Duo-View Pedestrian Behavior Prediction via Multi-modal Cross-Attentive Fusion

Zhengming Zhang
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
zhan3988@purdue.edu

Taotao Jing
Department of Computer Science
Tulane University
New Orleans, LA 70118
tjing@tulane.edu

Zhengming Ding
Department of Computer Science
Tulane University
New Orleans, LA 70118
zding1@tulane.edu

Renran Tian*
Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27607
rtian2@ncsu.edu

Abstract

Predicting pedestrian behavior is a crucial component in autonomous driving technology, fostering safer navigation and accident prevention for autonomous vehicles. Presently, research in pedestrian behavior modeling bifurcates into two distinct approaches: the egocentric view and the bird-eye view. Both perspectives offer unique advantages and drawbacks, yet there's a discernible absence of work integrating these two views. In this paper, we introduce a novel **M**ulti-modal **C**ross-**A**ttentive **F**usion algorithm (**MCAF**) that concurrently models trajectories from both perspectives, utilizing visual and spatial modalities in conjunction with interaction data and maps. We incorporate six different modalities from the two views (egocentric and bird-eye view), which include high-definition map (HD map), target and surrounding trajectories, egocentric image, egocentric trajectory, and ego-vehicle actions. Based on the nuScenes dataset, we construct a pedestrian trajectory dataset (nuScenes-DuoView) that encapsulates both views. Our findings indicate that this approach achieves superior performance to current methods, demonstrating an 8% and 12% improvement in Final Displacement Error (FDE) in the egocentric and bird-eye views, respectively. Additionally, the ablation study substantiates the benefits of fusing these two views.

1 Introduction

Predicting pedestrian behavior is critical for autonomous driving, particularly in dense urban environments where dynamic and uncertain interactions are common [26]. Unlike vehicles that generally

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI: Harnessing Artificial Intelligence for Smart Cities.

follow traffic rules, pedestrians exhibit spontaneous and intention-driven behaviors that are harder to anticipate. Visual cues, such as body posture and gaze, along with contextual information from the scene, are essential for accurate prediction [24, 5]. Two primary modeling perspectives exist: ego-centric and bird's-eye view (BEV). Ego-centric models, using on-board cameras and sensors, offer rich visual details of nearby pedestrians but struggle with depth estimation and occlusions. In contrast, BEV models—often created via LiDAR or fused sensors—offer a global scene layout useful for motion planning, but they lack fine-grained behavioral cues like gaze or subtle body motion [1]. Most existing approaches focus on one view and predict either trajectories or actions, limiting the richness and robustness of the output. Prior work also often treats prediction as a single-task problem, missing the opportunity to leverage shared representations across multiple outputs.

In this paper, we propose a multi-task learning framework that integrates ego-centric and BEV features to jointly predict pedestrian trajectories in both views and classify pedestrian actions. Using the nuScenes dataset [1], we build a multi-modal benchmark to support this framework. Our model outperforms existing methods, reducing final displacement errors by 8% and 12% in the ego and BEV views respectively. Extensive ablation studies also confirm the complementary benefits of combining views in a unified multi-task model.

2 RELATED WORK

Bird's-Eye View Pedestrian Trajectory Prediction: BEV trajectory prediction is traditionally studied on datasets like ETH [22], UCY [28], and Stanford Drone [16], which offer top-down views but limited visual detail. Generative models, particularly GAN-based [11, 29, 13], have been widely adopted. Others use RNNs for goal-conditioned prediction [31, 19, 32], or Transformers to capture spatio-temporal dynamics [10, 34]. Map fusion strategies vary, including rasterized HD maps [3, 37], vector/graph-based maps [36, 30], and grid-based classification [8, 6]. State-of-the-art performance has been achieved with transformer-based architectures[21].

Egocentric View Pedestrian Trajectory Prediction: Egocentric trajectory prediction focuses on datasets such as PIE [23], JAAD [14], PSI [4], and TITAN [18], which provide annotated trajectories and behavioral labels from a driver's perspective. Depth ambiguity and occlusions remain key challenges in this view. Multimodal prediction is increasingly popular, using cues from appearance, pose, and context [27, 25]. RNNs have traditionally dominated [23, 32], but Transformers are now preferred for better temporal modeling and modality fusion [33]. Incorporating priors like reachability maps [17] or intention estimates [9] further improves performance. Multi-task learning, combining trajectory prediction with action or intention labels, has also proven beneficial [25].

