Where Will They Go? Modelling Multimodal Pedestrian Manoeuvres from Ego-centric Videos

Yuxuan Xie[1], Nicolas Pugeault[1], Chongfeng Wei[2], Hubert P. H. Shum[3], and Edmond S. L. Ho[1*]

***Abstract*— Pedestrian trajectory prediction from an ego-centric camera is challenging since it depends on complex interactions with vehicles and scene context, as well as the intention of the pedestrian. By modelling correlation and intent from the historical and future trajectories of the pedestrian, it will usually result in a multimodal (i.e. multiple modes) distribution. Existing stochastic predictors often sample multiple futures from a single unimodal distribution, which can yield sub-optimal 'mixed-mode' trajectories that lie between distinct motion patterns and become implausible in real scenes. In this paper, we propose MMPM, a mode-aware framework that separately models future trajectory distributions into semantically meaningful modes based on the pedestrian's crossing behavior. MMPM consists of two modules: behavior-aware Pedestrian Interaction Module (PIM) that jointly captures pedestrian–vehicle and pedestrian–environment interactions by introducing gaze, head and hand gesture, and a CVAE-based Mode-aware Trajectory Predictor (MTP) module to model the future trajectory distributions on two modes, crossing and non-crossing the road, separately. A query-based decoder further enforces mode consistency during decoding. Experiments on PIE and JAAD datasets show that our method surpasses state-of-the-art baselines. Our proposed MTP is model-agnostic, which can be integrated into existing frameworks such as BiTrap-NP and SGNet-ED to further improve future trajectory prediction performance. We additionally introduce a data-driven validation protocol that matches predictions to spatio-temporally consistent ground-truth trajectories, demonstrating improved frame-wise displacement errors over previous work.**

I. INTRODUCTION

Pedestrian trajectory prediction is an essential aspect of tasks such as autonomous driving. In recent years, this problem has become an active research area and has achieved considerable progress by learning key factors that influence pedestrian motion, including interactions among agents [1], the impact of surrounding context [2], [3], and goal-driven behaviors [4], [5].

To model the dynamics of the interactions between the pedestrian, other road users and the environment, the historical features of agents are typically aligned and processed using attention [6], [7], [8], [9] or recurrent models [10], [11] to capture the spatio-temporal information. Visual inputs, e.g. scene-level images [12] or semantic segmentation [13], [1], are commonly used as spatial clues for spatio-temporal modeling. Goal-driven approaches [14], [4], [5] guide the decoding process by predicting pedestrians' final goals. However, the task still remains challenging due to the intrinsic complexity in modeling the 'one-to-many' mapping between the historical observations and feasible future trajectories. On the other hand, encouraging results have been reported in modeling multimodality for vehicle trajectory prediction [15], [16].

To the best of our knowledge, none of the aforementioned approaches explicitly model the multimodality of pedestrian trajectories in ego-centric datasets: *Given the same or similar historical trajectories, a pedestrian may have multiple possible future trajectories corresponding to distinct goals, with distinct motion patterns.* Existing stochastic prediction methods [5], [4], [17], [18] attempt to generate multiple future trajectories by sampling from a single Gaussian distribution via generative models. However, this stochastic prediction paradigm may not effectively capture different motion patterns [15], since using a single Gaussian to model the underlying multimodal data distribution is an ill-posed problem.

In this paper, we propose a model, namely Modelling Multimodal Pedestrian Manoeuvres (MMPM), that models the distribution of the future pedestrian trajectories in each *mode* individually. Our framework consists of two interacting components: a Pedestrian Interaction Module (PIM) and a conditional variational autoencoder (CVAE)-based Mode-aware Trajectory Predictor (MTP). Specifically, PIM aims to learn a representation based on the interactions among the pedestrian, other road users, the autonomous vehicle and the environment. We further enhance the representation by incorporating the gesture and gaze of the pedestrian as behavioral features, which have not been explored in prior pedestrian trajectory prediction approaches, to provide additional cues for our model to predict the pedestrian's intention. Next, during the training stage of MTP, we treat the ground truth action labels, i.e. *crossing* or *non-crossing* in this study, as the modes and we model the distribution of future trajectories under each mode separately. At inference, MTP starts with predicting the **mode** based on the output of PIM, then samples future trajectories from the learned prior for each mode according to the ratio of predicted crossing probability.

Experimental results show that the proposed method surpasses the state-of-the-art methods [19], [20] on both the benchmark PIE and JAAD datasets. Our proposed MTP is model-agnostic, which can be integrated into existing frameworks such as BiTrap-NP [4] and SGNet-ED [5] to further improve future trajectory prediction performance.

*Corresponding author, email: Shu-Lim.Ho@glasgow.ac.uk

[1]School of Computing Science, University of Glasgow, Glasgow, United Kingdom

[2]James Watt School of Engineering, University of Glasgow, Glasgow, United Kingdom

[3]Department of Computer Science, Durham University, Durham, United Kingdom

Furthermore, to further evaluate the effectiveness of modeling different modes separately, we design a data-driven validation protocol to compare the predicted future trajectories with real-world data from the dataset spatio-temporally, and we show that our method results in up to a 4.73 % decrease in frame-wise displacement error over previous work.

