2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

DualAT: Dual Attention Transformer for End-to-End Autonomous Driving

Zesong Chen*, Ze Yu*, Jun Li, Linlin You, and Xiaojun Tan✉

***Abstract*— The effective reasoning of integrated multimodal perception information is crucial for achieving enhanced end-to-end autonomous driving performance. In this paper, we introduce a novel multitask imitation learning framework for end-to-end autonomous driving that leverages a dual attention transformer (DualAT) to enhance the multimodal fusion and waypoint prediction processes. A self-attention mechanism captures global context information and models the long-term temporal dependencies of waypoints for multiple time steps. On the other hand, a cross-attention mechanism implicitly associates the latent feature representations derived from different modalities through a learnable geometrically linked positional embedding. Specifically, the DualAT excels at processing and fusing information from multiple camera views and LiDAR sensors, enabling comprehensive scene understanding for multitask learning. Furthermore, the DualAT introduces a novel waypoint prediction architecture that combines the temporal relationships between waypoints with the spatial features extracted from sensor inputs. We evaluate our approach on both the Town05 and Longest6 benchmarks using the closed-loop CARLA urban driving simulator and provide extensive ablation studies. The experimental results demonstrate that our approach significantly outperforms the state-of-the-art methods.**

I. Introduction

The end-to-end autonomous driving paradigm [1]–[4] has recently garnered significant attention due to its capacity to collectively optimize perception, prediction, and planning tasks, substantially reducing the amount of required computing resources. Camera-LiDAR fusion technology can yield more comprehensive and robust perception information for autonomous vehicles, enhancing the subsequent prediction and planning tasks. Many previously developed approaches [5]–[7] achieve feature alignment through the transformation of geometric projections between the image space and the LiDAR projection space, such as bird's-eye-view (BEV). The geometric projection establishes a hard one-to-one association between sparse point clouds and dense image pixels, which not only leads to the loss of numerous image features but also heavily depends on high-quality calibration between the two sensors. However, achieving accurate calibration is often challenging due to the inherent spatial-temporal misalignment problem [8]. Furthermore, CNN-based methods are limited to aggregating the multimodal features contained within a local neighborhood [9]. However, in complex traffic scenarios, the ego agent must be capable of attending to multiple regions, such as traffic lights above and other dynamic agents ahead. Attention mechanisms [10] are naturally suitable for global context reasoning. In this paper, we introduce a dual attention transformer (DualAT) to achieve enhanced agent perception. As shown in Fig. 1, self-attention captures the global context, enabling agents to process various informational elements in complex traffic scenarios, such as intersections with traffic lights. Furthermore, cross-attention merges the features derived from two distinct modalities in the BEV space by establishing a soft association using a learnable geometrically linked positional embedding. This innovative approach effectively addresses the misalignment issue stemming from the hard associations formed by sensor calibration.

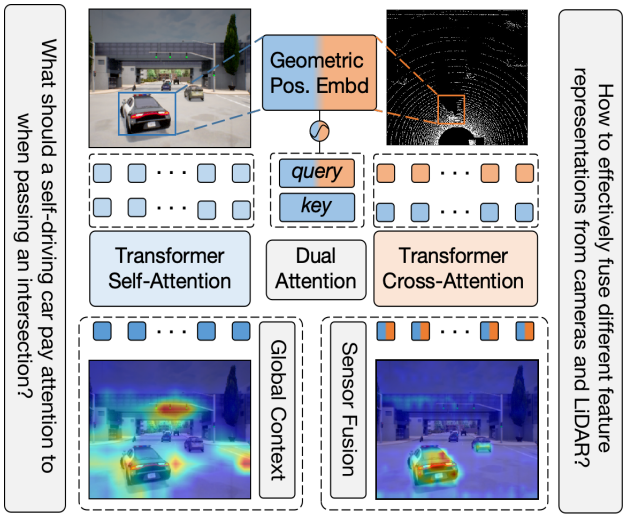


Fig. 1: The agent needs to obtain multimodal global perception information to effectively handle complex traffic scenarios. Our DualAT can integrate the capture of global context information and the fusion of LiDAR and multiview camera data, providing comprehensive and robust perception information for safe navigation.

In end-to-end autonomous driving cases [4], [9], [11], [12], a common waypoint prediction method involves using a gated recurrent unit (GRU) [13] network to predict a series of waypoints. The network is initialized with a low-dimensional feature, which is obtained by extracting the sensor input through a backbone network. However, this approach may result in the loss of perception information

This work was supported by the Key-Research and Development Prgoram of Guangdong Province under Grand (2020B0909030005, 2019B090913001).

