b) considers driving conditions under rain-induced poor visibility. Despite water puddles and reflections obscuring lane markings, EchoP accurately infers the roadway layout, generating a right-turn trajectory more faithful to the lane than the GT. Figure 3 (c) illustrates navigation through a complex multi-way junction. Here, EchoP closely follows the GT trajectory without crossing road boundaries, whereas GenAD drifts into non-drivable areas. Figure 3 (d) shows a scenario of low-light night driving. Under dim illumination, EchoP closely matches the GT trajectory, while GenAD exhibits significant deviations. These examples validate that the

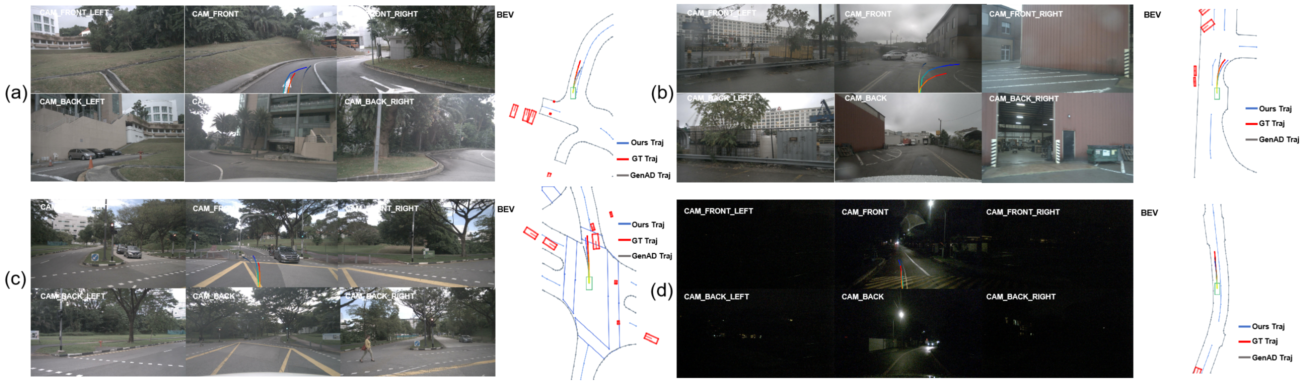


Fig. 3. **Visualization of planning trajectories in different scenes.** We visualize several driving scenes, overlaying the trajectories predicted by our method alongside those of one of the one-shot planners, *i.e.*, GenAD [4], and the ground-truth trajectories. Map context is rendered directly from the dataset annotations. In each figure, the green icon denotes the ego vehicle, while red icons mark surrounding vehicles and other dynamic objects. Video is in **Suppl.**