3 METHOD

3.1 Construction of the nuScenes-DuoView Dataset

To address the lack of dual-perspective pedestrian data, this study introduces nuScenes-DuoView, a new dataset that combines egocentric and bird's-eye views for pedestrian trajectory prediction. Built upon the nuScenes dataset [1], it integrates synchronized camera, LiDAR, and HD map data, capturing diverse pedestrian behaviors across multiple sensor modalities. From the nuScenes data, we extracted all pedestrian instances and filtered those with limited visibility—removing those further than 50 meters away, shorter than 3 seconds in view, or heavily occluded. This refinement resulted in 4,918 high-quality pedestrian trajectories out of 11,187 original instances. We then projected LiDAR data to top-down views and aligned them with egocentric frames at 10 Hz via interpolation. Each instance includes surrounding agent context, HD map patches, and binary action labels (walking or standing), enabling robust multi-view learning and social context modeling. More details of the dataset are discussed in Appendix A.

3.2 Problem Formulation

We model pedestrian trajectory prediction as a supervised time-series forecasting problem across two views: *egocentric* and *bird's-eye*. The objective is to learn a mapping from historical multimodal observations to future pedestrian trajectories and actions in both views.

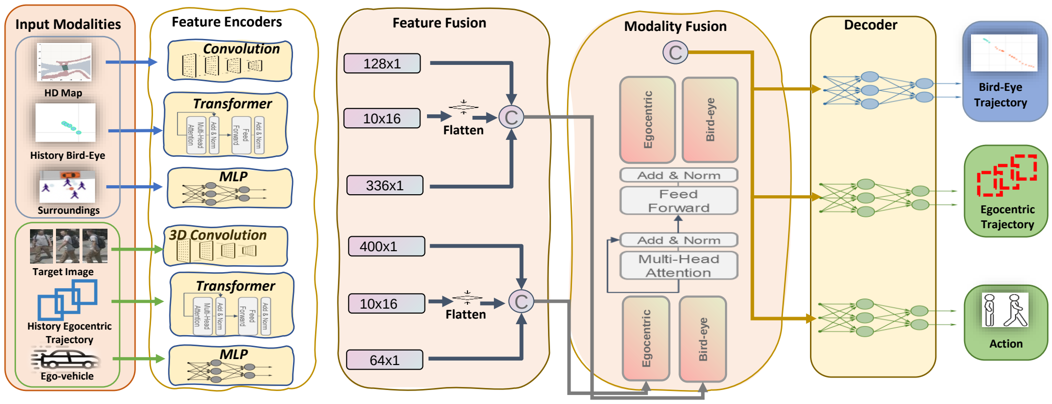

Figure 1: Illustration of our detailed model architecture.

Let $\mathbf{x}_{0:t} = \left(\mathbf{b}_{0:t}, \{\mathbf{e}^i_{0:t}\}_{i=1}^K, \{\mathbf{I}^i_{0:t}\}_{i=1}^K, \mathbf{m}, \mathbf{v}_{0:t}, \mathbf{o}_{0:t}\right)$ denote the historical observations up to time t, where each input is defined as follows. $\mathbf{b}_{0:t}$: bird's-eye view pedestrian positions in global map coordinates; $\mathbf{e}^i_{0:t}$: pedestrian bounding boxes from the i-th egocentric camera; $\mathbf{I}^i_{0:t}$: image stream from the i-th egocentric camera (synchronized with $\mathbf{e}^i_{0:t}$); $\mathbf{m}$: rasterized high-definition map centered on the scene; $\mathbf{v}_{0:t}$: ego-vehicle states (position, velocity, acceleration); and $\mathbf{o}_{0:t}$: bird's-eye view trajectories of nearby traffic participants.

The target prediction is defined as:

$$\mathbf{y}_{t+1:t+\tau} = \left(\mathbf{b}_{t+1:t+\tau}, \{\mathbf{e}^i_{t+1:t+\tau}\}_{i=1}^K, \mathbf{a}_{t+\tau}\right),$$

where $\mathbf{a}_{t+\tau}$ denotes the pedestrian action label at the final prediction step.