*Denotes equal contribution.

The authors are with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China. ✉X. Tan is the corresponding author. Email: {chenzs5, yuze5}@mail2.sysu.edu.cn, {stslijun, youllin, tanxj}@mail.sysu.edu.cn

related to small objects, such as pedestrians and bicycles. In this paper, we propose a novel waypoint prediction architecture based on a transformer model. Unlike previously developed autoregressive waypoint prediction frameworks based on GRUs, our model generates all future waypoints in parallel by utilizing masked self-attention to model the long-term temporal dependencies of waypoints. Additionally, cross-attention is utilized to query perception features that are more pertinent to waypoint prediction, which can produce intermediate attention maps that aid in interpreting how planning decisions are generated. Note that we perform cross-attention calculations on the high-resolution perception features obtained by deconvolution layers, these calculations do not impose a computational resource burden since waypoint queries are small and fixed, and the computational complexity of attention scales linearly with the feature resolution. The contributions of this paper are summarized as follows.

- We propose a DualAT sensor fusion network, where a cross-attention mechanism learns to map features from multiple camera views to the BEV, employing a learnable geometrically linked positional embedding. Additionally, a self-attention mechanism captures the global context of multimodal features.
- We propose a novel DualAT-based waypoint prediction network, which combines temporal relationships between waypoints with spatial features extracted from sensor input. Furthermore, we visualize the intermediate attention map of DualAT to improve the interpretability of end-to-end predictions.
- We demonstrate the state-of-the-art performance of our approach on the Longest6 and Town05 benchmarks and provide extensive ablation studies to verify the effectiveness of the proposed modules.

II. Related Works

A. End-to-End Autonomous Driving

In end-to-end autonomous driving, there are generally two forms of learning driving policies: direct control actions and trajectories/waypoints. CIL [14] and CILRS [15] utilize distinct action prediction branches based on various high-level commands to accomplish navigation tasks in urban environments. LBC [16] employs knowledge distillation to train sensor agents that are capable of imitating the driving actions of privileged information agents. The advantage of waypoint-based prediction lies in its consideration of relatively long time horizons in the future, which can effectively reduce the risk of collisions. Transfuser [9] employs a GRU network to autoregressively generate a series of waypoints, and this method has gained popularity in the waypoint prediction domain. TCP [17] utilizes a dual-branch GRU network that not only predicts waypoints but also employs them to guide action predictions across multiple time steps. Interfuser [12] integrates a transformer module before the GRU to avoid the loss of perception feature space information caused by the pooling operation.

B. Sensor Fusion for Autonomous Driving

Deepfusion [18] employs cross-attention to match multiple image pixels with a LiDAR voxel and merge their features. Transfusion [19] conducts cross-attention at the object level to combine corresponding image features for objects predicted in BEV. GuideFormer [20] has developed two SwinT [21] architecture branches, employing a cross-attention mechanism within a local window to guide the completion of sparse depth images using complementary semantic information derived from RGB images. LAV [4] uses pointpainting [22] to incorporate the RGB features from the image into the original point cloud. Transfuser [9], [11] employs a CNN-transformer architecture to integrate multimodal features at multiple scales. Similar to DETR [23], Interfuser [12] implements a transformer-based encoder-decoder architecture to handle multiview camera and LiDAR features. ThinkTwice [24] employs LSS [25] to transform 2D images into BEV space and concatenate them with LiDAR BEV features.

C. Multitask Learning

Supervising a model with additional tasks can enhance its learning ability and accelerate its training process by leveraging shared features. MT-CILRS [26] extends the CILRS framework by incorporating an independent convolutional block attention module(CBAM) [27] for each vision task. NEAT [28] obtains a BEV representation for joint trajectory planning and BEV semantic prediction by iterating the intermediate attention map multiple times. LP2 [29] applies a joint design philosophy to the tasks of localization, perception, and prediction to alleviate the effects of localization errors. TransMEF [30] employs a transformer-based framework for multiexposure image fusion, completing three self-supervised reconstruction tasks through multitask learning. BEVFusion [31] is a versatile framework for multitask and multisensor perception that fuses multiview cameras and LiDAR sensors in a shared BEV space, enabling multitask 3D perception.

III. Method

The overall architecture of our method is shown in Fig. 2. In this section, we provide overviews of its two important components: the multimodal fusion-based masked transformer and the multitask learning module.

