OmniNAV: Omniscient Navigation via Unified LiDAR–Camera BEV Fusion for End-to-End Autonomous Driving

Firas Jendoubi, Achref Djaber, Redouane Khemmar, Romain Rossi, Madjid Haddad

OmniNAV: Omniscient Navigation via Unified LiDAR–Camera BEV Fusion for End-to-End Autonomous Driving

1st Firas Jendoubi
Univ Rouen Normandie, ESIGELEC
Normandie Univ, IRSEEM UR4353
F-76000 Rouen, France
Firas.Jendoubi@groupe-esigelec.org

2nd Achref Djaber
Univ Rouen Normandie, ESIGELEC
Normandie Univ, IRSEEM UR4353
F-76000 Rouen, France
Achref.Djaber@groupe-esigelec.org

3rd Redouane Khemmar
Univ Rouen Normandie, ESIGELEC
Normandie Univ, IRSEEM UR4353
F-76000 Rouen, France
Redouane.Khemmar@esigelec.fr

4th Romain Rossi
Univ Rouen Normandie, ESIGELEC
Normandie Univ, IRSEEM UR4353
F-76000 Rouen, France
Romain.Rossi@esigelec.fr

5th Madjid Haddad
SEGULA Technologies
92000 Nanterre, France
Madjid.Haddad@segula.fr

***Abstract*—Achieving reliable autonomous driving requires a unified spatial understanding of dynamic agents and static environments from heterogeneous sensor data. This paper presents OmniNAV, an omniscient end-to-end autonomous driving framework that unifies LiDAR and multi-view camera inputs into a shared Bird's-Eye-View (BEV) representation. Through a unified BEV fusion backbone, the system ensures consistent spatial reasoning and cross-task alignment across all stages of the driving pipeline. OmniNAV jointly performs multi-modal perception, including object detection and tracking, detailed map segmentation, motion forecasting, and safety-aware trajectory planning, within a single differentiable architecture. Experiments on the nuScenes benchmark demonstrate that OmniNAV achieves state-of-the-art performance across perception, prediction, and planning tasks, combining high accuracy, robustness, and reliable decision-making for omniscient navigation. Code is available at : https://anonymous.4open.science/r/OmniNAV-218E.**

***Index Terms*—Autonomous driving, BEV fusion, LiDAR–camera fusion, Multi-sensor perception, Motion forecasting, Planning, Multi-Task Learning, End-to-End learning.**

I. Introduction

Autonomous driving requires a comprehensive understanding of the surrounding environment, combining perception, motion prediction, and planning within a unified decision-making loop. Achieving this capability depends on interpreting both static map structures and dynamic agents under diverse and complex conditions. Traditionally, modular autonomous driving pipelines have decomposed this process into separate stages such as object detection, tracking, motion forecasting, and trajectory planning. While effective in isolation, such separation often leads to information loss, error accumulation, and limited interaction between perception and decision modules.

Recent advances in End-to-End (E2E) learning aim to unify these components within a single differentiable framework. Models such as UniAD [1] have shown the potential of jointly optimizing perception, prediction, and planning for more coherent and interpretable driving behaviors. However, most E2E frameworks rely mainly on camera inputs, which suffer from depth uncertainty and degraded performance under poor lighting or occlusion. Moreover, fusing LiDAR and camera data remains challenging due to spatial misalignment and differences in feature density across modalities.

The Bird's-Eye-View (BEV) representation has emerged as an effective solution for unified scene understanding. By transforming multi-view features into a common top-down space, BEV-based methods enable consistent reasoning across heterogeneous inputs and downstream tasks. Yet, achieving stable and well-aligned BEV fusion across modalities remains difficult. UniBEV [2] demonstrated that shared BEV encoders with channel-weighted fusion can improve alignment and robustness, establishing a strong foundation for multi-sensor perception.

Building on these insights, we propose OmniNAV, an omniscient end-to-end autonomous driving framework that unifies LiDAR and multi-view camera inputs within a shared BEV representation. Our unified BEV fusion backbone enforces spatial consistency, adaptive weighting across modalities, and coherent multi-task interaction. This design enhances both geometric and semantic understanding, allowing joint optimization of perception, motion forecasting, and trajectory planning in a spatially aligned and interpretable manner. Evaluations on the nuScenes benchmark show that OmniNAV achieves state-of-the-art performance across key autonomy tasks. The results confirm that unified BEV fusion and multi-task learning are key enablers for reliable, scene-aware navigation and robust end-to-end autonomous driving.

Our main contributions are summarized as follows:

- We propose a unified multi-sensor BEV fusion backbone that integrates LiDAR and multi-view camera data into a shared spatial representation, enabling robust scene understanding and consistent feature alignment across all end-to-end tasks.
- We introduce a spatially consistent multi-task design that reinforces interactions between perception, prediction, and planning tasks, improving overall accuracy and scene understanding.
- We demonstrate that OmniNAV surpasses state-of-the-art models in perception and prediction while maintaining competitive planning performance, validating the benefits of unified BEV fusion for holistic autonomous driving.