Formally, the learning task is to seek a function f:

$$\mathsf{f} : \mathbf{x}_{0:t} \to \mathbf{y}_{t+1:t+\tau},$$

leveraging both global scene context from the bird's-eye view and fine-grained cues from multi-camera egocentric observations.

3.3 Model Architecture

Our model jointly predicts pedestrian trajectories in both bird's-eye and egocentric views, along with pedestrian actions, in a single run (Figure 1). It consists of three main modules: **feature encoding**, **multi-modal fusion**, and **multi-task prediction**. The **feature encoding** is an internal modality fusion, which involves simple concatenation, and the **multi-modal fusion** is conducted through a transformer encoder to effectively integrate information from different sources and making accurate predictions. Such design can effectively learn how to combine the different sources of information to make accurate predictions.

Feature Encoding. We encode six modalities: (1) bird's-eye trajectories and (2) egocentric trajectories (from K cameras), both using transformers to capture temporal dependencies; (3) rasterized HD maps via a CNN; (4) pedestrian-centered egocentric video clips using a pre-trained 3D CNN [2]; (5) ego-vehicle states and (6) nearby object trajectories via MLPs. Each egocentric trajectory $\mathbf{e}^i_{0:t}$ and image stream $\mathbf{I}^i_{0:t}$ is tied to its camera index i for precise multi-camera modeling.

Multi-Modal Fusion. We adopt a two-level fusion strategy. First, feature-level fusion aggregates related modalities into two embeddings: *bird's-eye* (pedestrian trajectory, nearby objects, map) and *egocentric* (image, pedestrian trajectory, ego state). Second, a transformer encoder applies cross-attention between these two embeddings to integrate global context and fine-grained cues while keeping computation efficient. Instead of stacking several layers of transformers to consider multiple modalities simultaneously, which significantly increases model complexity, we use the feature-wise fusion technique before the transformer towards a more efficient and stable mechanism.

Multi-Task Prediction. Separate MLP heads predict (1) bird's-eye trajectories, (2) egocentric trajectories, and (3) pedestrian actions. We use uncertainty-based adaptive loss weighting [12] to balance tasks, allowing the model to prioritize more confident predictions while still learning from all objectives. We defined the weight w_i for the loss of task i as $w_i = \frac{1}{\sigma_i^2}$, where σ_i is the standard deviation of the predicted probability distribution for task i.

4 Evaluation

4.1 Implementation

Both bird's-eye and egocentric trajectory encoders use two-layer transformers (2 heads, 128-d embeddings). The HD map encoder is a two-layer CNN (5×5 kernels, 128 channels). Egocentric pedestrian videos are processed by a pre-trained I3D-ResNet50, outputting 400-class probability features. Ego-vehicle actions and surrounding object trajectories are encoded with two-layer MLPs (256, 128). All activations are ReLU. Fusion is performed with a single transformer layer (4 heads, 256-d embeddings), followed by task-specific MLP decoders.

4.2 Evaluation Protocol and Metrics

The nuScenes-DuoView dataset is split 70%/10%/20% for train/val/test from 800 scenarios, with each pedestrian sequence segmented into 3-second samples (1s observation, 2s prediction), yielding 7,267/693/1,406 samples. With a 10 Hz annotation rate, each sample has 10 observation and 20 prediction frames. Performance is measured using Average Displacement Error (ADE) and Final Displacement Error (FDE) [25, 4]. For egocentric view, we report ADE/FDE for both bounding box centers and corners. Action prediction is evaluated with precision, recall, F1, and accuracy [15].

4.3 Comparisons with State-of-the-Art Methods

To ensure a fair comparison of existing methods on the nuScenes-DuoView dataset, we implemented an early stop criterion monitoring center FDE with a tolerance of 20 epochs. Additionally, we utilized a learning rate of 1e-3 with a cosine learning rate schedule (decay to 1e-5 in 100 epochs). These standardized settings allowed us to evaluate each model's performance consistently and effectively.

Model\Metric	ADE ↓	FDE ↓
MTP [7]	2.98	5.59
PCENet [20]	1.15	2.08
TF [10]	0.82	1.47
P2T [6]	1.24	2.29
NSP [35]	0.81	1.39
Ours	**0.77**	**1.36**

Table 1: Comparative Analysis of Models in Bird's-eye View.