A. Multimodal Fusion-Based Masked Transformer

Given a set of $n = 3$ monocular views $(I_k, K_k, R_k, t_k)_{k=1}^{n}$ consisting of an input image $I_k \in \mathbb{R}^{H\times W\times 3}$, intrinsic camera rotation $K_k \in \mathbb{R}^{3\times 3}$, extrinsic rotation $R_k \in \mathbb{R}^{3\times 3}$, and an offset $t_k \in \mathbb{R}^3$ between the origins of the camera and ego-vehicle coordinate systems. In addition, $\mathbf{F}_I \in \mathbb{R}^{N_I\times C}$ and $\mathbf{F}_B \in \mathbb{R}^{N_B\times C}$ refer to the flattened image and BEV feature.

1) BEV masks: The BEV grid involves the combined FOVs of multiple cameras. We intend to split the BEV grid into different views based on the BEV masks, each of which corresponds to the FOV of a single camera. A perspective

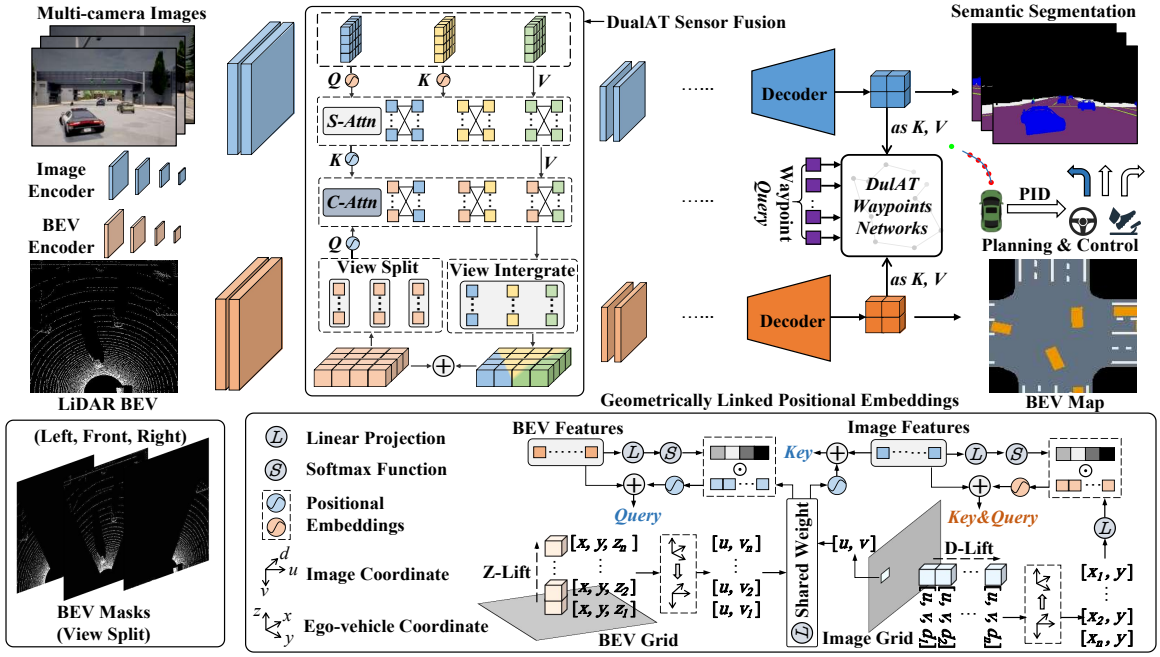


Fig. 2: **The overall architecture of the DualAT.** For the image branch, we utilize three cameras (left, front, and right), with each camera possessing an input image resolution of 224×300 and a receptive field of 60°. This configuration provides a total coverage of 180°. In the LiDAR branch, we transform the acquired LiDAR point cloud data into a histogram on a 2-dimensional 192×192 BEV grid. This grid provides a resolution of 4 pixels/m, which is equivalent to 48 m in front of the ego vehicle and 24 m on each side. We employ a pretrained ResNet-34 [32] for the image encoder and ResNet-18 for the BEV encoder. Both decoders consist of 5 deconvolutional layers with 2x upsampling operations. Note that we utilize the features following the second deconvolutional layer as the inputs of the waypoint prediction network.