The remainder of this paper is organized as follows. Section II reviews related work on BEV fusion and unified end-to-end autonomous driving frameworks. Section III details the proposed OmniNAV architecture and methodology. Section IV presents the experimental setup and results, including both quantitative and qualitative analyses. Finally, Section V concludes the paper and discusses potential directions for future work.

II. Related Work

A. BEV-Based Representation and Fusion

Bird's Eye View (BEV) perception has become a cornerstone in autonomous driving, offering a unified spatial representation that simplifies reasoning across perception, prediction, and planning tasks.

Camera-based BEV methods transform multi-view images into a unified top-down representation, enabling structured spatial reasoning for downstream perception tasks. Early approaches such as LSS [3] and BEVDet [4] rely on predicted depth distributions to lift image features into 3D frustums, which are then projected into BEV space for downstream tasks like map segmentation and 3D object detection. To capture temporal information, methods like BEVDet4D [5] and SoloFusion [6] integrate historical BEV features with the current frame, while BEVFormer [7] leverages spatio-temporal transformers to model dynamic scene evolution. Although these methods have advanced camera-centric BEV perception, their performance at long range remains limited due to the inherent depth ambiguities of monocular vision.

LiDAR-based BEV approaches address this issue by directly encoding point clouds into structured BEV grids. Techniques such as VoxelNet [8], PointPillars [9], and CenterPoint [10] provide accurate localization but lack dense semantic cues and become sparse at distance.

Multimodal BEV fusion has therefore emerged to combine complementary strengths of LiDAR and cameras. BEVFusion [11] introduced a unified BEV space by concatenating image BEV features (via LSS) with LiDAR voxel features, improving robustness across modalities. MetaBEV [12] refined this with deformable attention-based fusion, while SuperFusion [13] proposed multi-stage integration. Recently, ReliFusion [14] incorporated reliability-aware weighting to dynamically emphasize sensor features depending on their confidence, improving robustness under adverse conditions. GraphBEV [15] addresses misalignment caused by inaccurate calibration between LiDAR and cameras. It introduces a LocalAlign module using neighbor-aware depth features and a Global-Align module to correct BEV misalignment. UniBEV [2] represents a significant step forward by explicitly addressing the alignment challenge in multimodal fusion. Rather than performing late fusion in separate coordinate spaces, UniBEV introduces a unified BEV feature representation where both LiDAR and camera inputs are projected into the same spatial grid. A central innovation in UniBEV is the Channel-wise Normalized Weighting (CNW) fusion module, which adaptively balances modality contributions on a per-channel basis. Instead of naïve concatenation or summation, CNW computes normalized weights that reflect the reliability of LiDAR's geometric cues and the semantic richness of camera features. This fine-grained fusion yields more discriminative and robust BEV representations, maintaining strong performance even when one sensor modality is degraded or missing.

B. End-to-End Autonomous Driving Frameworks

Recent advances in end-to-end autonomous driving have increasingly emphasized joint modeling of perception, prediction, and planning, moving away from traditional modular pipelines to reduce error accumulation and improve system robustness. At the forefront of these frameworks is UniAD [1], which integrates multiple driving tasks into a single transformer-based architecture. UniAD operates on a shared BEV representation, allowing all downstream modules to access consistent spatial information. The framework employs query-based attention mechanisms, where learnable queries represent agents, maps, and the ego-vehicle. Within UniAD, TrackFormer [16] handles object detection and multi-agent tracking by associating agent identities across time, enabling robust temporal reasoning. MapFormer [17] models the road topology and semantic structure of lanes, dividers, and intersections, providing a panoptic map segmentation that informs both prediction and planning. MotionFormer [1] forecasts multi-agent trajectories by modeling interactions among agents and with the environment, producing joint predictions that account for complex scene dynamics. To complement this, OccFormer [1] predicts multi-step occupancy grids with agent identities preserved, improving safety and environmental understanding. The Planner [1] module then leverages the ego-vehicle query to generate collision-free trajectories, explicitly reasoning about predicted occupancy and agent behavior.