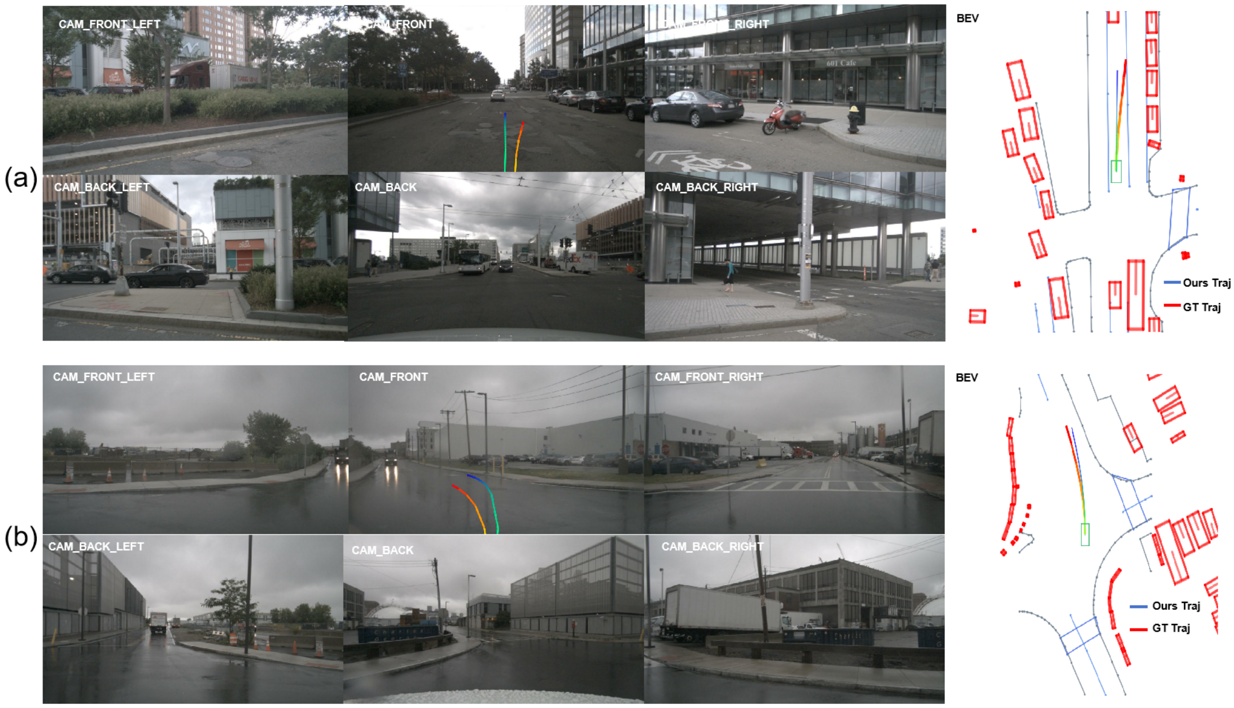


Fig. 4. **Visualization of failure cases.** We highlight two distinct planning failure cases. The first arises when the navigation cue in the ground-truth annotation is ambiguous, and the second occurs in open spaces where the predicted trajectory deviates during a turn. In each figure, the green icon denotes the ego vehicle, while red icons mark surrounding vehicles and other dynamic objects.

CFC cycle in EchoP intrinsically corrects trajectory predictions, enhancing robustness across diverse conditions.

Failure cases. In Figure 4, we show two representative failure modes. The first stems from ambiguous navigation cues in the nuScenes annotation: in Fig. 4 (a), the annotation marks the command as "go straight," yet the ground-truth trajectory actually includes a rightward lane change. When such inconsistencies arise, the planner can be misled and produce a biased path. The second limitation appears on very wide roads that lack explicit lane markings. Although the CFC cycle improves scene understanding, an open roadway grants considerable freedom even for human drivers. As a result, EchoP generates a trajectory that deviates from the ground truth, as in Fig. 4 (b), while still remaining within a plausible driving corridor and preserving the correct turning direction.

Real-world limitations and future directions. In very complex scenarios (e.g., weakly structured open areas, highly multi-modal interactions, and noisy command annotations), a single recorded reference may be insufficient to represent the set of safe and reasonable behaviors. As a result, trajectory/waypoint regression to a single ground-truth target can over-penalize plausible alternatives, especially in open-space or interaction-heavy cases. Future work may explore (i) improving supervision reliability (e.g., route-level command-consistency modeling), (ii) integrating optional geometric priors when available (e.g., drivable-area or road-boundary cues), and (iii) adopting multi-modal planning and corridor-aware evaluation to better reflect non-unique safe futures; deployment-level lightweight safety guardrails in closed-loop systems are also promising complements.

V. CONCLUSION

The Echo Planning framework introduces a self-correcting Current → Future → Current BEV cycle loop (CFC cycle) that endows end-to-end planners with an intrinsic notion of temporal consistency. By predicting a future trajectory, inverting it to reconstruct the present BEV, and penalising any mismatch through a cycle loss, the framework detects and suppresses implausible motion without extra labels or auxiliary heads. Extensive nuScenes and Bench2Drive experiments confirm that this bidirectional feedback reduces displacement error and collision rate beyond state-of-the-art one-shot planners, while incurring no annotation overhead and no additional inference overhead. Echo Planning therefore provides a practical, safety-oriented upgrade for modern autonomous-driving stacks. Overall, this work highlights bidirectional trajectory–scene consistency as a deployable training principle for end-to-end planning, which may guide future robust planner design and benchmark diagnostics for complex scenarios. We also note that extending this cycle-consistency principle to longer-horizon planning and other navigation settings (*e.g.*, UAV navigation [74]) remains an interesting direction for future work.

Acknowledgment. Z. Zheng acknowledges supports from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, the Jiangsu Provincial Science and Technology Program (Grant No. SBZ20250900116), the University of Macau Advanced Research Institute in Hengqin, and the Macao Science and Technology Development Fund Grant FDCT/0043/2025/RIA1.

REFERENCES

[1] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549.

[2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.

[3] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.

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

[5] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "Sparsedrive: End-to-end autonomous driving via sparse scene representation," *arXiv preprint arXiv:2405.19620*, 2024.

[6] X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "Para-drive: Parallelized architecture for real-time autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 449–15 458.

[7] P. Li and D. Cui, "Navigation-guided sparse scene representation for end-to-end autonomous driving," in *The Thirteenth International Conference on Learning Representations*, 2025.

[8] W. Wang, Y. Yang, and F. Wu, "Towards data-and knowledge-driven ai: A survey on neuro-symbolic computing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 2, p. 878–899, Feb. 2025. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3483273

[9] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," *arXiv preprint arXiv:2501.11260*, 2025.