We summarize our contributions as follows:

1) We propose the Pedestrian Interaction Module (PIM) which leverages pedestrians' behavior, through gesture and gaze, to learn a behavior-aware interaction representation between pedestrians and their surroundings.
2) We propose Mode-aware Trajectory Predictor (MTP) to model the multimodal future trajectory distributions separately based on pedestrians' crossing behavior.
3) We introduce a novel data-driven validation protocol to assess the validity of multimodal trajectory prediction.

II. Related work

A. Pedestrian Trajectory Prediction

Bird's-eye view. Predominant approaches involve predicting trajectories from bird's-eye-view videos captured by surveillance cameras or drones. Prior research has focused on multi-agent scenarios, learning diverse trajectories, goal, or group behavior to model the interaction between agents [8], [21], [17], [22], [23], [18], [24]. Another stream of work [25], [26] is based on modeling vehicle-pedestrian interactions based on Reinforcement Learning and simulation. These approaches typically model interactions based on pedestrians' positions and their relative movement within the observation horizon, using distance-based cues. However, pedestrians' non-verbal communications with others, such as gestures and nodding, can strongly indicate their intended behaviors and trajectory pattern. Such approaches cannot effectively leverage these visual cues to predict future trajectories due to viewpoint limitations.

Ego-centric view. Ego-centric methods rely on data from an on-board moving camera. Previous methods use multi-stream inputs [7], [9], [13], [27], [20], [28], [29], while some works utilize historical trajectories and goal guidance [4], [5] to achieve notable stochastic results. BR-GAN [30] uses YoloV3 on bird-eye-view video clips to identify pedestrian behavior as walking, grouping and standing then combine it with other temporal features in a late-fusion paradigm. TAGRN-SAR [19] utilizes action intents into autoregressive trajectory decoding. Considering actions reflect the dynamics of pedestrians, the actions combined with Gaussian noise are aggregated with encoder output to enhance the model's awareness of future dynamics. Compared to aforementioned methods, we not only introduce gaze and gesture, which are highly related to interaction instead of using dynamic-related actions, but also utilize crossing actions as pedestrians' modes, enhancing the model's ability to perceive and model the multimodality of future trajectory.

Multimodality in trajectory prediction Mainstream methods [10], [18], [17], [4], [5] adopt generative models, such as CVAE or diffusion architecture for stochastic prediction, and incorporate losses for encouraging the model to sample diverse trajectories with a single distribution. Although these methods achieve state-of-the-art performance, we consider that a unimodal distribution is insufficient to capture the multimodality of pedestrian motion properly. Inspired by recent advancements in multimodal vehicle trajectory prediction [31], [16], [15], we propose to explicitly define and separate multiple mode distributions, enabling the model to better understand the multimodal future of pedestrian. We further validate through extensive experiments that our approach mitigates the above issue.

III. Methodology

A. Problem Formulation

We formulate trajectory prediction as a sequence generation problem which generates the future trajectory of an agent based on the observed history. For a target agent *tar*, given its historical trajectories $\mathbf{X}_{\text{tar}} = \{\mathbf{X}^{t-T_h+1}, \ldots, \mathbf{X}^{t-1}, \mathbf{X}^t\}$ over T_h frames, we aim to predict its future trajectory $\mathbf{Y} = \{\mathbf{Y}^{t+1}, \mathbf{Y}^{t+2}, \ldots, \mathbf{Y}_{\text{tar}}^{t+T_f}\}$ over T_f future frames, where X^t are the input features at t-th time step.

To model the interaction between *tar* and the surroundings, multiple modalities are considered in this work:

1) the pedestrian's normalized bounding-box coordinates $\mathbf{p}_{\text{obs}}^t$,
2) the pedestrian's displacement relative to the last observed time step t, $\mathbf{x}_{\text{obs}}^t$,
3) the pedestrian's velocity $\mathbf{v}_{\text{obs}}^t$, defined as the displacement of the pedestrian with respect to the previous time step.

The above features are computed on the image plane from the top-left and bottom-right corners of the pedestrian bounding box to capture the dynamic pedestrian motion.

In addition, we use behavioral features of pedestrian which have not been explored in the previous work:

4) *gaze* and *gestures* $\mathbf{b}_{\text{obs}}^{\text{t}} = (\text{g}_{\text{obs}}^{\text{t}}, \text{l}_{\text{obs}}^{\text{t}})$, depending on availability in the dataset.

Furthermore, we include

5) the ego-vehicle motion $\mathbf{V}\text{ego}^{\mathbf{t}} = (v_{ego}^t, v_{ego,x}^t, v_{ego,z}^t)$, i.e., the vehicle GPS speed and its decomposed velocities along the x- and z-axes, and
6) the context $\mathbf{c}$ at the last observation, represented by the observed semantic segmentation map classified into four categories, pedestrians, motorcyclists/bicyclists, vehicles, and static context.

B. Modelling Multimodal Pedestrian Manoeuvres (MMPM)

Figure 1 illustrates our proposed mode-aware framework for pedestrian trajectory prediction, MMPM, where we incorporate new behavioral features into interaction modeling in the Pedestrian Interaction Module (PIM) (Section III-C) and separately model future trajectory distributions under different pedestrian modes in the Mode-aware Trajectory Predictor (MTP) (Section III-D).

