2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Modeling 3D Pedestrian-Vehicle Interactions for Vehicle-Conditioned Pose Forecasting

Guangxun Zhu[1], Xuan Liu[1], Nicolas Pugeault[1], Chongfeng Wei[2], and Edmond S. L. Ho[1*]

***Abstract*— Accurately predicting pedestrian motion is crucial for safe and reliable autonomous driving in complex urban environments. In this work, we present a 3D vehicle-conditioned pedestrian pose forecasting framework that explicitly incorporates surrounding vehicle information. To support this, we enhance the Waymo-3DSkelMo dataset with aligned 3D vehicle bounding boxes, enabling realistic modeling of multi-agent pedestrian–vehicle interactions. We introduce a sampling scheme to categorize scenes by pedestrian and vehicle count, facilitating training across varying interaction complexities. Our proposed network adapts the TBIFormer architecture with a dedicated vehicle encoder and pedestrian–vehicle interaction cross-attention module to fuse pedestrian and vehicle features, allowing predictions to be conditioned on both historical pedestrian motion and surrounding vehicles. Extensive experiments demonstrate substantial improvements in forecasting accuracy and validate different approaches for modeling pedestrian–vehicle interactions, highlighting the importance of vehicle-aware 3D pose prediction for autonomous driving.**

I. Introduction

Autonomous driving has attracted significant attention for its potential to revolutionize transportation, but it still faces numerous challenges in complex urban environments [1]. Among these, pedestrian prediction is particularly critical, as pedestrians are common, vulnerable, and highly dynamic road users, whose flexible and sometimes unpredictable behaviors pose significant challenges for perception systems [2]. Accurately understanding these behaviors is therefore essential for safe navigation and planning.

A main challenge in pedestrian prediction lies in modeling their interactions with other agents. Pedestrian behavior is influenced by other road users (e.g., pedestrians, vehicles, cyclists) [3], and insufficient or overly coarse modeling of these interactions can result in inaccurate trajectory predictions, potentially leading autonomous vehicles to adopt inappropriate responses. Most prior studies [4], [5], [6], [7], [8], [9], [10] model pedestrian interactions using 2D representations such as locations and 2D poses. For handling data captured from the bird's-eye view, SocialCircle [4] introduces an angle-based representation to capture relative spatial layouts for interaction modeling, and SocialMOIF [5] leverages multi-order intention fusion to enhance trajectory prediction. TrajCLIP [6] proposed to improve the continuity in the predicted pedestrian trajectories using contrastive learning to improve the consistency between the feature spaces of the historical and future trajectories. Social Value Orientation (SVO), which is a concept in Psychology for a person's preferences for allocating resources between themselves and others, has been incorporated into Reinforcement Learning (RL) frameworks [7], [8] for modeling different behaviours between pedestrians and drivers. While these methods explicitly consider social interactions, they remain limited to 2D numerical coordinates, which restricts their ability to capture fine-grained 3D motion dynamics.

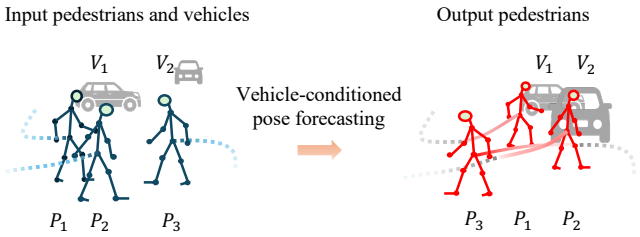

Fig. 1: Illustration of our Vehicle-conditioned pedestrian pose forecasting. Pedestrian predictions are not only based on their historical motion but are also influenced by surrounding vehicle information.

Another stream of existing research in pedestrian-vehicle interactions focused on analysing ego-centric view data available in public datasets such as TITAN [11], JAAD [12], [13] and PIE [14]. A recent work, namely PedVLM [15], predicts the intention (i.e., crossing or not crossing) of the pedestrian by feeding the RGB image and the corresponding optical flow computed from the ego-centric view alongside the textual description of the scene to a large visual-language models (VLM). PedFormer [9] fuses multimodal features, including 2D coordinates of pedestrians from the ego-centric view, ego-vehicle motion and semantic segmentation of the scene for pedestrian behavior prediction. Further to multi-modal sensor fusion, PIP-Net [10] takes advantage of fusing features extracted from multi-camera views (left, front, and right) captured using on-board cameras provided in the new Urban-PIP dataset to better model the contextual information when modeling pedestrian-vehicle interactions.