Building on UniAD, other frameworks explore complementary approaches. SparseDrive [18] focuses on efficiency, using sparse scene representations and parallel motion planning to reduce computational overhead while maintaining high prediction and planning performance. DriveTransformer [19] simplifies the pipeline by interacting directly with raw sensor inputs, bypassing dense BEV construction. TransFuser and other multimodal fusion methods combine camera and LiDAR inputs to enhance perception and prediction, although some

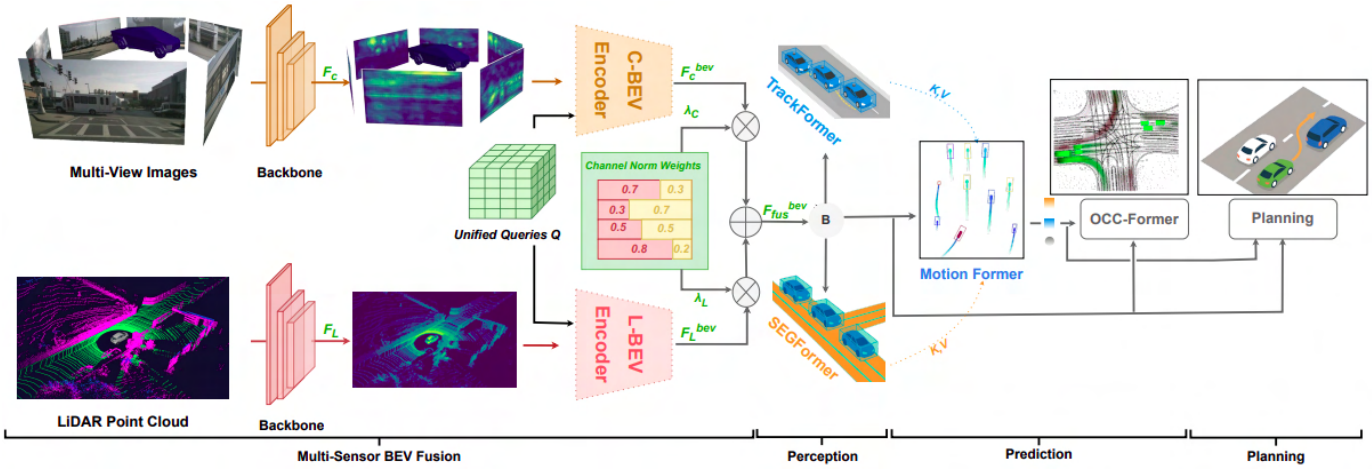


Fig. 1: **OmniNAV system diagram.** Multi-view images and LiDAR are processed by separate backbones, lifted to BEV via camera (C-BEV) and LiDAR (L-BEV) encoders driven by unified queries Q. Channel-Norm Weights produce per-channel coefficients λ_c, λ_ℓ for fusion into $F_{\text{fus}}^{\text{bev}}$. The fused BEV conditions *TrackFormer*, *SEGFormer* (Mapping), *MotionFormer*, *OCC-Former*, and the *Planning* module.

lack full BEV alignment or task integration. Collectively, these works illustrate the evolution from modular to holistic, multi-task autonomous driving systems, with UniAD representing a comprehensive reference for perception, tracking, forecasting, and planning. Our proposed OmniNAV builds directly upon this framework, introducing a unified LiDAR–camera BEV fusion approach that further enhances perception accuracy, motion prediction, and scene understanding in complex driving scenarios.

III. Methodology

In this section, we present OmniNAV, a unified end-to-end autonomous driving architecture that integrates LiDAR and multi-view camera data into a shared Bird's-Eye-View (BEV) representation. As shown in Fig. 1, the framework includes modality-specific backbones for feature extraction, BEV encoders that project each modality into a spatially aligned representation, and a fusion module based on CNW that combines LiDAR and camera BEV features into a single unified map. The fused BEV serves as the core representation for downstream modules handling detection and tracking, map segmentation, motion forecasting, occupancy estimation, and trajectory planning. Together, these components form a cohesive perception-to-planning pipeline that enables OmniNAV to reason jointly across tasks and deliver robust, spatially consistent autonomous navigation.

A. BEV Feature Representation

Feature Extraction. For the LiDAR branch, we adopt a SECOND-based backbone to extract features from voxelized point clouds. This backbone consists of multiple convolutional layers with batch normalization and carefully designed strides to efficiently capture 3D spatial structures. A SECONDFPN neck upsamples and fuses multi-scale features, producing the final LiDAR feature map F_L, as shown in Eq. (1):

$$F_L = \text{SECONDFPN}(\text{SECOND}(V)) \tag{1}$$

where V represents the voxelized LiDAR input.

For the camera branch, a ResNet-101 [20] backbone with deformable convolutions in the later stages extracts high-level semantic features from multi-view images I. A feature pyramid network (FPN) [21] then consolidates these features into a uniform spatial resolution F_C for BEV projection, as defined in Eq. (2):

$$F_C = \text{FPN}(\text{ResNet101}(I)) \tag{2}$$

These modality-specific feature extractors ensure that both LiDAR and camera inputs provide complementary spatial and semantic information for downstream BEV encoding and fusion.