Next, our Mode-aware Trajectory Predictor generates two trajectory modalities, by dividing the output of PIM into two

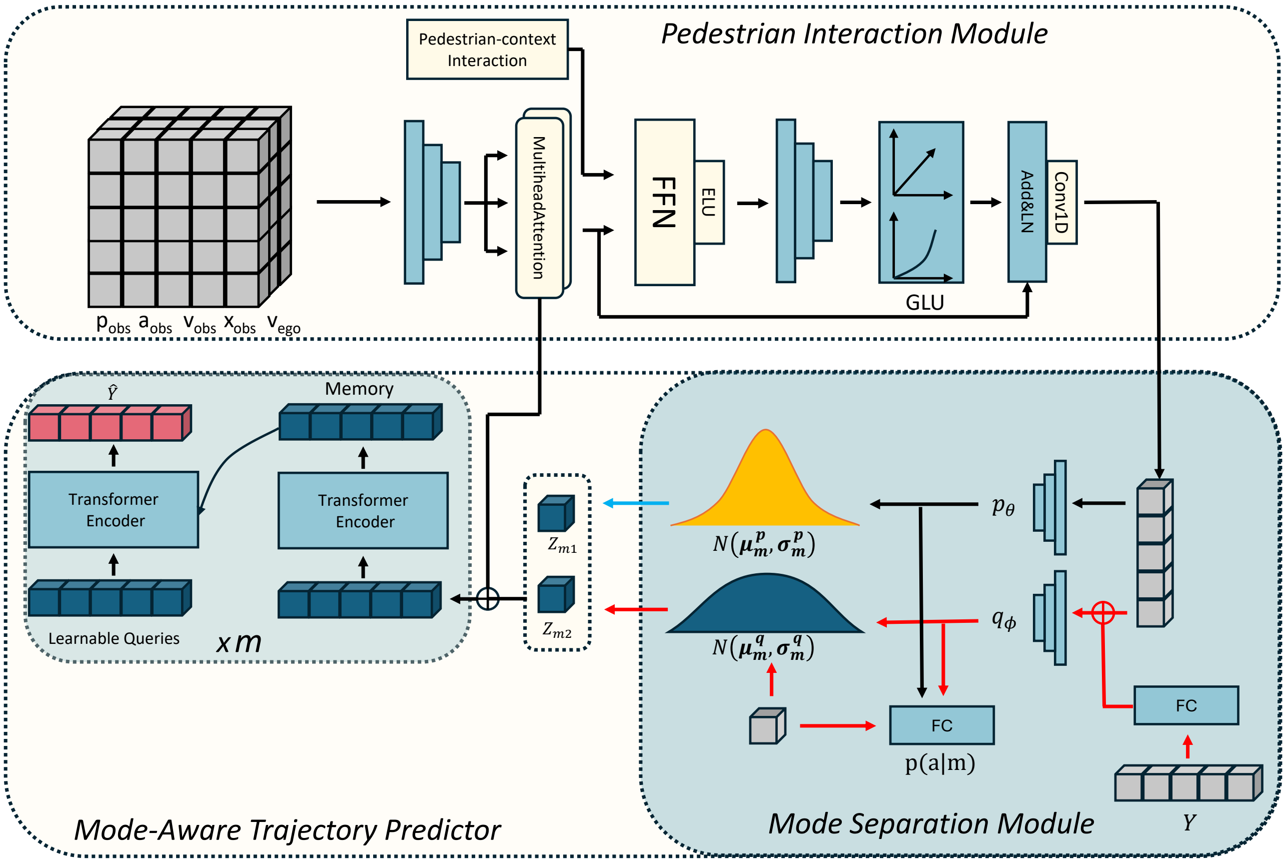

Fig. 1. Our proposed framework: Pedestrian Interaction Module, Mode-aware Trajectory Prediction module and m decoders for each separated mode. red line denotes training-only while blue line denotes test-only.

distinct distributions, representing crossing and non-crossing actions of pedestrians. A query-based decoder then samples a total of k samples from both distributions, weighted by the probability of predicted crossing, and decodes the final trajectory prediction.

C. Pedestrian Interaction Module (PIM)

We design PIM to capture interaction among pedestrian, ego-vehicle and context, by integrating new gesture and gaze features into other motion cues. This enables MMPM to be aware of pedestrian's intention in addition to learning the interaction dynamics. It is suggested in the literature [32], [10] that the interaction relies on low-level motion, such as direction, speed, relative position, etc. However, we argue that interaction is influenced not only by such motion cues but also by behavior that reveal the pedestrian's future intention. For instance pedestrian behavior such as nodding or hand gestures demonstrates the communication with other agents and hint the intention. To model the influence of behavior, PIM is an encoder based on attention which consists of two components: a pedestrian-vehicle interaction block, and a pedestrian-context interaction block.