Although some approaches may leverage a combination of multiple modalities, such as images, vehicle ego-motion, 2D pedestrian poses, or vehicle 2D bounding boxes, they are limited in capturing fine-grained pedestrian behaviors and motion dynamics, which may result in inaccurate predictions and unsafe planning decisions for autonomous vehicles. More recently, several works have explored the use of 3D human poses to better capture motion dynamics [16], [17].

*Corresponding author 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

Compared to 2D data, 3D poses provide richer spatial structures and depth information, thereby providing a stronger basis for fine-grained interaction modeling. Nevertheless, these approaches focus only on pedestrian–pedestrian interactions and do not consider the influence of multiple types of agents in autonomous driving scenarios, primarily due to the scarcity of 3D pose datasets captured in real driving environments. For example, many approaches rely on MuPoTS-3D [18], which provides 3D poses estimated only from RGB images and is collected in non-driving contexts, limiting realism. Wang et al. [19] attempted to synthesize three-person interactions by combining single- and two-person motion sequences from the high-quality CMU-Mocap dataset [20], recorded using an optical motion capture system. However, this produces a synthetic dataset with limited diversity and lacks realistic multi-agent interactions in traffic scenarios. The most recent dataset, JRDB-GlobMultiPose (JRDB-GMP) [21], is collected in real-world environments using a moving robot platform, yet the 3D poses are obtained purely via RGB-based pose estimation, which affects the motion quality due to depth ambiguity and reduced robustness under occlusion. Moreover, these datasets lack aligned 3D data from real autonomous driving scenarios, making it difficult for models trained on them to generalize to realistic interactions in such environments.

To address this challenge, Crosato et al. [22] proposed a VR platform for capturing interactions between a pedestrian and a vehicle controlled by human subjects in 3D in a virtual environment. However, the data was not captured in real-world and only contains limited scenarios with 1 pedestrian and 1 vehicle. More recently, Waymo-3DSkelMo [23] was constructed from the Waymo Open Dataset Perception Benchmark (hereafter referred to as Waymo) [24], a large-scale autonomous driving dataset, to provide 3D pedestrian skeletal motion data. Specifically, 3D human shapes (estimated using LiDAR-HMR [25] with SMPL [26]) and motion priors (Neural Motion Fields (NeMF) [27]) were used to reconstruct high-quality and natural 3D skeletal motion from the raw LiDAR range images in Waymo. Waymo-3DSkelMo significantly increases the number of 3D skeletal poses from approximately 10k in Waymo to over 2.4 million and is aligned with other data modalities, such as 3D vehicle bounding boxes and LiDAR point clouds, making it possible to model full-scene interactions. However, only 3D skeletal motion is benchmarked in Waymo-3DSkelMo, which limits its usage for broader autonomous driving applications.

In this paper, we extend the Waymo-3DSkelMo dataset by incorporating the 3D bounding box information of each vehicle from the Waymo dataset, covering both moving and parked vehicles, as both dynamic and static vehicles are common and safety-critical agents that can strongly influence pedestrian behaviour. Through aligning the vehicle and skeletal motion in a common 3D space, 3D data representing the interactions between multiple pedestrians and vehicles can be obtained. We further propose a sampling scheme to divide the scenes into different categories to facilitate the training of interaction modeling networks with different complexity levels (i.e., different numbers of vehicles and pedestrians). Finally, we propose a 3D Vehicle-conditioned pedestrian pose forecasting network by incorporating the vehicle information and adapting the TBIFormer [28] architecture. In this network, pedestrian predictions not only rely on their historical motion but are also conditioned on surrounding vehicle information, as illustrated in Fig. 1. Extensive experimental results are obtained to demonstrate the benefits of including the vehicle information in the 3D pose forecasting of the pedestrians.

Our contributions can be summarized as follows:

- We enhance the Waymo-3DSkelMo dataset by incorporating 3D vehicle bounding boxes and introducing a scene-level sampling scheme that categorizes interactions based on the number of pedestrians and vehicles, enabling more realistic and structured modeling of pedestrian–vehicle interactions.
- We propose a new 3D pedestrian pose forecasting network that incorporates vehicle information, allowing pedestrian predictions to be conditioned on both historical pedestrian motion and surrounding vehicles.
- We provide extensive benchmarking results, demonstrating the benefits of incorporating vehicle information for accurate 3D pose forecasting, and validating different approaches to modeling interactions between pedestrians and vehicles. *Our implementation code is publicly available* [3].