transformation from the ego-vehicle coordinate to the image coordinate can be described as follows:

$$d_c P^{(I)} = K_k R_k (P^{(E)} - t_k), \tag{1}$$

where $P^{(E)} = (x, y, z)$ represents any ego-vehicle coordinate. $P^{(I)} = (u_i, v_i, 1)$, and d_c represents the depth value of each image pixel (u_i, v_i) in the camera coordinate system. Denote $P^{(B)} = (u_b, v_b, z)$ as the BEV coordinate system, which represents the location of the BEV grid. z represents the height value of each BEV pixel (u_b, v_b), which can be considered a constant since we only consider the 2D BEV grid and set $z = 1$ in this work. $P^{(B)}$ and $P^{(E)}$ can be mutually transformed through simple rotation and translation operations. A corresponding mask can be constructed for the BEV grid based on the pixel range of each camera:

$$\mathbf{Mask}^{(B)}(u_b, v_b) = \begin{cases} 1 & u_i \in W_k, v_i \in H_k \\ 0 & other \end{cases}, \tag{2}$$

where $W_k = [w_k^{min}, w_k^{max})$ and $H_k = [h_k^{min}, h_k^{max})$ represent the width and height ranges of camera k, respectively. We visualize the BEV masks as shown in the bottom-left corner of Fig. 2. When sensor fusion is completed, the split views are integrated back into a joint view.

2) Geometrically linked positional embeddings: The significant discrepancy in feature distribution between the RGB and 3D point cloud poses a challenge for the transformer in accurately establishing correlations between these two distinct modalities. In this work, we utilize learnable geometrically linked positional embeddings to establish latent spatial connections between the two modalities. In cross-attention, we project the ego-vehicle coordinate onto the image coordinate system, thus creating soft associations in the perspective space between the features derived from two modalities to guide attention learning. In self-attention, to align with cross-attention, we reverse-project the image coordinates into the ego-vehicle coordinate system, enabling soft associations among the image features in the BEV space. These two positional embeddings $\mathbf{E}^{(I)} \in \mathbb{R}^{N_I \times C}$, $\mathbf{E}^{(B)} \in \mathbb{R}^{N_B \times C}$ are computed as follows:

$$\mathbf{E}^{(I)} = softmax(\mathbf{L}(\mathbf{F}_I)) \odot \mathbf{L}(R_k^{-1} K_k^{-1} d_c P^{(I)} + t_k), \tag{3}$$

$$\mathbf{E}^{(B)} = softmax(\mathbf{L}(\mathbf{F}_B)) \odot \mathbf{L}(K_k R_k (P^{(E)} - t_k)/d_c), \tag{4}$$

where $\mathbf{L}(\cdot)$ represents the linear projection operation. The BEV space lacks the height dimension, while the perspective space lacks the depth dimension. Inspired by [25], [33], we lift both the Z-axis and D-axis to supplement the missing dimensions. Finally, we integrate the features of the complementary dimensions through a weighted averaging operation. Note that our model does not rely heavily on depth or height estimation but provides additional information for attention learning by encoding spatial relationships into positional embeddings. The bottom-right corner of Fig. 2 illustrates the calculation process of positional embeddings in detail.

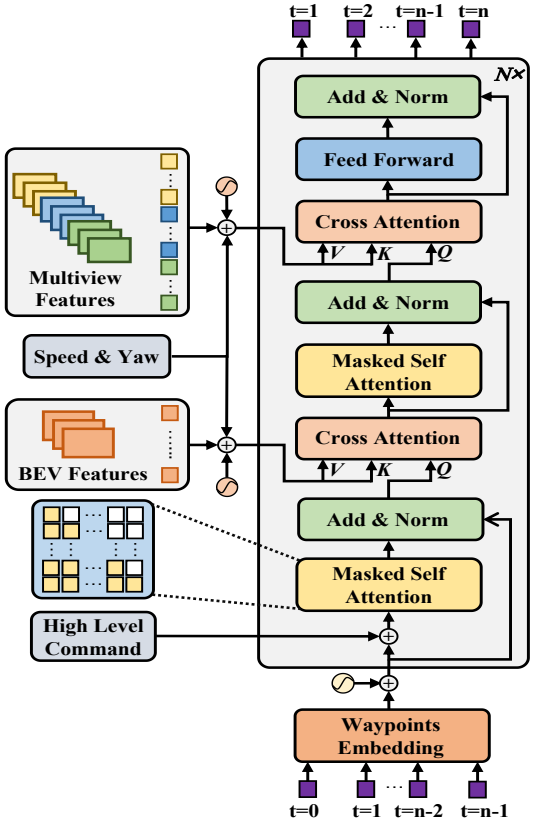


Fig. 3: **Illustration of the waypoint prediction network.** We introduce a novel waypoint prediction network based on a transformer. This network utilizes masked self-attention to capture long-term temporal dependencies in trajectories and employs cross-attention to fuse critical perception information for guiding the waypoint generation process.