Motivated by [13], the pedestrian-context interaction block aims to aggregate scene information based on the pedestrian historical trajectory and and ego-vehicle motion. We split the resized semantic map into n patches of size $p_z \times p_z$. After applying multi-head attention over the patch tokens, we feed the attended context tokens together with the last hidden state of scaled pedestrian trajectories $\mathbf{x}^i_{\text{obs}}$, and $\mathbf{v}^{\mathbf{i}}_{\text{ego}}$ into a global attention unit to obtain spatial interaction features for the target pedestrian, namely context vector c:

$$\begin{aligned} A &= \text{MultiHead}(e) + e \\ E &= \text{MultiHead}(S_m) + S_m \\ g &= \text{softmax}(h_t \odot \text{W}_1 E) \\ C &= \tanh(\text{W}_c [g \oplus h_t]) \end{aligned} \tag{1}$$

where e denotes the output of embedding layers, W denote a Feed Forward Network (FFN), S_m represents embedding of the segmentation patches, h_t is the final hidden state of GRU, and $\oplus$ refers to concatenation.

In the pedestrian-context interaction block (Figure 2), ego-vehicle and pedestrian motion embeddings are concatenated first then pass through a multi-head attention layer. The embeddings are further refined using a GRN [33] network to better learn the spatial-temporal dependencies. This is comparable to [19], but we improve it by leveraging C to condition the embeddings. The context vector determines the extent to which information in the embeddings is transformed, allowing the model to control the temporal feature update according to the pedestrian-context interaction, rather than only relying on temporal features for selection. We then employ a global pooling layer, implemented as 1D convolution, to encode the multimodal latent representation:

$$\begin{aligned} I &= \mathrm{Conv1D}\big(\mathrm{LayerNorm}\big(A+\mathrm{GLU}(n_1)\big)\big), \\ \eta_1 &= \mathrm{MLP}(\eta_2), \\ \eta_2 &= \mathrm{ELU}\big(\mathrm{FFN}(A)+\mathrm{FFN}(c)\big). \end{aligned} \tag{2}$$

where MLP is a three-layer Multilayer Perceptron (MLP), ELU refers to Exponential Linear Unit activation function [34], and GLU refers Gated Linear Units [35].

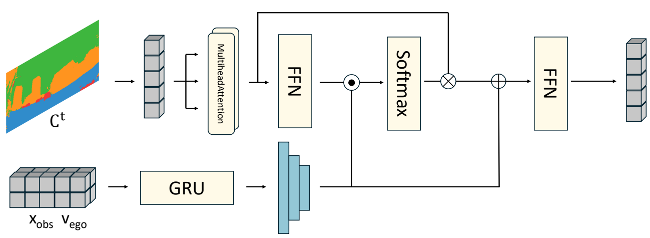

Fig. 2. The architecture of Pedestrian-context interaction component.

D. Mode-aware Trajectory Predictor

1) Mode separation.: Considering the future trajectories vary, and previous works simply predict future trajectory with a single distribution, which are insufficient to fully model multimodality of trajectory, we define the crossing action of pedestrian, i.e, crossing or non-crossing as modes M, and propose a Mode-aware Trajectory Predictor that generates future trajectory distributions $p(M \mid \mathbf{X}), M = \{m_i \mid i = 1,2\}$ with probabilities $p(a \mid m_i)$, where the two modes correspond to *crossing* and *non-crossing* respectively. The overall trajectory distributions are represented as:

$$p(\mathbf{Y} \mid \mathbf{X}) = \sum_{m=1}^{M} p(m_i \mid \mathbf{X})\, p(\mathbf{Y} \mid m_i, \mathbf{X})\, p(a \mid m_i). \tag{3}$$

We exploit a CVAE framework to learn multimodal future trajectories. Specifically, we employ a recognition network $q_\phi(M \mid \mathbf{X}, \mathbf{Y})$, a prior network $p_\theta(M \mid \mathbf{X})$, where ϕ and θ denote the parameters of these networks.

The ground-truth future trajectory $\mathbf{Y}$ is provided to the recognition network together with $\mathbf{X}$ in the training phase. In details, the embedding emb_y are concatenated with I and fed into an MLP, which outputs the parameters of Gaussian distributions $N(\mu_m^q, \sigma_m^q)$ for each mode $m \in M$, while the prior network only takes e as input and outputs $N(\mu_m^p, \sigma_m^p)$. We sample z_{mi} from the recognition distributions and decode future trajectories conditioned on $\mathbf{X}$.

At inference time, $\mathbf{Y}$ is unavailable. Therefore, the decoder samples z_{mi} from each mode's distribution $N(\mu_m^p, \sigma_m^p)$ from prior network p_θ, and then decodes the corresponding future trajectories $\hat{Y}^{Tf}$.

2) Decoder.: To alleviate error accumulation and ensure the mode consistency during decoding, we employ M query-style decoder [36] for samples from each mode instead of a shared weight decoder. Before entering the Transformer encoder, $\mathbf{X}$ as condition and z_{mi} are concatenated, and the decoder then uses a fixed length of learnable queries to attend to the encoder outputs.

3) Auxiliary crossing prediction and probability-guided sampling.: To explicitly differentiate modes, we assume that different modes should reflect whether a pedestrian will cross the road. We therefore introduce an auxiliary action prediction task. Specifically, we use $q_\phi(M \mid \mathbf{X}, \mathbf{Y})$ or $p_\theta(M \mid \mathbf{X})$ as input to an MLP to predict the crossing probability for weighting the sampling process.