BEV Encoding. Inspired by UniBEV, we define a grid of learnable BEV queries $Q \in \mathbb{R}^{H\times W\times N}$ covering the spatial extent of the scene. Each query is associated with D reference points along the height axis, forming 3D coordinates $R \in \mathbb{R}^{D\times H\times W\times 4}$ in homogeneous form $(x, y, z, 1)$. Here, H and W denote the BEV spatial resolution along the horizontal and vertical axes, respectively, and N represents the feature embedding dimension of each BEV query. In our implementation, we set the BEV grid size to 200×200, providing a balanced trade-off between spatial detail and computational efficiency. These references allow the encoder to capture volumetric context at each BEV location, ensuring accurate geometric alignment across modalities.

Reference points are projected into sensor-specific coordinate systems as defined in Eqs. (3) and (4):

$$R_i^C = P_C(R, P_i), \quad i = 1, \ldots, V \tag{3}$$

$$R^L = P_L(R) \tag{4}$$

where $P_C(\cdot)$ denotes the camera projection function that maps the 3D reference points R into 2D image coordinates using the projection matrix P_i of the i-th camera, and P_L projects R into the LiDAR feature map coordinate system. V denotes the total number of camera views.

BEV features are aggregated using deformable cross-attention, as shown in Eqs. (5) and (6):

$$F_{BEV,C} = \sum_{i=1}^{V} \sum_{z=1}^{D} \text{DeformAttn}(Q, R_i^C(z), F_i^C) \tag{5}$$

$$F_{BEV,L} = \sum_{z=1}^{D} \text{DeformAttn}(Q, R^L(z), F_L) \tag{6}$$

where Q are the BEV queries, F_i^C and F_L denote the camera and LiDAR feature maps, and z indexes depth levels. This step enables each query to integrate multi-scale spatial information from both modalities within a unified BEV space.

After several layers of deformable attention, the BEV maps $F_{BEV,C}$ and $F_{BEV,L}$ preserve the same spatial resolution $H \times W \times N$ as Q, ensuring cross-modal consistency.

For fusion, a Channel-Normalized Weighting strategy is applied instead of concatenation, avoiding zero-padding when a modality is absent. Each channel is assigned learnable per-modality weights $\lambda_C, \lambda_L \in \mathbb{R}^N$, normalized by a softmax. This adaptive weighting allows the model to emphasize the most reliable modality per channel while maintaining a stable feature dimensionality across inputs.

The fused BEV feature map is then computed as in Eq. (7):

$$F_{\text{fus}}^{\text{bev}} = F_{BEV,C} \odot \lambda_C + F_{BEV,L} \odot \lambda_L \tag{7}$$

where $\odot$ denotes channel-wise multiplication with broadcasting over spatial dimensions. This adaptive fusion mechanism allows the network to emphasize the more reliable modality per channel while maintaining consistent feature dimensions even when one sensor is unavailable. When only a single modality is available, the softmax assigns full weight to that modality. This flexible strategy enhances robustness and improves cross-modal feature alignment for downstream perception and motion forecasting tasks.

B. Perception

The BEV features generated by the encoders are processed by a perception module consisting of a dynamic agent branch and a static map branch.

TrackFormer. Inspired by TrackFormer [16], detection and multi-object tracking are jointly performed using a query-based transformer. Detection queries handle newly observed agents, while track queries propagate past states to ensure temporal consistency. Cross-attention with BEV features captures spatial context, and an additional ego-vehicle query explicitly models the self-driving car for planning.

SegFormer (Mapping). For static scene understanding, we adopt a Panoptic SegFormer-inspired [17] design where map queries encode road elements such as lanes, dividers, crossings, and drivable areas. Refined across transformer layers with intermediate supervision, these queries provide structured map representations that support motion forecasting and planning.

C. Prediction

MotionFormer. MotionFormer is the core spatio-temporal prediction module of UniAD, designed to model interactions between dynamic agents and the static environment within the BEV space. It receives agent queries from TrackFormer and semantic map queries from MapFormer, enabling unified reasoning across spatial and temporal contexts. Through transformer-based attention, MotionFormer captures three essential types of interactions: agent–agent, agent–map, and goal–agent. The agent–agent attention models social dynamics among surrounding entities, the agent–map interaction enforces structural and lane-level constraints, and the goal–agent attention refines motion endpoints toward realistic driving intentions. Each query aggregates temporal cues from previous frames, producing coherent and socially consistent trajectory forecasts. This unified reasoning mechanism bridges perception and planning, making MotionFormer a key component for predictive scene understanding in end-to-end autonomous driving.

OccFormer. OccFormer handles future occupancy prediction by modeling how the scene evolves over time in the BEV space. It takes BEV features and agent representations as input to predict multi-step occupancy maps while preserving agent identities. Through transformer-based attention, it captures spatial and temporal dependencies, estimating which areas will be occupied by dynamic or static entities. These predictions provide dense scene understanding that complements MotionFormer's trajectory outputs and guide the planner toward safe and collision-free decisions.