II. Dataset

In this section, we first introduce the enhancement of the Waymo-3DSkelMo [23] dataset by incorporating vehicle information in Section II-A. Next, we present the scene segmentation strategy in Section II-B, which is designed to sample diverse data for the experiments.

A. Incorporating Vehicle Information

We employ Waymo-3DSkelMo [23] and Waymo Open Dataset (Perception) [24] to train and validate pedestrian–vehicle interaction models. In particular, Waymo-3DSkelMo contains 2,438,145 3D skeletal poses of the pedestrian reconstructed from the raw LiDAR range images provided by the Waymo dataset. The naturalness of the reconstructed pedestrian motions was further enhanced using NeMF [27] as the human motion prior. Waymo-3DSkelMo provides 3D keypoints of all pedestrians in real-world scenes, which can be spatially and temporally aligned with the vehicle 3D bounding boxes in Waymo in this study. Waymo-3DSkelMo contains **837 scenes** with an overall duration of **4 hours**, covering a diverse set of urban scenarios and capturing a large number of pedestrian–vehicle interactions.

B. Scene Segmentation

Because different scenes contain varying numbers of agents with diverse spatial and temporal distributions, it is difficult to utilize all scenes as a single training dataset due

[3]Project repository: https://github.com/GuangxunZhu/VehCondPose3D

TABLE I: Number of segmented scenes (training + validation) with different numbers of pedestrians and vehicles within 0–15 m in the dataset.

# of Veh.	Number of Pedestrian(s)			
	1	2	3	4
0	17762 + 8236	8236 + 2232	5160 + 1333	3503 + 875
1	15446 + 7534	7534 + 2005	4975 + 1355	3541 + 977
2	14336 + 7321	7321 + 1782	5077 + 1233	3810 + 883
3	12982 + 6815	6815 + 1733	4903 + 1245	3764 + 902
4	10683 + 6192	6192 + 1272	4351 + 944	3393 + 738

to the large difference in spatial density. As a result, further segmentation of the scenes is necessary to provide more meaningful training data for downstream tasks.

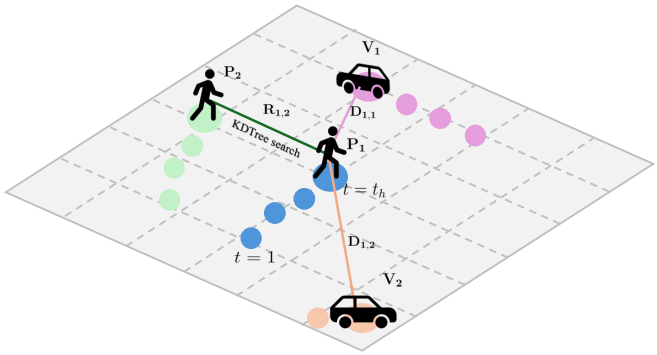

Fig. 2: Illustration of pedestrians and vehicles in a scene with their trajectories and inter-agent distances.

In particular, we segmented the scenes based on the distance between pedestrians and vehicles nearby. Specifically, as illustrated in Fig. 2, we first apply a KDTree [29], a data structure that enables efficient nearest-neighbor searches in multi-dimensional space, to quickly find pedestrian groups that are closest to each other based on the average root joint positions over a temporal window. In this work, we focus on the extracted scenes with the number of pedestrians between 1 and 3 since the number of scenes decreases sharply above this range as indicated in the statistics presented in Table I. A sliding window is then employed to select temporally overlapping segments. For each segment, we compute the maximum pairwise distance between pedestrians' root positions at each timestamp and then take the minimum over all timestamps, denoted as R:

$$R = \min_{t=1,\ldots,T} \left(\max_{i,j \in \{1,\ldots,N_p\}, i \neq j} \|\mathbf{r}_i(t) - \mathbf{r}_j(t)\|_2 \right) \quad (1)$$

Here, $\mathbf{r}_i(t)$ represents the 3D root position of the i-th pedestrian at time t, N_p is the number of pedestrians in the segment, and T is the total number of frames in a time window. We retain the segment if R is below a threshold of 18 m which is the maximum distance used in the TBIFormer [28] training dataset CMU-Mocap (UMPM) [20], [30].