During the training, regardless of the predicted probability, we use the ground-truth action label to select the corresponding mode distribution and its decoder for sampling and decoding. This strategy encourages each branch to specialize in a single semantic mode and prevents mode-collapse. At inference time, we sample from all mode distributions according to the predicted crossing probability to ensure explicit multimodal outputs. Given a total of k samples and a predicted crossing probability P_c we draw $Pc \cdot k$ latent samples from the crossing-mode distribution and $k - Pc \cdot k$ samples from the non-crossing-mode distribution. The decoded trajectories from both modes are then merged as the final prediction. We show that this sampling yields more valid multimodal samples compared with existing methods in Section IV-E.

E. Objectives

As discussed above, we predict future displacements relative to the position $\mathbf{x}^t$, instead of directly predicting future absolute positions or frame-wise velocity. This helps reduce initial misalignment as well as reconstruction loss, and also mitigates overall trajectory drift. We apply the Best-of-Many RMSE loss[37] as the reconstruction term L_{rec} to encourage diversity, and a KL divergence term L_{kl} to bridge the gap between the prior network $p_\theta(M \mid \mathbf{X})$ and the recognition network $q_\phi(M \mid \mathbf{X}, \mathbf{Y})$ which is a closed form solution [37]. The cross-entropy loss L_{ce} is used for the action prediction head. The KL divergence loss and the overall loss L are formulated as follow:

$$L_{\text{KL}} = \sum_{m \in M} \omega_m \,\mathrm{KL}\big(q_\phi(z \mid X, Y, m) \,\big\|\, p_\theta(z \mid X, m)\big) \tag{4}$$

$$L = \alpha L_{\text{rec}} + \beta L_{\text{ce}} + \gamma L_{\text{kl}} \tag{5}$$

where α, β, γ are the weights for each loss term.

IV. Experiments

A. Datasets and Metrics

We evaluate our approach on two public ego-centric datasets, PIE [38] and JAAD [39], following the default train/test splits. Since ego-motion is not available in JAAD, we use the driver's action annotation as an alternative. We sample the sequences at 30 FPS with a window overlap ratio of 0.5, using 0.5s of observation to predict 1.5s into the future. Following [4], we evaluate our method on the image plane in pixels using the following metrics: MSE on bounding-box coordinates, reported at 0.5s, 1.0s, and 1.5s to reflect short-term to long-term prediction performance. We also report C_{MSE} and CF_{MSE} at 1.5s, which measure the average displacement and the final displacement of the bounding-box center coordinates, respectively.

B. Implementation Details

We set all embedding layers to $d = 64$, and mixed latent variable size to $d_m = 64$, compressed by a 3-layer MLP. We use 4 attention heads for all multi-head attention and Transformer blocks, and the depth of Transformer is set to 2. The loss weights α, β, γ are set to 1, 0.1, 0.1, respectively, where L_{kl} is activated after a warm-up of the first 10% of training. We train the model with batch size $B = 32$, terminated after 100 epochs, using a learning rate of $\eta = 5 \times 10^{-4}$ with a 10% cosine warm-up schedule and the AdamW [40] optimizer.

C. Quantitative Results

We conducted extensive experiments to compare our proposed model with previous state-of-the-art pedestrian trajectory prediction methods on ego-centric datasets, including Bitrap [4], SGnet [5], ABC+ [41], ENCORE [20], and TAGRN-SAR [19], and results are reported in Table I. Bottom rows show the results of our best model and variants for ablation.Ours-NM and Our-NB refer to our model without MTP and without incorporating behavior features. Compared with previous methods, our model without MTP outperforms the state-of-the-art TAGRN-SAR, achieving a 4% improvement in CF_{MSE}, and further reduces the error by 4.17% after introducing MTP, with more performance gains on the other metrics. In particular, the 1.5s prediction error C_{MSE} is improved by 25% and 7.41%, respectively. The incorporation of behavior features also significantly improves our method, leading to 25% and 14.01% gains in short-term prediction and final displacement compared to our model without behavior features. On the JAAD dataset, Table II reports that our model consistently achieves state-of-the-art performance, obtaining up to a 35% improvement in short-term prediction without the module. Both CF_{MSE} and C_{MSE} are further reduced, outperforming the top-performer ABC+. Overall, our framework demonstrates superior short-term prediction performance compared with previous methods, while the MTP improves all metrics and is particularly beneficial for long-term prediction on JAAD, confirmed by more accurate predictions at 1.5s.

TABLE I
COMPARISON OF STOCHASTIC RESULTS (20 SAMPLES) TO SOTA ON THE PIE DATASET.LOWER IS BETTER FOR ALL METRICS AND THE BEST RESULTS ARE **BOLD**.

	MSE			C_{MSE}	CF_{MSE}
	0.5s	**1s**	**1.5s**	**1.5s**	**1.5s**
BiTrap-NP [4]	23	48	102	81	261
SGNet-ED [5]	16	39	88	66	206
ABC+ [41]	16	38	87	65	191
ENCORE [20]	15	33	70	49	155
TAGRN-SAR [19]	16	31	61	36	100
Ours-NB	16	34	63	36	107
Ours-NM	13	31	56	27	96
Ours	**12**	**29**	**51**	**25**	**92**

TABLE II
COMPARISON OF STOCHASTIC RESULTS (20 SAMPLES) TO SOTA ON THE JAAD DATASET.