D. Planning

The Planner module, originally introduced in UniAD, is responsible for generating the future trajectory of the ego vehicle using the high-level representations produced by the preceding modules. It takes as input the ego query from MotionFormer, which is enriched with contextual information from perception, prediction, and occupancy reasoning. By leveraging an attention-based interaction mechanism with the BEV feature space, the planner continuously refines its understanding of the scene, outputs a smooth and feasible trajectory that respects map constraints and dynamically avoids obstacles identified by OccFormer. This design ensures coherent and safety-oriented motion generation, effectively closing the perception–prediction–planning loop within the end-to-end OmniNAV framework.

IV. Experimental Results & Analysis

A. Experiment Setup

Dataset & Metrics. For training and evaluation, we utilize the nuScenes dataset [22], a large-scale benchmark designed for autonomous driving research. It provides multi-sensor data collected in diverse urban environments, including six cameras, a LiDAR, and radar sensors, as well as high-definition

TABLE I: Comprehensive evaluation of OmniNAV across all autonomy tasks, including Detection, Tracking, Mapping, Motion Forecasting, Occupancy Prediction and Planning.

Detection		**Tracking**			**Mapping**		**Motion Forecasting**			**Occupancy Prediction**				**Planning**	
mAP↑	NDS↑	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
0.60	0.66	0.570	0.947	1491	0.302	0.675	0.858	1.270	0.186	64.0	47.4	55.3	41.2	1.154	0.941

maps and annotations for 3D detection, tracking, and motion prediction.

To assess performance, we adopt a comprehensive set of task-specific metrics. For perception tasks, we follow the official nuScenes protocol, using mean Average Precision (mAP) and NuScenes Detection Score (NDS) for detection, Average Multi-Object Tracking Accuracy (AMOTA) and Precision (AMOTP) for tracking, and Intersection over Union (IoU) for mapping. For motion prediction, we rely on End-to-end Prediction Accuracy (EPA), Average Displacement Error (ADE), Final Displacement Error (FDE), and Miss Rate (MR) to measure the quality of predicted trajectories. Future occupancy prediction is evaluated using Future Video Panoptic Quality (VPQ) and IoU at both near-range (30 × 30 m) and far-range (100 × 100 m), as proposed in FIERY. Finally, for planning evaluation, we use Displacement Error (DE) and Collision Rate (CR).

Training Details. The proposed model is trained in two stages. In the first stage, we jointly optimize the BEV encoders, backbone, and perception components (including detection, tracking, and mapping modules). This stage is trained for 20 epochs on a regional computing server CRIANN equipped with 8 NVIDIA A100 GPUs (80 GB each), and requires approximately 5 days.

In the second stage, we enable the full end-to-end pipeline by jointly training all perception, prediction, and planning modules. This stage is trained for 21 epochs under the same hardware configuration, with a total training time of about 5 days. We observe that the best performance is achieved at epoch 18, after which convergence saturates.

B. Quantitative Results

Table I summarizes OmniNAV end-to-end performance on nuScenes across detection, tracking, mapping, motion forecasting, occupancy prediction, and planning. Arrows indicate the desired direction (↑ higher is better; ↓ lower is better) for each metric, giving a compact view of system-wide trade-offs.

TABLE II: Detection results on nuScenes: mAP and NDS.

Method	**mAP↑**	**NDS↑**
BEVFormer [7]	0.46	0.4
BEVFusion [11]	0.59	0.58
UniAD [1]	0.39	0.49
OmniNAV (Ours)	**0.60**	**0.66**

TABLE III: Multi-object tracking results on nuScenes.

Method	**AMOTA↑**	**AMOTP↓**	**Recall↑**	**IDS↓**
Immortal Tracker [23]	0.378	1.119	0.478	936
ViP3D [24]	0.217	1.625	0.363	–
QD3DT [25]	0.242	1.518	0.399	–
MUTR3D [26]	0.294	1.498	0.427	3822
UniAD [1]	0.359	1.320	0.467	**906**
OmniNAV (Ours)	**0.570**	**0.947**	**0.644**	1491

Table II and Table III detail the perception performance of OmniNAV in terms of detection and multi-object tracking. For detection, our model achieves the highest mAP (0.60) and NDS (0.66), outperforming BEVFormer, BEVFusion, and UniAD. This confirms that the unified BEV fusion effectively enhances the spatial alignment between LiDAR and camera features, leading to more accurate localization and recognition of objects in complex scenes.

In tracking, OmniNAV shows a substantial improvement over existing methods, achieving an AMOTA of 0.57 and an AMOTP of 0.947. These results indicate that the model not only detects objects reliably but also maintains their temporal consistency across frames, a crucial aspect for understanding dynamic environments. The increase in Recall demonstrates better agent continuity, while the slightly higher IDS count reflects the model's sensitivity to dense and occluded traffic scenarios, which will be further addressed by introducing temporal attention mechanisms in future work. Overall, OmniNAV establishes new state-of-the-art results for joint detection and tracking under a unified BEV-based perception framework.