3) DualAT sensor fusion: We fuse multimodal features using a cross-attention mechanism and capture the global context using a self-attention mechanism. We operate on grid-structured intermediate feature maps. Specifically, we transform a feature sequence $\mathbf{F} \in (\mathbf{F}_I, \mathbf{F}_B)$ into the query, key, and value matrices using linear projections as follows:

$$\mathbf{Q} = \mathbf{F}\mathbf{M}^q + \mathbf{E}, \ \mathbf{K} = \mathbf{F}\mathbf{M}^k + \mathbf{E}, \ \mathbf{V} = \mathbf{F}\mathbf{M}^v, \quad (5)$$

where $\mathbf{M}^q, \mathbf{M}^k, \mathbf{M}^v \in \mathbb{R}^{C\times C}$ are projection matrices, and $\mathbf{E} \in (\mathbf{E}^{(I)}, \mathbf{E}^{(B)})$ is the positional embedding that was discussed in Sec. III-A.2. The self-attention is computed as follows:

$$S\text{-}Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}})\mathbf{V} \quad (6)$$

For self-attention, the queries, keys, and values $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ are calculated only within a single modality, while cross-attention is computed between two modalities. The cross-attention between these modalities is computed as follows:

$$C\text{-}Attn_{I\to B}(\mathbf{Q}_B, \mathbf{K}_I, \mathbf{V}_I) = softmax(\frac{\mathbf{Q}_B\mathbf{K}_I^T}{\sqrt{C}})\mathbf{V}_I \quad (7)$$

The cross-attention mechanism attends to $\mathbf{V}_I$ by considering the cross-modal similarity between $\mathbf{Q}_B$ and $\mathbf{K}_I$. Our learning model can easily identify this similarity because we combine BEV masks and geometrically linked positional embeddings as auxiliary information for the attention computation. Finally, by adding the image features $\mathbf{V}_I$ to the BEV features $\mathbf{F}_B$ through cross-modal dependencies, the LiDAR branch integrates both geometric and RGB semantic information. Note that the fusion module is applied at each ResNet layer, resulting in multimodal fusions at four different resolutions.

B. Multitask Learning

1) Waypoint prediction network: Fig. 3 shows the details of our proposed waypoint prediction network. Given a series of n waypoints $\mathbf{W}^{(in)} = (w_0, w_1, w_2, \ldots, w_{n-2}, w_{n-1})$, our goal is to predict n waypoints in the future: $\mathbf{W}^{(out)} = (w_1, w_2, w_3, \ldots, w_{n-1}, w_n)$. Specifically, we use the previous waypoints to predict the next waypoint: $w_0 \to w_1, w_1 \to w_2, \ldots, w_{n-2} \to w_{n-1}, w_{n-1} \to w_n$. Note that $w_0 = (0, 0)$ is fixed and represents the origin of the ego-vehicle coordinate system. The learning model must account for the correlations among waypoints. We use a self-attention mechanism to model these long-term temporal dependencies. Moreover, we use a $\mathbf{Mask}^{(W)}$ to prevent the model from receiving ground-truth waypoints during the training phase. The masked self-attention (MSA) is calculated as follows:

$$MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}} + \mathbf{Mask}^{(W)})\mathbf{V}, \quad (8)$$

$$\mathbf{Mask}^{(W)}(i, j) = \begin{cases} 0 & j \le i \\ -\infty & j > i \end{cases}, \quad (9)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n\times C}$ are the C-dimensional waypoints embedding obtained by encoding the 2-dimensional waypoints using a shared MLP layer. $\mathbf{Mask}^{(W)} \in \mathbb{R}^{n\times n}$ ensures that the t_{th} waypoint can only calculate attention weights with itself and the previous $t-1$ waypoints $(1, 2, ..., t-1)$. Immediately afterward, all waypoint embeddings utilize cross-attention(Eq. 7) to query their corresponding intermediate image and BEV features.

2) Semantic segmentation and BEV map: For semantic segmentation, we consider seven classes: (1) unlabeled, (2) vehicles, (3) roads, (4) red lights, (5) pedestrians, (6) lane markings, and (7) sidewalk objects. For the BEV map, we consider four classes: (1) unlabeled, (2) vehicles, (3) roads, and (4) lane markings.

3) Loss function: We train the waypoint prediction network using the L_1 loss:

$$\mathcal{L}_{\text{pt}} = \sum_{t=1}^{n} ||\mathbf{w}_t - \mathbf{w}_t^{\text{gt}}||_1, \quad (10)$$

where $\mathbf{w}_t^{\text{gt}}$ represents the ground-truth waypoint for time step t acquired from the expert. We utilize the cross-entropy loss for semantic segmentation and the focal loss [38] for BEV map prediction. We train the end-to-end network by calculating the weighted average of all losses:

$$\mathcal{L} = \lambda_{\text{pt}}\mathcal{L}_{\text{pt}} + \lambda_{\text{Seg}}\mathcal{L}_{\text{Seg}} + \lambda_{\text{Map}}\mathcal{L}_{\text{Map}}, \quad (11)$$

TABLE I: **Performance on the Longest6 benchmark.** * denotes the main metric, and ↑ means that higher values are better. Expert denotes the distillation from privileged agents' outputs or features. Seg and Depth denote the 2D semantic segmentation and depth estimation. Box denotes the LiDAR 3D object detection.