Since the number of surrounding vehicles varies across pedestrian samples, this inconsistency introduces challenges for training and validation. To address this, we select the set of vehicles V_{selected} whose average distance to the nearest

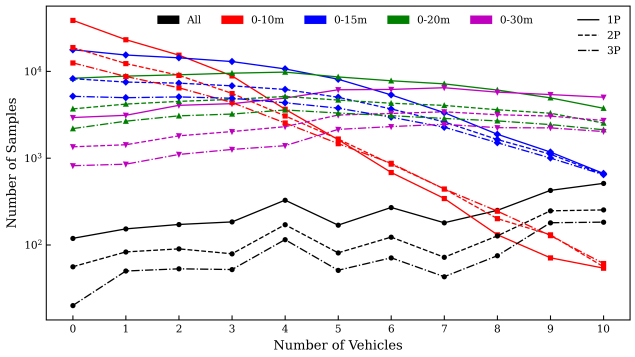

Fig. 3: Number of pedestrian–vehicle interaction training samples under varying numbers of surrounding vehicles and different distance thresholds for scenarios ranging from one to three pedestrians.

pedestrian over the frames in the time window T is below a predefined threshold τ:

$$V_{\text{selected}} = \left\{ i \in V \;\middle|\; \frac{1}{|T|} \sum_{t \in T} \min_{j \in P} \|r_j(t) - v_i(t)\|_2 \leq \tau \right\}$$

where $r_j(t)$ and $v_i(t)$ denote the positions of pedestrian j and vehicle i at frame t, respectively.

As shown in Fig. 3, the number of training samples associated with different vehicle counts varies under different distance thresholds. From Fig. 3, we can see that if the distance threshold is too small, the number of available samples decreases sharply as the number of vehicles increases. Conversely, if the threshold is too large, the number of samples becomes nearly uniform across different vehicle counts, but intuitively, vehicles that are too far away are unlikely to interact meaningfully with pedestrians. Therefore, to maximize the amount of realistic interaction data while keeping the experimental setup manageable, we select scenes with 1–4 vehicles within a 0–15 m range for one-, two-, and three-person scenes, resulting in a total of 3×4 experimental conditions. This choice reflects a practical trade-off between training set size and empirically observed performance, which we will further justify with evidence from prior studies in future work. The number of segmented scenes for each condition is summarized in Table I.

In summary, the processed dataset for training and validation includes all possible pedestrian combinations together with the actual number of vehicles within a fixed range. Although the proposed scene segmentation approach may limit the potential for capturing pedestrian–pedestrian interactions, the inclusion of real vehicle information makes it well-suited for evaluating pedestrian–vehicle interactions.

III. Method

In this section, we will introduce our proposed vehicle-conditioned pedestrian pose forecasting network as illustrated in Fig. 4. We first define the problem in Section III-A then present the overall framework in Section III-B,

and further describe its core components in the following subsections.

A. Problem Definition

Consider P pedestrians with observed 3D skeletal poses over $T+1$ frames, denoted as $X_p^{1:T+1} = \{x_p^1, x_p^2, \ldots, x_p^{T+1}\}$ for $p = 1, \ldots, P$. To capture motion dynamics, we represent pedestrian motion as frame-to-frame displacements $y^i = x^{i+1} - x^i$, forming the displacement sequence $Y^{1:T} = \{y^1, \ldots, y^T\}$. Similarly, for K surrounding vehicles, we observe their 3D trajectories $V^{1:T+1} = \{v_1^{1:T+1}, \ldots, v_K^{1:T+1}\}$ in the same coordinate space, where each v_k^t contains the 8 corner points of the vehicle's 3D bounding box at frame t, and convert them to displacement sequences $U^{1:T} = \{u^1, \ldots, u^T\}$ with $u_k^i = v_k^{i+1} - v_k^i$. The goal is to leverage both pedestrian and vehicle displacement sequences, $(Y^{1:T}, U^{1:T})$, to predict the N future pedestrian displacements $Y^{T+1:T+N}$, which are then transformed back into 3D poses $X^{T+2:T+N+1}$.