TABLE IV: Mapping results on the nuScenes dataset (Lanes, Drivable, Divider, Crossing).

Method	**Lanes↑**	**Drivable↑**	**Divider↑**	**Crossing↑**
VPN [27]	18.0	76.0	–	–
LSS [3]	18.3	73.9	–	–
BEVFormer [7]	23.9	77.5	–	–
BEVerse [28]	–	–	30.6	17.2
UniAD [1]	31.3	69.1	25.7	13.8
OmniNAV (Ours)	**48.4**	**81.1**	**45.0**	**36.8**

As shown in Table IV, OmniNAV significantly outperforms prior methods in all segmentation categories, achieving the best IoU for lanes, drivable areas, dividers, and crossings. This improvement confirms that the unified BEV representation captures richer spatial and semantic details, enhancing static scene modeling and providing a reliable foundation for motion forecasting and planning.

Table V presents the motion forecasting results of OmniNAV on the nuScenes benchmark. Compared with PnPNet and ViP3D, OmniNAV achieves a notably higher End-to-End Prediction Accuracy (EPA = 0.626), reflecting more stable and context-aware trajectory estimation. Although UniAD shows

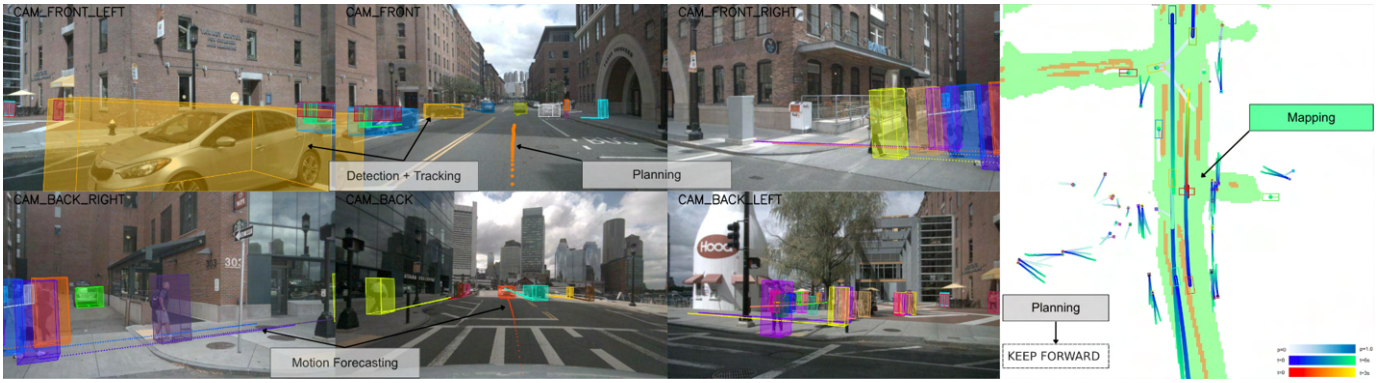


Fig. 2: Qualitative results of the OmniNAV system on a nuScenes example. The six synchronized multi-view camera inputs illustrate perception outputs including 3D object detection, tracking, and motion forecasting for surrounding agents, along with the planned trajectory of the ego vehicle. The right side shows the unified BEV representation combining detection, tracking, mapping, and trajectory planning, demonstrating consistent spatial reasoning across modalities.

TABLE V: Motion forecasting results.

Method	minADE (m)↓	minFDE (m)↓	MR↓	EPA↑
PnPNet [29]	1.15	1.95	0.226	0.222
ViP3D [24]	2.05	2.84	0.246	0.226
UniAD [1]	**0.71**	**1.02**	**0.151**	0.456
OmniNAV (Ours)	0.858	1.270	0.186	**0.626**

slightly lower displacement errors and Miss Rate due to its temporal fusion attention, OmniNAV focuses on unified spatial fusion within the BEV domain, enabling strong scene-level reasoning and consistent multi-agent forecasting. Despite the lack of temporal modeling, OmniNAV delivers competitive motion prediction through robust spatial feature alignment, indicating that integrating temporal attention could further refine trajectory accuracy and consistency.

TABLE VI: Occupancy prediction results.

Method	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑
FIERY [30]	59.4	36.7	50.2	29.9
StretchBEV [31]	55.5	37.1	46.0	29.0
ST-P3 [32]	–	38.9	–	32.1
BEVerse [28]	61.4	40.9	54.3	36.1
UniAD [1]	63.4	40.2	54.7	33.5
OmniNAV (Ours)	**64.0**	**47.4**	**55.3**	**41.2**

TABLE VII: Planning results on nuScenes.