	MSE			C_{MSE}	CF_{MSE}
	0.5s	**1s**	**1.5s**	**1.5s**	**1.5s**
BiTrap-NP [4]	38	94	222	177	565
SGNet-ED [5]	37	86	197	146	443
ABC+ [41]	40	89	189	145	409
ENCORE [20]	32	85	210	167	554
Ours-NM	26	81	197	145	448
Ours	**24**	**73**	**177**	**141**	**401**

TABLE III
RESULTS OF INTRODUCING MODE SEPARATION MODULE TO SOTA.

	MSE			C_{MSE}	CF_{MSE}
	0.5s	**1s**	**1.5s**	**1.5s**	**1.5s**
BiTrap-NP [4]	23	48	102	81	261
SGNet-ED [5]	16	39	88	66	206
BiTrap-NP-M	17	46	99	77	252
SGNet-ED-M	**13**	**34**	**91**	**63**	**196**

TABLE IV
PER-FRAME ERROR VALIDITY EVALUATION TO BASELINES ON PIE DATASET.

	All Scenes MSE	**Two Side MSE**	**One Side MSE**
BiTrap-NP [4]	37.39	35.03	39.74
SGNet-ED [5]	36.75	35.13	38.39
Ours	**35.62**	**33.67**	**37.57**

To further assess the effectiveness of modeling each mode separately in our proposed method, we plug our MTP module into both Bitrap and SGnet, without modifying any other layers or replacing the decoder. As in Table III, we observe consistent improvements across all metrics for both baselines, although the performance gains are slightly smaller than those achieved by our model after introducing MTP. We assume that this is because the relatively simple encoders in Bitrap and SGnet cannot capture pedestrian interactions as effectively as ours, limiting the benefit of introducing MTP. Nevertheless, these results indicate that our MTP can be a model-agnostic component to improve existing methods, and explicitly modelling pedestrians' modes helps models predict trajectories more accurately.