We replicated the performance of five recent models for predicting pedestrian trajectories from a bird's-eye view. These models include NSP [35], PCEnet [20], Transformer TF [10], MTP [7], and P2T [6]. We reproduced the results of these models on the nuScenes-DuoView dataset with the original implementation of each model and selected the smallest model size due to the dataset's limited size. The results in Table 1 indicate that our model surpasses the current method in bird's-eye view pedestrian trajectory prediction. Despite the performance, our model stands out due to its simplicity and straightforwardness when compared to the others, which incorporates numerous hand-crafted designs. We hypothesize that the performance enhancement is attributed to the egocentric view, which furnishes additional spatial information through visuals and different angles.

Model	ADE	FDE	ADE$_{bbox}$	FDE$_{bbox}$
PIE [23]	48	59	89	105
Bitrap [32]	49	57	92	112
SGNet [31]	46	55	89	102
Ours	**28**	**41**	**61**	**86**

Table 2: Comparative Analysis of Models in Ego View.

Similarly, we replicated the performance of three models from an egocentric view, including PIE [23], SGNet [31], and Bitrap [32]. The model performances are summarized in Table 2. Our model excels above the rest, outperforming others by approximately 40% in ADE. Except our model, no existing methods achieved an ADE below 30. This superior performance may be from the guiding role of the bird's-eye view. The egocentric view presents a more complex task as it incorporates the movements of both the ego vehicle and the target pedestrian. Conversely, the bird's-eye view is independent of the ego vehicle's movement. Such information assists the model in distinguishing the pedestrian's kinetic information from the ego vehicle's movement.

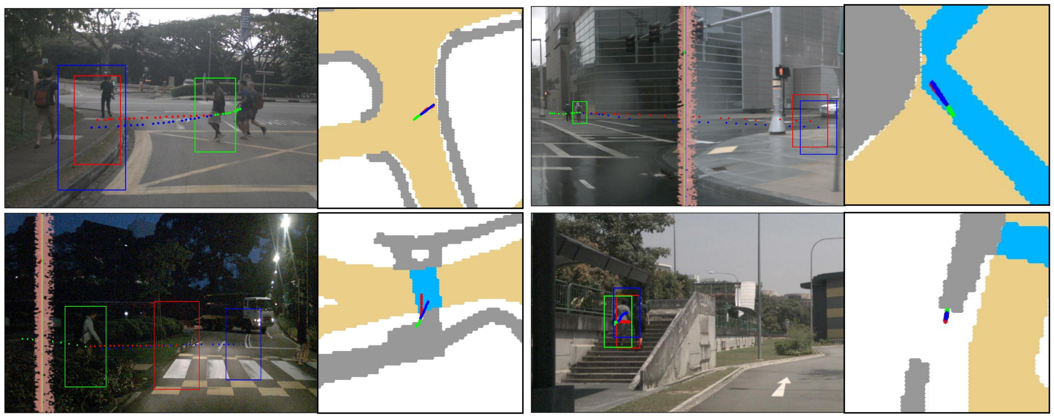

Figure 2: For the egocentric view (left), the target pedestrian is highlighted within a green bounding box, with their actual trajectory represented by green dots. The prediction is indicated by red dots, while the ground-truth is marked by blue dots. The final locations of the predicted and actual bounding boxes are represented by red and blue boxes, respectively. In the bird-eye view on the HD map (right), the original ground truth input is shown as a green line, the model's prediction is a red line, and the actual path taken is represented by a blue line. The map's features are color-coded: sidewalks are gray, crosswalks are blue, and drivable areas are yellow. The vertical dash line separates cameras.

5 Case Study

In this section, we assessed the performance of our proposed model by utilizing representative examples that adhere to the format displayed in Figure 2. Every demonstration is divided into two parts: the egocentric view to the left and the bird-eye view to the right. The images have been cropped to enhance visual clarity. The vertical blue dashed line serves as a marker for the boundary between different cameras. For the egocentric view, the target pedestrian is highlighted within a green bounding box, with their actual trajectory represented by green dots. The model's predicted path is indicated by red dots, while the true trajectory is marked by blue dots. The final locations of the predicted and actual bounding boxes are represented by red and blue boxes, respectively. In the bird-eye view on the HD map, the original ground truth input is shown as a green line, the model's prediction is a red line, and the actual path taken is represented by a blue line. The map's features are color-coded: sidewalks are gray, crosswalks are blue, and drivable areas are yellow.