Method	L2(m)↓				Collision Rate(%) ↓			
	1s	2s	3s	**Avg.**	1s	2s	3s	**Avg.**
FF [33]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO [34]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3 [32]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [1]	**0.48**	**0.96**	**1.65**	**1.03**	0.05	0.17	0.71	0.31
OmniNAV (Ours)	0.71	1.32	2.07	1.37	**0.01**	**0.00**	**0.34**	**0.11**

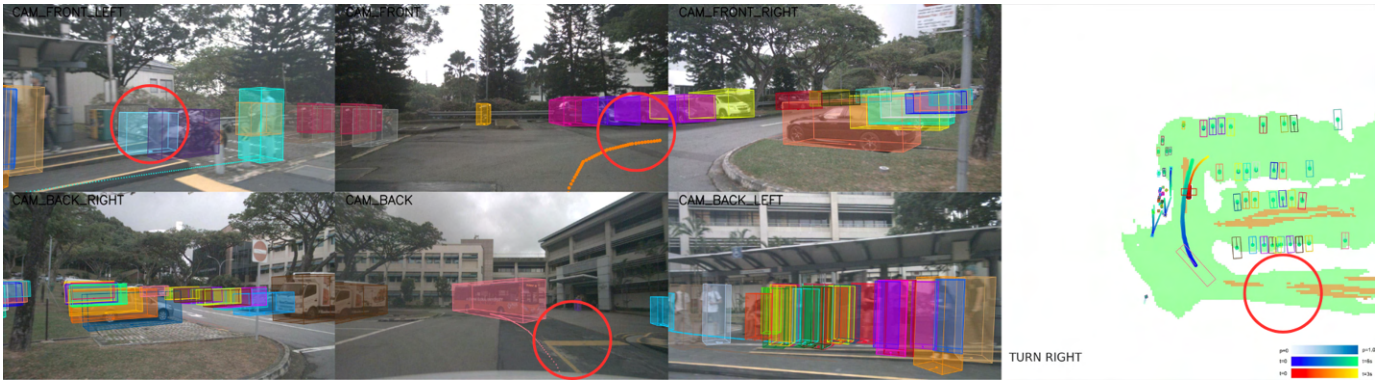
Table VI and Table VII present the results of OmniNAV on occupancy prediction and planning tasks. For occupancy prediction, our model achieves the highest scores across all metrics, with notable gains in IoU and VPQ for both near and far ranges. This improvement indicates that OmniNAV effectively models the spatial distribution of dynamic elements and maintains temporal coherence within the BEV space. The superior VPQ performance demonstrates enhanced consistency in predicting future occupancy, benefiting from the precise LiDAR–camera fusion and strong feature alignment in the unified representation.

For the planning task, OmniNAV achieves the lowest collision rate among all compared methods, demonstrating exceptional safety and robustness in navigation. Although its average L2 displacement error is slightly higher than that of UniAD, this difference reflects a deliberate trade-off between precision and safety. While UniAD produces slightly more accurate trajectory alignment, OmniNAV generates safer, more stable motion plans that minimize the risk of collisions, particularly in complex traffic conditions. This outcome highlights the effectiveness of our unified perception–prediction backbone in supporting reliable decision-making. Future work will incorporate temporal attention mechanisms to refine trajectory smoothness and further reduce displacement errors without compromising safety.

C. Qualitative Results

Figure 2 presents an example of the qualitative results produced by the proposed OmniNAV system on the nuScenes dataset. The visualization demonstrates that the model is capable of accurately detecting and tracking surrounding agents, forecasting their future motion, and generating a safe planned trajectory for the ego vehicle. The BEV representation further highlights precise mapping of drivable areas and lane boundaries, showing consistent integration between perception, prediction, and planning outputs.

Figure 3 illustrates a qualitative comparison between OmniNAV and UniAD. OmniNAV successfully detects smaller or partially occluded objects such as motorcycles that UniAD fails to recognize. In addition, our model provides more complete and reliable mapping results in certain regions where UniAD misses lane or drivable area boundaries. Regarding planning, our model generates smoother and more realistic



(a) Qualitative results from OmniNAV (Ours).

(b) Qualitative results from UniAD.

Fig. 3: Qualitative comparison between OmniNAV and UniAD on the nuScenes dataset. OmniNAV results use segmented bounding boxes to visualize detected objects. Red circles highlight key differences between both models, where OmniNAV demonstrates improved detection completeness, better spatial segmentation, and more coherent trajectory planning compared to UniAD.

ego trajectories that better align with the road geometry and surrounding traffic context, confirming the advantages of its unified BEV fusion design.