Method	Cameras	LiDAR	Auxiliary Losses	Reference	DS*↑	RC↑	IS↑
WOR [34]	1	✗	Expert	ICCV 2021	17.3 ± 3.0	43.5 ± 3.0	0.54 ± 0.06
NEAT [28]	1	✗	Map	ICCV 2021	24.1 ± 3.3	59.9 ± 0.5	0.49 ± 0.02
TCP [17]	1	✗	Expert	NeurIPS 2022	42.9 ± 0.6	61.8 ± 4.2	0.71 ± 0.04
LAV [4]	4	✓	Expert+Seg+Map+Box	CVPR 2022	48.4 ± 3.4	80.7 ± 0.8	0.60 ± 0.04
Transfuser [11]	3	✓	Dep+Seg+Map+Box	TPAMI 2022	47.3 ± 5.7	**93.4**± 1.2	0.50 ± 0.06
InterFuser [12]	4	✓	Map+Box	CoRL 2022	47.0 ± 6.0	74.0 ± 1.0	0.63 ± 0.07
Perception PlanT [35]	3	✓	Dep+Seg+Map+Box	CoRL 2022	57.7 ± 5.0	88.2 ± 0.9	0.65 ± 0.06
CaT [36]	3	✗	Expert+Map	CVPR 2023	58.4 ± 2.2	78.8 ± 1.5	**0.77**± 0.02
DualAT	3	✓	Seg+Map	Ours	**67.2**± 2.3	89.5 ± 3.9	0.76 ± 0.05

TABLE II: **Performance on the Town05 Long benchmark.**

Method	DS*↑	RC↑	IS↑
CILRS [15]	7.8 ± 0.3	10.3 ± 0.0	0.75 ± 0.05
LBC [16]	12.3 ± 2.0	31.9 ± 2.2	0.66 ± 0.02
Transfuser [11]	31.0 ± 3.6	47.5 ± 5.3	**0.77** ± 0.04
Roach [37]	41.6 ± 1.8	**96.4** ± 2.1	0.43 ± 0.03
LAV [4]	46.5 ± 2.3	69.8 ± 2.3	0.73 ± 0.02
TCP [17]	57.2 ± 1.5	80.4 ± 1.5	0.73 ± 0.02
InterFuser [12]	51.6 ± 3.4	88.9 ± 2.5	0.59 ± 0.05
DualAT (Ours)	**60.7** ± 3.0	91.8 ± 3.6	0.66 ± 0.02

where λ is the weight of each task and is adjusted by observing the validation errors of multiple experiments.

IV. Experiments

A. Experimental Setup

1) Dataset: We implemented our approach using the open-source CARLA simulator [39] (version 0.9.10.1), which covers 8 towns and 21 different weather conditions. We use an expert agent with access to privileged information from the CARLA simulator to collect a dataset consisting of 253K frames across all kinds of towns and weather conditions, captured at a rate of 2 FPS.

2) Benchmark: To validate the ability of a sensor-only agent (without HD maps) to effectively manage congested traffic scenarios and extend its applicability to unseen towns, we evaluate our approach using the Longest6 [11] and Town05 Long [9] benchmarks. The Longest6 benchmark contains 6 towns, each of which includes 6 routes, totaling 36 routes, with an average length of 1.7 km. The Town05 Long benchmark consists of 10 long routes ranging from 1000 to 2000 m. On the Longest6 benchmark, we use all towns for training. On the Town05 Long benchmark, we retain Town05 for evaluation purposes and use all other towns for training.

3) Metrics: Three official metrics from the CARLA Leaderboard are used. **Route completion (RC)** denotes the completed percentage of the total route distance. The **infraction score (IS)** is the weighted route completion score, which is computed by accounting for infractions made along the route. The main metric is the **driving score (DS)**, which is calculated as the multiplication of the route completion and infraction scores.

TABLE III: **Ablation for the waypoint decoder. Veh** denotes collisions with vehicles, and **Red** denotes red light infractions.