The model demonstrated good performance in the top-row cases. In top-left, the model anticipated that the pedestrian was about to cross with consistency between the egocentric and bird-eye views. Similarly, for the top-right case, the model correctly predicted the pedestrian's crossing on the crosswalk. Some scenarios where model's performance is not accurate are shown in the bottom row. In bottom-left, while the model expected the pedestrian to cross, it underestimated walking speed in the egocentric view and predicted an inconsistent trajectory change in the bird-eye view. In the bottom-right situation, the model inaccurately predicted the pedestrian's trajectory in the egocentric view, possibly due to insufficient data on the pedestrian climbing the floor.

6 Conclusion

In conclusion, our research has established the significance and effectiveness of a holistic, multimodal approach in predicting pedestrian behavior. By integrating the unique advantages of both the egocentric and bird-eye view perspectives, we have managed to generate more accurate trajectory predictions. The integration of six different modalities from both views further enriches the predictive capabilities of our model. Through the utilization of the adapted nuScenes-DuoView dataset, we have been able to validate our approach, with results indicating considerable improvements.

7 Acknowledgment

This work was supported by the National Science Foundation [Grant No.2516587].

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning*, pages 86–99. PMLR, 2020.

[4] Tina Chen, Renran Tian, Yaobin Chen, Joshua Domeyer, Heishiro Toyoda, Rini Sherony, Taotao Jing, and Zhengming Ding. Psi: A pedestrian behavior dataset for socially intelligent autonomous car. *arXiv preprint arXiv:2112.02604*, 2021.

[5] Tina Chen, Renran Tian, and Zhengming Ding. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3103–3109, 2021.

[6] Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020.

[7] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020.

[8] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Home: Heatmap output for future motion estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 500–507. IEEE, 2021.

[9] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[10] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, pages 10335–10342. IEEE, 2021.

[11] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018.

[12] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[13] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016.

[15] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1258–1268, 2021.

[16] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[17] Osama Makansi, Ozgun Cicek, Kevin Buchicchio, and Thomas Brox. Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4354–4363, 2020.

[18] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020.

[19] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

[20] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 759–776. Springer, 2020.

[21] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*, 2022.

[22] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 452–465. Springer, 2010.

[23] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[24] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, 3(1):61–70, 2017.

[25] Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021.

[26] Amir Rasouli and John K Tsotsos. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE transactions on intelligent transportation systems*, 21(3):900–918, 2019.

[27] Amir Rasouli, Tiffany Yau, Mohsen Rohani, and Jun Luo. Multi-modal hybrid architecture for pedestrian action prediction. *arXiv preprint arXiv:2012.00514*, 2020.

[28] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer, 2016.

[29] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019.

[30] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

[31] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):2716–2723, 2022.

[32] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters*, 6(2):1463–1470, 2021.

[33] Ziyi Yin, Ruijin Liu, Zhiliang Xiong, and Zejian Yuan. Multimodal transformer networks for pedestrian trajectory prediction. In *IJCAI*, pages 1259–1265, 2021.

[34] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

[35] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 376–394. Springer, 2022.

[36] Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 532–539. IEEE, 2021.

[37] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904. PMLR, 2021.

A Appendix A: The Novel nuScenes-DuoView Dataset

To bridge the gap in existing research, we deliver a novel dataset that combines egocentric and bird-eye views for pedestrian trajectory prediction. This dataset is based on the nuScenes autonomous driving perception dataset, which provides a wide range of sensor data, such as LiDAR, six cameras, and radar, along with HD maps containing semantic information, for 1,000 diverse traffic scenarios. Our dataset includes synchronized images and annotated pedestrian trajectories from both viewpoints, facilitating the exploration of fusion techniques that integrate information from multiple angles. With this resource, our goal is to advance pedestrian trajectory prediction research by harnessing the advantages of combining diverse perspectives.

To create our dataset, named 'nuScenes-DuoView,' we extracted all pedestrian instances from the LiDAR data in the nuScenes dataset and recorded the corresponding ego-vehicle actions. As the dataset was not originally collected for pedestrian trajectory prediction, we further refined it by applying filters. Initially, we excluded pedestrian instances that remained within the frontal three cameras of the ego vehicle for less than three seconds and were more than 50 meters away throughout their trajectories. Subsequently, we manually removed pedestrians that were occluded and not visible in the images, as the pedestrian selection primarily relied on LiDAR data, which offers a higher position and a longer range. In total, the nuScenes-DuoView dataset comprises 4,918 pedestrians out of the original 11,187 instances.