[10] Z. Chen, M. Ye, S. Xu, T. Cao, and Q. Chen, "Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–256.

[11] T. Ye, W. Jing, C. Hu, S. Huang, L. Gao, F. Li, J. Wang, K. Guo, W. Xiao, W. Mao *et al.*, "Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving," *arXiv preprint arXiv:2308.01006*, 2023.

[12] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[13] Y. Zhang, D. Qian, D. Li, Y. Pan, Y. Chen, Z. Liang, Z. Zhang, S. Zhang, H. Li, M. Fu *et al.*, "Graphad: Interaction scene graph for end-to-end autonomous driving," *arXiv preprint arXiv:2403.19098*, 2024.

[14] S. Huang, Z. Shen, Z. Huang, Y. Liao, J. Han, N. Wang, and S. Liu, "Anchor3dlane++: 3d lane detection via sample-adaptive sparse 3d anchor regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, p. 1660–1673, Mar. 2025. [Online]. Available: https://doi.org/10.1109/TPAMI.2024.3508798

[15] R. Zhang, H. Zhang, H. Yu, and Z. Zheng, "Approaching outside: Scaling unsupervised 3d object detection from 2d scene," in *European Conference on Computer Vision*. Springer, 2024, pp. 249–266.

[16] R. Zhang, H. Zhang, and Z. Zheng, "Harnessing uncertainty-aware bounding boxes for unsupervised 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 9230–9240.

[17] J. Lu, M. Nie, B. Zhang, R. Peng, X. Cai, H. Xu, F. Wen, W. Zhang, and L. Zhang, "Translating images to road network: A sequence-to-sequence perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025.

[18] R.-J. Zhu, Z. Wang, L. Gilpin, and J. Eshraghian, "Autonomous driving with spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 136 782–136 804, 2024.

[19] D. Zhang, G. Wang, R. Zhu, J. Zhao, X. Chen, S. Zhang, J. Gong, Q. Zhou, W. Zhang, N. Wang *et al.*, "Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving," *arXiv preprint arXiv:2404.06892*, 2024.

[20] Z. Song, C. Jia, L. Liu, H. Pan, Y. Zhang, J. Wang, X. Zhang, S. Xu, L. Yang, and Y. Luo, "Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 432–22 441.

[21] Y. Suo, Z. Zheng, X. Wang, B. Zhang, and Y. Yang, "Jointly harnessing prior structures and temporal consistency for sign language video generation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1–18, 2024.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[23] D. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Neural Information Processing Systems,Neural Information Processing Systems*, Jan 1988.

[24] F. Codevilla, M. Muller, A. Lopez, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018. [Online]. Available: http://dx.doi.org/10.1109/icra.2018.8460487

[25] J. Cheng, R. Xin, S. Wang, and M. Liu, "Mpnp: Multi-policy neural planner for urban driving," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 549–10 554.

[26] M. Bojarski, D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Apr 2016.

[27] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9329–9338.

[28] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. [Online]. Available: http://dx.doi.org/10.1109/cvpr46437.2021.00700

[29] D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta, "Parting with misconceptions about learning-based vehicle motion planning," in *Conference on Robot Learning*. PMLR, 2023, pp. 1268–1281.

[30] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6119–6132, 2022.

[31] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," in *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[32] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 983–21 994.

[33] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. 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*, 2024, pp. 14 864–14 873.

[34] L. Liu, C. Jia, G. Yu, Z. Song, J. Li, F. Jia, P. Wu, X. Hao, and Y. Luo, "Guideflow: Constraint-guided flow matching for planning in end-to-end autonomous driving," 2025.

[35] B. Sun, Y. Cao, Y. Wang, R. Wang, J. Shang, X. Feng, J. Lu, J. Shi, S. Yang, X. Yan, and Z. Song, "Minddrive: An all-in-one framework bridging world models and vision-language model for end-to-end autonomous driving," 2025.

[36] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Vadv2: End-to-end vectorized autonomous driving via probabilistic planning," 2024.

[37] S. Chen, T. Cheng, X. Wang, W. Meng, Q. Zhang, and W. Liu, "Efficient and robust 2d-to-bev representation learning via geometry-guided kernel transformer," *arXiv preprint arXiv:2206.04584*, 2022.

[38] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. [Online]. Available: http://dx.doi.org/10.1109/iccv48922.2021.01499

[39] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[40] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[41] B. Liao, S. Chen, B. Jiang, T. Cheng, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Lane graph as path: Continuity-preserving path-wise modeling for online lane graph construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 334–351.