Waypoint decoder	DS↑	RC↑	IS↑	Veh↓	Red↓
MLP+GRU	26.9	65.3	0.51	0.14	0.02
Transformer+GRU	30.3	**78.4**	0.50	0.13	0.02
DualAT head	**43.7**	69.8	**0.68**	**0.04**	**0.01**

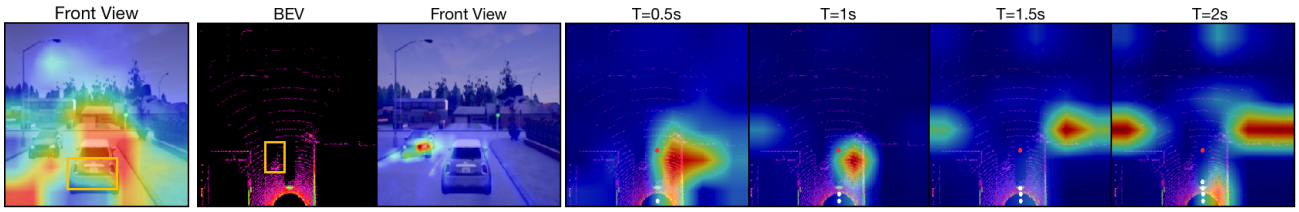
B. Comparison with the State-of-the-Art Methods

The results are shown in Table I and Table II. On the Longest6 benchmark, the DualAT achieves the highest RS and ranks second in terms of the RC and IS metrics. This indicates that the agent effectively handles nearly all driving scenarios while maintaining a high infraction score (low infraction rate), highlighting the robustness of our approach. On the Town05 benchmark, the DualAT achieves the highest RS and produces the second-best RC. These are similar to the results obtained on the Longest6 benchmark. However, the DualAT performs moderately well in terms of the IS metric, which is attributed to the omission of stop signs by our expert agent during the data collection process, resulting in the models acquired through imitation learning also disregarding stop signs and thus causing infractions by directly crossing them. Moreover, some methods [4], [12], [37] have stop signal training labels, so they perform a stop operation on the controller when the agent detects a stop signal. Compared with these methods, our approach have a natural disadvantage, but we finally obtain the highest DS, showing that our method performs better on other infractions.

C. Ablation Study

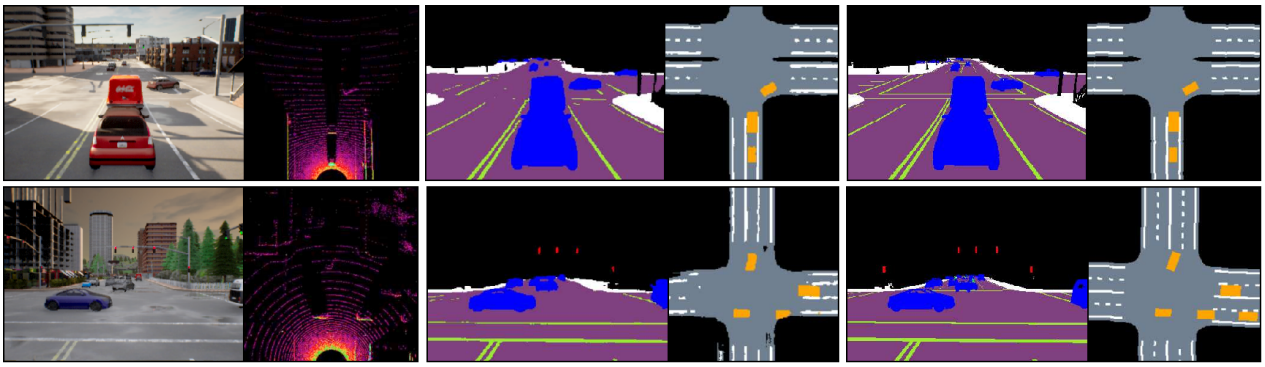
In this section, we conduct ablation studies on three critical components, the waypoint decoder, attention block, and multitask learning, using the **Town05 Long benchmark**.

1) Waypoint decoder: To demonstrate the effectiveness of our waypoint decoder, we develop a basic perception module: the sensor inputs undergo downsampling through a ResNet-34 backbone. We consider the two other most prevalent waypoint decoder architectures, MLP+GRU and Transformer+GRU(InterFuser [12] head), alongside our newly proposed DualAT. The results are presented in Table III.



(a) Self-Attention (b) LiDAR-Camera Cross-Attention (c) Waypoint Attention for Multiple Time Steps

Fig. 4: Visualizations of the dual attention mechanism. The dual attention mechanism in the sensor fusion module is shown in (a) and (b). Note that the yellow box is the query. The cross-attention in the waypoint decoder is shown in (c). Note that the white point is the predicted waypoint, and the red point is the target point.