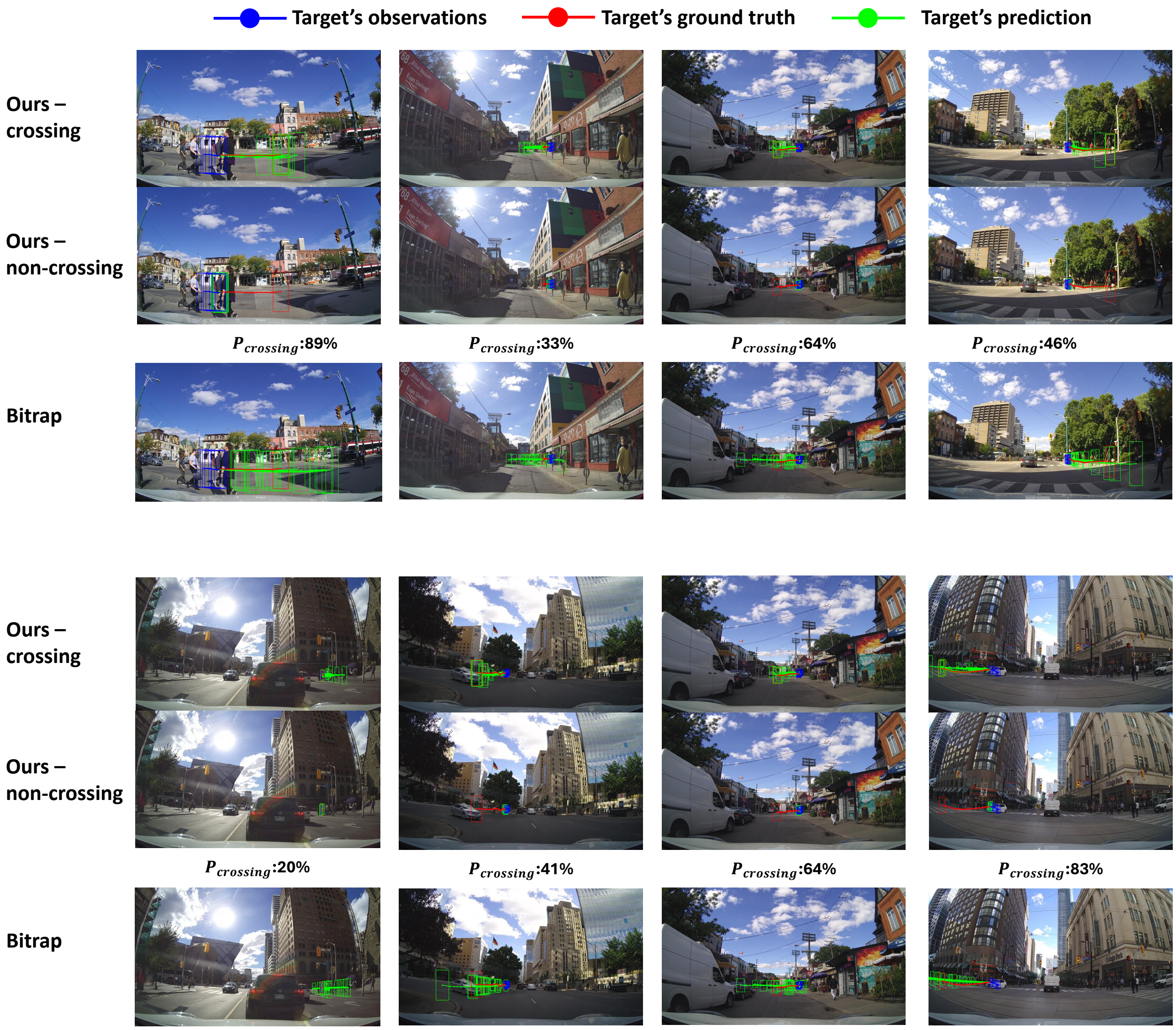

Fig. 3. Qualitative results from 2 different scenarios (Top and bottom). For each scenario, the first row demonstrates the prediction of crossing mode in our method, and the second row shows non-crossing prediction. All sampled trajectories of Bitrap are plotted together in the third row. The colors correspond to pedestrian's historical trajectory, future ground truth, prediction.

D. Qualitative Results

Fig. 3 shows a qualitative comparasion on PIE between our best model and Bitrap whose source is available. To investigate the performance on each mode, we force the model to predict 10 crossing and 10 non-crossing trajectories among 20 generated samples while showing the actual crossing probability which indicates the sampling ratio. The results show that the modes we defined consistently generate clearly different trajectories, corresponding to being stationary or crossing the road. Besides, within each mode, the final destinations are more concentrated or clustered, indicating that the learned mode effectively constrains the sampling process. In contrast, although Bitrap can produce diverse trajectories, some samples appear to lie between two sampled trajectories which are not exactly valid in the scene. This may be caused by the limitations of sampling from mixed modes. The results demonstrate our method's capacity to separate and model pedestrians' motion modes and pedestrians' intention awareness, resulting in more valid samples.

E. Data-driven Trajectory Validation

To verify whether modeling the distributions of pedestrian modes separately can produce more 'realistic' trajectories, we quantify validity by the distance between the predicted trajectory and any ground truth trajectories, and propose a valid trajectory evaluation protocol on the PIE dataset. We argue that this requires strict spatiotemporal consistency, so we further segment the test set into clips of varying lengths based on the ego-vehicle GPS speed, while keeping the original dataset split unchanged. We empirically use 10 km/h as the threshold [20] to segment the long videos into different clips. We then match all predictions of each pedestrian within a clip to ground-truth trajectories.

Specifically, taking the last observed position as the center, we search all ground-truth future trajectories within a radius of 150 pixels, and compare them against all predicted trajectories by the minimum displacement error. To be consistent with the benchmark, we also adopt MSE as the metric. Considering its sensitivity to outliers, we remove errors outside the inter-quartile range from all best matches. We further categorize each circle as two-side and one-side , depending on if both crossing and non-crossing ground-truth trajectories are available and matched against predictions. As shown in Table IV, our method achieves smaller errors against real samples than the baselines, and remains closer to ground-truth trajectories in two-side circles. The scarcity of comparable ground-truth trajectories within the search radius and the Best-of-many loss term, may cause this higher error compared with the prediction task, as the rest of other samples are not regularized while there are only a few ground truth to be matched.

V. Conclusion

We proposed a mode-aware trajectory prediction framework, MMPM, based on a novel mode-aware modelling for trajectory prediction, and the Pedestrian Interaction Module (PIM) which jointly considers pedestrian-vehicle and pedestrian-context interactions. Our method models pedestrian-context interactions by learning spatial correlations between the pedestrian and the scene context, and captures pedestrian-vehicle interactions by modeling the temporal dependencies between pedestrian and ego-vehicle motions, yielding multimodal motion representations. The Mode-aware Trajectory Predictor (MTP) module is then proposed to split multimodal pedestrian modes, facilitating the learning of trajectory distributions. MTP is model-agnostic, which can be integrated into existing frameworks to further improve future trajectory prediction performance. In addition, we propose a new experimental protocol to validate the validity of stochastic trajectory prediction. Our method achieves state-of-the-art performance over previous trajectory prediction methods, and extensive experiments demonstrate the effectiveness of our prediction and explicit modes.

There are limitations in our current approach that suggest potential directions for future work. Our method relies on the crossing annotations provided in PIE and JAAD. Since entire long pedestrian sequences are labeled by such annotations, the crossing behavior may not have started or may have already ended within the time horizon. The data split of our and previous methods does not consider this mislabeling when modeling action and trajectory prediction jointly. In addition, binary modes may not fully cover the multimodality of pedestrian motion yet. Future work could explore alternative approaches for identifying modalities, adopting either predefined or dynamically separated modalities to reduce reliance on annotations while introducing more diverse modes, e.g., going straight, or crossing to the left/right/front side of the road.

References

[1] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15 580–15 590, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:236956521

[2] Z. Sui, Y. Zhou, X. Zhao, A. Chen, and Y. Ni, "Joint intention and trajectory prediction based on transformer," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 7082–7088.