[42] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," *arXiv preprint arXiv:2208.14437*, 2022.

[43] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "Reasonnet: End-to-end driving with temporal and global reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 723–13 733.

[44] Z. Liu, S. Chen, X. Guo, X. Wang, T. Cheng, H. Zhu, Q. Zhang, W. Liu, and Y. Zhang, "Vision-based uneven bev representation learning with polar rasterization and surface estimation," in *Conference on Robot Learning*. PMLR, 2023, pp. 437–446.

[45] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[46] H. Ju, S. Huang, S. Liu, and Z. Zheng, "Video2bev: Transforming drone videos to bevs for video-based geo-localization," *ICCV*, 2025.

[47] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 674–21 683.

[48] B. Graham, M. Engelcke, and L. v. d. Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2018.00961

[49] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. [Online]. Available: http://dx.doi.org/10.1109/iccv48922.2021.00315

[50] C. Yuan, Z. Zhang, J. Sun, S. Sun, Z. Huang, C. D. W. Lee, D. Li, Y. Han, A. Wong, K. P. Tee *et al.*, "Drama: An efficient end-to-end motion planner for autonomous driving with mamba," *arXiv preprint arXiv:2408.03601*, 2024.

[51] J. Philion and S. Fidler, *Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D*, Jan 2020, p. 194–210. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-58568-6_12

[52] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, and H. Lu, "Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2321–2336, 2022.

[53] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020–2036, 2025.

[54] Z. Yang, N. Song, W. Li, X. Zhu, L. Zhang, and P. H. Torr, "Deepinteraction++: Multi-modality interaction for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 6749–6763, 2025.

[55] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 229–241, 2019.

[56] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3621–3631.

[57] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: Adaptive space-time tokenization for videos," *Advances in neural information processing systems*, vol. 34, pp. 12 786–12 797, 2021.

[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/cvpr.2018.00745

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Neural Information Processing Systems,Neural Information Processing Systems*, Jun 2017.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[61] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8660–8669.

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

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

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

[65] M. Guo, Z. Zhang, Y. He, K. Wang, and L. Jing, "End-to-end autonomous driving without costly modularization and 3d manual annotation," *arXiv preprint arXiv:2406.17680*, 2024.

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

[67] Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan, "Enhancing end-to-end autonomous driving with latent world model," *arXiv preprint arXiv:2406.08481*, 2024.

[68] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, and X. Wang, "Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 037–12 047, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274192736

[69] B. Zhang, N. Song, X. Jin, and L. Zhang, "Bridging past and future: End-to-end autonomous driving with historical prediction and planning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6854–6863.

[70] B. Sun, B. Zhang, J. Lu, X. Feng, J. Shang, R. Cao, M. Zheng, C. Wang, S. Yang, Y. Cao, and Z. Song, "Focalad: Local motion planning for end-to-end autonomous driving," 2025.

[71] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

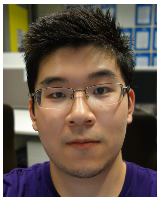
[72] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang, "Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes," 2023. [Online]. Available: https://arxiv.org/abs/2305.10430

[73] Z. Song, L. Liu, H. Pan, B. Liao, M. Guo, L. Yang, Y. Zhang, S. Xu, C. Jia, and Y. Luo, “Diver: Reinforced diffusion breaks imitation bottlenecks in end-to-end autonomous driving,” 2025.
[74] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, “Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching,” in *European Conference on Computer Vision*. Springer, 2024, pp. 213–231.

Jintao Sun received the M.S. degree from The George Washington University, America, in 2021 and the B.E. degree from the Harbin Institute of Technology, China, in 2019. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, Beijing Institute of Technology, China. His current research interests include the image retrieval, vision and language and the autonomous driving.

Hu Zhang is a CERC Research Fellow in CSIRO DATA61, Australia. He received the Ph.D. degree from the University of Technology Sydney in 2022 and the B.S. degree from University of Science and Technology of China in 2017. His research interests include 3D computer vision in Autonomous Driving, 3D reconstruction, imbalanced data learning.

Gangyi Ding received the B.E. degree from Peking University, Beijing, China, in 1988 and the Ph.D. degree from the Beijing Institute of Technology, Beijing, in 1993. He is currently a Professor with the School of Computer Science and Technology, Beijing Institute of Technology. In 1993, he joined the faculty, Beijing Institute of Technology. His research interests include computer simulation, software engineering, and digital performance.



Zhedong Zheng is an Assistant Professor with the University of Macau. He received the Ph.D. degree from the University of Technology Sydney in 2021 and the B.S. degree from Fudan University in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He served as the senior PC for IJCAI and AAAI, and the area chair for ACM MM’24 and ICASSP’25.