(a) Sensor Inputs (b) Auxiliary Task Prediction (c) Ground Truth

Fig. 5: Visualizations of the auxiliary tasks. We choose two cases to showcase the auxiliary task predictions produced in the Longest6 (above) and Town05 (below) benchmark evaluations. Note that these static scenes are included in the Longest6 training set but are not included in Town05.

TABLE IV: **Ablation for the attention block.** Note that self-attention is applied to a single modality, while cross-attention achieves multimodal fusion.

Attention Block		DS↑	RC↑	IS↑
No Attention		41.8	81.3	0.51
+ Self-Attention (No sensor fusion)	Image-Only	44.6	**94.4**	0.49
	Image&BEV	48.1	90.2	0.54
+ Cross-Attention (With sensor fusion)	No Pos. Embd	50.6	91.8	0.56
	Default Config	**56.4**	89.2	**0.65**

2) Attention block: The DualAT achieves camera-LiDAR fusion by leveraging the transformer attention mechanism. We separate the two types of attention and explore the benefits that they contribute to the overall system. The results are displayed in Table IV.

3) Auxiliary loss: We consider semantic segmentation and BEV map as auxiliary tasks in this work. Table V shows the impact of adding each auxiliary task on driving performance.

D. Visualization

Fig. 4 shows the attention maps in the sensor fusion module and the waypoint decoder, allowing for a visual observation of how the model jointly processes multisensor information and utilizes key perception information to make reasonable planning decisions. In Fig. 5, we visualize the prediction results of the auxiliary tasks conducted on the Longest6 and Town05 benchmarks, demonstrating that the model can jointly optimize multiple tasks and achieve robust performance.

TABLE V: **Ablation for the auxiliary loss.**

Auxiliary Losses	DS↑	RC↑	IS↑	Red↓
None	57.7	87.4	0.66	0.04
+ Semantics	59.3	89.6	0.66	0.02
+ BEV map	**62.7**	**90.7**	**0.69**	**0.02**

V. Conclusion

In this paper, we introduce a dual-attention transformer for end-to-end autonomous driving that is capable of efficiently processing multisensor information to perform multitask learning. Our method achieves state-of-the-art results on the Longest6 and Town05 benchmarks. Predicting the future motion of other agents is our future work, as this will enable the ego agent to establish safety constraints and make more interpretable end-to-end planning decisions. The DualAT has already demonstrated its ability to predict object occupancy levels using single-frame sensor inputs. With the use of historical frames, we believe our model can accurately estimate the future motion of all agents.

REFERENCES

[1] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[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," *arXiv preprint arXiv:2303.12077*, 2023.

[4] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 222–17 231.

[5] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision*, 2018, pp. 641–656.

[6] S. Fadadu, S. Pandey, D. Hegde, Y. Shi, F.-C. Chou, N. Djuric, and C. Vallespi-Gonzalez, "Multi-view fusion of sensor data for improved perception and prediction in autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2349–2357.

[7] Y. Zhou, P. Sun, Y. Zhang, D. Anguelov, J. Gao, T. Ouyang, J. Guo, J. Ngiam, and V. Vasudevan, "End-to-end multi-view fusion for 3d object detection in lidar point clouds," in *Conference on Robot Learning*. PMLR, 2020, pp. 923–932.

[8] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation," *IEEE Transactions on Multimedia*, 2023.

[9] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[12] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*. PMLR, 2023, pp. 726–737.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[14] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[15] 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.

[16] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020, pp. 66–75.

[17] 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.

[18] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le, A. Yuille, and M. Tan, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 17 182–17 191.

[19] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 1090–1099.

[20] K. Rho, J. Ha, and Y. Kim, "Guideformer: Transformers for image guided depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6250–6259.

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[22] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[24] 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.

[25] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[26] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamaki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2902–2911.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision*, 2018, pp. 3–19.

[28] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 793–15 803.

[29] J. Phillips, J. Martinez, I. A. Barsan, S. Casas, A. Sadat, and R. Urtasun, "Deep multi-task learning for joint localization, perception, and prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 4679–4689.

[30] L. Qu, S. Liu, M. Wang, and Z. Song, "Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2126–2134.

[31] 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.

[32] 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.

[33] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[34] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 590–15 599.

[35] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," in *Conference on Robot Learning*. PMLR, 2023, pp. 459–470.

[36] J. Zhang, Z. Huang, and E. Ohn-Bar, "Coaching a teachable student," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7805–7815.

[37] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 222–15 232.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[39] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.