[3] X. Zhai, Z. Hu, D. Yang, L. Zhou, and J. Liu, "Social aware multi-modal pedestrian crossing behavior prediction," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4428–4443.

[4] Y. Yao, E. M. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation," *IEEE Robotics and Automation Letters*, vol. 6, pp. 1463–1470, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220845847

[5] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232380341

[6] Y. Lian, K. Zhang, M. Li, and J. Lin, "Hierarchical transformer-based red-light running prediction model for two-wheelers with multitask learning," *IEEE Transactions on Intelligent Vehicles*, pp. 1–15, 2024.

[7] Z. Su, G. Huang, S. Zhang, and W. Hua, "Crossmodal transformer based generative framework for pedestrian trajectory prediction," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2337–2343, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250507293

[8] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9793–9803, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232352504

[9] S. Saadatnejad, Y. Gao, K. Messaoud, and A. Alahi, "Social-transmotion: Promptable human trajectory prediction," *ArXiv*, vol. abs/2312.16168, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266551223

[10] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214802528

[11] B. Ivanovic and M. Pavone, "Modeling multimodal dynamic spatiotemporal graphs," *ArXiv*, vol. abs/1810.05993, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53115162

[12] L. Neumann and A. Vedaldi, "Pedestrian and ego-vehicle trajectory prediction from monocular camera," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 199–10 207, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235719781

[13] A. Rasouli and I. Kotseruba, "Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9844–9851, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252907322

[14] L. F. Chiara, P. Coscia, S. Das, S. Calderara, R. Cucchiara, and L. Ballan, "Goal-driven self-attentive recurrent networks for trajectory prediction," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2517–2526, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248377731

[15] D. Choi and K. Min, "Hierarchical latent structure for multi-modal vehicle trajectory forecasting," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250425801

[16] Z. Wang, H. Miao, S. Wang, R. Wang, J. Wang, and J. Zhang, "C2f-tp: A coarse-to-fine denoising framework for uncertainty-aware trajectory prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12 810–12 817.

[17] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic trajectory prediction via motion indeterminacy diffusion," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 092–17 101, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247748591

[18] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5517–5526, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257631504

[19] C.-H. Chiu, Y.-C. Lin, and Y.-J. Chen, "Temporal attention with gated residual network for stepwise multimodal trajectory prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–9, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276948460

[20] A. Rasouli, "A novel benchmarking paradigm and a scale- and motion-aware model for egocentric pedestrian trajectory prediction," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5630–5636, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264145930

[21] P. S. Chib, A. Nath, P. Kabra, I. Gupta, and P. Singh, "Ms-tip: Imputation aware pedestrian trajectory prediction," in *International Conference on Machine Learning*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272330450

[22] R. Karim, S. M. A. Shabestary, and A. Rasouli, "Destine: Dynamic goal queries with temporal transductive alignment for trajectory prediction," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2230–2237, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263834714

[23] C. Xu, M. Li, Z. Ni, Y. Zhang, and S. Chen, "Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6488–6497, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248239810

[24] C. Wong, B. Xia, and X. You, "Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 005–19 015, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263829662

[25] L. Crosato, H. P. H. Shum, E. S. L. Ho, and C. Wei, "Interaction-aware decision-making for automated vehicles using social value orientation," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1339–1349, 2023.

[26] L. Crosato, C. Wei, E. S. L. Ho, and H. P. H. Shum, "Human-centric autonomous driving in an av-pedestrian interactive environment using svo," in *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 2021, pp. 1–6.

[27] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 21 050–21 061, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248789262

[28] Y. Lian, K. Zhang, M. Li, and J. Lin, "Hierarchical transformer-based red-light running prediction model for two-wheelers with multitask learning," *IEEE Transactions on Intelligent Vehicles*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270716362

[29] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7593–7602, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4406882

[30] S. M. Pang, J. X. Cao, M. Y. Jian, J. Lai, and Z. Y. Yan, "Br-gan: A pedestrian trajectory prediction model combined with behavior recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 609–24 620, 2022.

[31] B. Zhang, N. Song, and L. Zhang, "Demo: Decoupling motion forecasting into directional intentions and dynamic states," *ArXiv*, vol. abs/2410.05982, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:273228276

[32] T. Shu, Y. Peng, L. Fan, H. Lu, and S.-C. Zhu, "Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations," *Topics in cognitive science*, vol. 10 1, pp. 225–241, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:26105917

[33] B. Lim, S. Ö. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *ArXiv*, vol. abs/1912.09363, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209414891

[34] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv: Learning*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:5273326

[35] N. Shazeer, "Glu variants improve transformer," *ArXiv*, vol. abs/2002.05202, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211096588

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *ArXiv*, vol. abs/2005.12872, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218889832

[37] A. Bhattacharyya, B. Schiele, and M. Fritz, "Accurate and diverse sampling of sequences based on a "best of many" sample objective," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8485–8493, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49319880

[38] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6261–6270, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204959605

[39] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *ArXiv*, vol. abs/1609.04741, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:4816620

[40] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv*, vol. abs/1711.05101, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3312944

[41] M. Halawa, O. Hellwich, and P. Bideau, "Action-based contrastive learning for trajectory prediction," *ArXiv*, vol. abs/2207.08664, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250627278