D. Discussion

OmniNAV demonstrates strong performance across perception, prediction, and planning, confirming the effectiveness of unified LiDAR–camera BEV fusion for end-to-end autonomous driving. The unified backbone and Channel-Normalized Weighting enable robust spatial reasoning and balanced multi-modal integration, leading to significant improvements in detection, mapping, and occupancy prediction. The unified LiDAR–camera BEV fusion enhances both geometric precision and semantic understanding, leading to improved spatial reasoning across tasks. The incorporation of LiDAR data strengthens depth perception and long-range accuracy, while multi-view camera inputs enrich contextual awareness for motion forecasting and planning. Such capabilities make OmniNAV particularly suited for higher levels of driving autonomy (Levels 3–5), where perception reliability and safe decision-making are essential. Furthermore, the proposed architecture can serve as a foundation for future integration of advanced driver assistance features such as adaptive distance keeping and automatic emergency braking, bridging the gap between current ADAS systems and fully autonomous navigation. However, several challenges remain. The current framework primarily focuses on spatial fusion and does not yet incorporate temporal attention to model long-term dependencies across frames. Integrating temporal reasoning in the BEV domain could further improve motion consistency and planning accuracy. Additionally, exploring lighter architectures and advanced evaluation methods for planning performance will be essential for deployment in real-time systems. Future work will also consider extending the framework to additional tasks, such as intent prediction and scene-level interaction modeling, to achieve a more comprehensive driving policy.

V. Conclusion

In this paper, we introduced OmniNAV, a unified end-to-end autonomous driving framework that fuses LiDAR and multi-view cameras within a shared BEV representation. The proposed architecture jointly addresses perception, prediction, and planning through a unified fusion backbone that ensures strong spatial alignment and task consistency. Extensive experiments on the nuScenes benchmark show that OmniNAV achieves excellent performance across detection, tracking, mapping, mo-

tion forecasting, and planning tasks. The results demonstrate the effectiveness of unified BEV fusion for improving both spatial understanding and temporal reasoning within a single framework. Future work will focus on designing a lighter and more efficient version of the system, integrating temporal attention for enhanced motion modeling, and extending the framework to higher-level tasks such as behavior prediction and anomaly detection.

ACKNOWLEDGMENT

This work was performed, in part, on computing resources provided by CRIANN (Centre Regional Informatique et d'Applications Numeriques de Normandie, Normandy, France).

REFERENCES

[1] 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.

[2] S. Wang, H. Caesar, L. Nan, and J. F. Kooij, "Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2776–2783.

[3] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European conference on computer vision*. Springer, 2020, pp. 194–210.

[4] 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.

[5] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[6] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection," *arXiv preprint arXiv:2210.02443*, 2022.

[7] 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 multi-camera images via spatiotemporal transformers. arxiv 2022," *arXiv preprint arXiv:2203.17270*.

[8] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[10] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[11] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

[12] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for 3d detection and map segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8721–8731.

[13] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "Superfusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, 2022.

[14] R. Sadeghian, N. Hooshyaripour, C. Joslin, and W. Lee, "Reliability-driven lidar-camera fusion for robust 3d object detection," *arXiv preprint arXiv:2502.01856*, 2025.

[15] Z. Song, L. Yang, S. Xu, L. Liu, D. Xu, C. Jia, F. Jia, and L. Wang, "Graphbev: Towards robust bev feature alignment for multi-modal 3d object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 347–366.

[16] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8844–8854.

[17] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu, "Panoptic segformer: Delving deeper into panoptic segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1280–1289.

[18] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "Sparsedrive: End-to-end autonomous driving via sparse scene representation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8795–8801.

[19] X. Jia, J. You, Z. Zhang, and J. Yan, "Drivetransformer: Unified transformer for scalable end-to-end autonomous driving," *arXiv preprint arXiv:2503.07656*, 2025.

[20] 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.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[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] Q. Wang, Y. Chen, Z. Pang, N. Wang, and Z. Zhang, "Immortal tracker: Tracklet never dies," *arXiv preprint arXiv:2111.13672*, 2021, preprint, submitted 26 November 2021.

[24] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Vip3d: End-to-end visual trajectory prediction via 3d agent queries," *arXiv preprint arXiv:2208.01582*, 2022, revised version posted 19 June 2023 (v3).

[25] H.-N. Hu, Y.-H. Yang, T. Fischer, T. Darrell, F. Yu, and M. Sun, "Monocular quasi-dense 3d object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1992–2008, 2022.

[26] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Mutr3d: A multi-camera tracking framework via 3d-to-2d queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4537–4546.

[27] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[28] 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.

[29] R. Qian, Y. Li, H. Sun, C. Hu, Q. Wang, H. Chen, and D. Lin, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5881–5890.

[30] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance segmentation in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[31] A. K. Akan and F. Güney, "Stretchbev: Stretching future instance prediction spatially and temporally," in *European Conference on Computer Vision (ECCV)*, 2022.

[32] 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 (ECCV)*, 2022.

[33] 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.

[34] 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.