Next, we performed a LiDAR data projection onto a bird's-eye view and extracted the corresponding top-down image. The nuScenes dataset offers synchronized annotations at a rate of 2 Hz, alongside raw sensor inputs at 10 Hz. To match the higher sampling rate, we applied linear interpolation to the annotations. In addition to the bird's-eye view and egocentric images, we captured a high-definition (HD) map centered on the target pedestrian's location. Furthermore, we incorporated the trajectories of the ten nearest vehicles or pedestrians to provide social contextual information. Each pedestrian received an additional binary action label, indicating whether they were walking or standing.

Model	Task			Extra Modalities			
	Ego	Bird-I	Act	Img	HD Map	Surding	Ego
NSP		✓			✓		
PCENet		✓					
TF		✓				✓	
MTP		✓					
P2T		✓					
PIE	✓		✓	✓			✓
SGNet	✓						
Bitrap	✓						
Ours	✓	✓	✓	✓	✓	✓	✓

Table 1: Model comparison on the nuScenes-DuoView dataset.

B Appendix B: Ablation Study

We list the results of the ablation study in Table 2. The first columns indicate the model components, each component is an input modality.

From our analysis, it's apparent that when a single modality is used (either bird's-eye or egocentric), the performance is comparable to or even falls short of existing methods. However, when these two modalities are integrated, there's a substantial performance boost of 10% for bird's-eye view and 30% for the egocentric view. Visual features and trajectories of surrounding objects seem to contribute marginally to both views. Yet, HD maps play significant roles in improving predictions in the bird-eye view. Incorporating the ego-vehicle's actions proves to be highly beneficial for predictions in the egocentric view. This suggests that the appropriate selection and integration of modalities and features can greatly enhance the model's capabilities. Furthermore, it highlights the importance of considering the task's specific characteristics and requirements when choosing features.

Model						Bird-eye		Egocentric				Action	
Ego	Bird	Img	HD	EgoVeh	Surround	ADE	FDE	ADE	FDE	ADE	FDE	Acc	F1
✓						-	-	43	56	78	95	0.65	0.71
	✓					0.85	1.53	-	-	-	-	0.82	0.90
✓	✓					0.78	1.39	31	46	70	85	**0.87**	*0.91*
✓	✓	✓				*0.77*	1.37	32	45	72	88	0.86	0.89
✓	✓	✓	✓			**0.71**	**1.32**	30	44	66	89	0.85	0.87
✓	✓	✓	✓	✓		0.83	1.49	**28**	**41**	**61**	**86**	**0.87**	**0.92**
✓	✓	✓	✓	✓	✓	*0.77*	*1.36*	*29*	*43*	*63*	*87*	**0.87**	*0.91*

Table 2: Ablation Study of Our Model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we clearly describe that the paper contributes a new dataset and a new algorithm to address the gap of simultaneously predicting pedestrian behaviors from both ego-centric and bird' eye perspectives. The proposed method can booster behavior prediction performance in both perspectives compared to the SOTA methods in individual views.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Due to page limit, we don't include a limitation section in the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We don't make any theoretic contributions in the paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details necessary for reproducing the experimental results are fully described in either the main paper or appendix. The code will be released upon acceptance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In appendix, we describe the used dataset as well as the version of publicly available models we rely on in details.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please refer to the implementation details in appeix for all the details regarding hyperparameter choice, preprocessing, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We don't do statistical significance check simply because our model outperforms existing models by a large margin that don't need a statistical test to prove it.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We list the computing resources needed for the proposed method, whose parameters are publicly available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We closely follow the code of ethics by NeurIPS.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We are solving applications in autonomous driving scenario, which is highly impacted.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: we don't think the release of our data or models will have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data and the pretrained deep models are fully described. We give credits to the original creator in the main paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release our code upon acceptance, which include new assets such as trained model weights.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The research relies on public benchmark dataset and does not involve any research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The research relies on public benchmark dataset and does not involve any research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is not a part of the algorithm or involved as any important, original, or non-standard components in the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.