B. Overall Framework

As illustrated in Fig. 4, we extend TBIFormer [28], which effectively captures long-term temporal dependencies and fine-grained body-part dynamics, to incorporate vehicle information for pedestrian 3D pose forecasting. For the pedestrian branch, our network retains the core TBIFormer architecture, including stacked TBIFormer blocks and the Temporal Body Partition Module (TBPM). To leverage vehicle information, we introduce an additional **vehicle branch** that processes displacement sequences of vehicles in the same 3D coordinate space as pedestrians. Pedestrian and vehicle features are then fused through a **Pedestrian–Vehicle Interaction Cross-Attention (PVI-CA)** module, which models interactions between pedestrian body parts and vehicle groups. Notably, we extend the Trajectory-Aware Relative Position Encoding (TRPE) to the pedestrian–vehicle cross attention, providing discriminative spatial and temporal cues for accurate pose prediction. The fused representation is subsequently fed into the traditional Transformer decoder [31] to predict future pedestrian poses. Overall, our model explicitly integrates vehicle motion while preserving the temporal, spatial, and social modeling capabilities of TBIFormer.

C. Vehicle Encoder

To incorporate vehicle information into the same 3D space as pedestrians, we design a Vehicle Encoder that converts 3D vehicle bounding box data into a compact feature representation suitable for interaction modeling. Each vehicle is represented by the 8 corner points of its 3D bounding box. These corner points are first used to compute a displacement sequence, capturing the vehicle's temporal motion across consecutive frames. The displacement sequence is then transformed via Discrete Cosine Transformation (DCT) [32], which suppresses high-frequency components and yields a more compact representation in the displacement trajectory space, preserving the primary motion trends while reducing noise and redundancy. The resulting sequence is downsampled along the temporal dimension to length L. Following this, the 8 corner points are divided into $B_V = 12$ logical groups (12 edges of the bounding box) , a choice motivated by the ablation study in Section IV-E and analogous to the body-part division (Temporal Body Partition Module) used for pedestrians as in TBIFormer (i.e. 3D keypoints of each pedestrian are divided into 5 groups to represent the trunk and 4 limbs), allowing the network to model localized motion patterns within the vehicle structure. For N_V vehicles in the scene, concatenating all groups across all vehicles forms a Multi-Vehicle Feature sequence of length $U = N_v \times B_v \times L$ with feature dimension D. Temporal positional encoding (TPE) and identity encoding (IE) are applied to each sequence in the same manner as the pedestrian branch, ensuring temporal dynamics and individual vehicle identity are preserved. The Multi-Vehicle Feature sequence is then processed through a two-layer MLP to extract high-dimensional embeddings, ensuring the same feature dimension as the pedestrian branch, which facilitates seamless integration in the subsequent cross-attention module, enabling effective fusion of pedestrian and vehicle information for interaction-aware 3D pose forecasting.

D. Pedestrian–Vehicle Interaction Cross-Attention

To effectively incorporate vehicle information into pedestrian motion forecasting, we extend the original Trajectory-Aware Relative Position Encoding (TRPE) to model interactions between pedestrians and vehicles. The extended TRPE encodes trajectory-aware relational information between each pedestrian and surrounding vehicles, capturing both spatial and temporal context, as illustrated in Fig. 4.

The resulting trajectory-aware relative position embeddings are then integrated into a cross-attention mechanism, referred to as Pedestrian-Vehicle Interaction Cross-Attention (PVI-CA). Given the pedestrian features H^p from the pedestrian branch and vehicle features H^v from the vehicle branch, the cross-attention computes:

$$Q = H^p W_Q, \quad K = H^v W_K, \quad V = H^v W_V, \tag{2}$$

$$\text{PVI-CA}(Q, K, V) = \text{softmax}\Big(\frac{QK^\top + B_{\text{TRPE}}}{\sqrt{d_z}}\Big)V, \tag{3}$$

where B_{TRPE} is the contextual bias derived from the extended TRPE between pedestrian and vehicle features (see Fig. 4). By integrating both spatial and trajectory-aware relational information, PVI-CA allows the model to selectively attend to vehicles that are most relevant to each pedestrian's motion.

E. Decoder

As illustrated in Fig. 4, we follow the standard Transformer decoder design [31]. Specifically, the joint coordinates of the last observed pedestrian sub-sequence are concatenated and down-sampled by a 1D convolution to form global body query tokens, while the fused pedestrian–vehicle features from the PVI-CA module serve as keys and values. The decoder encodes the relations between the current queries and historical context, conditioned on

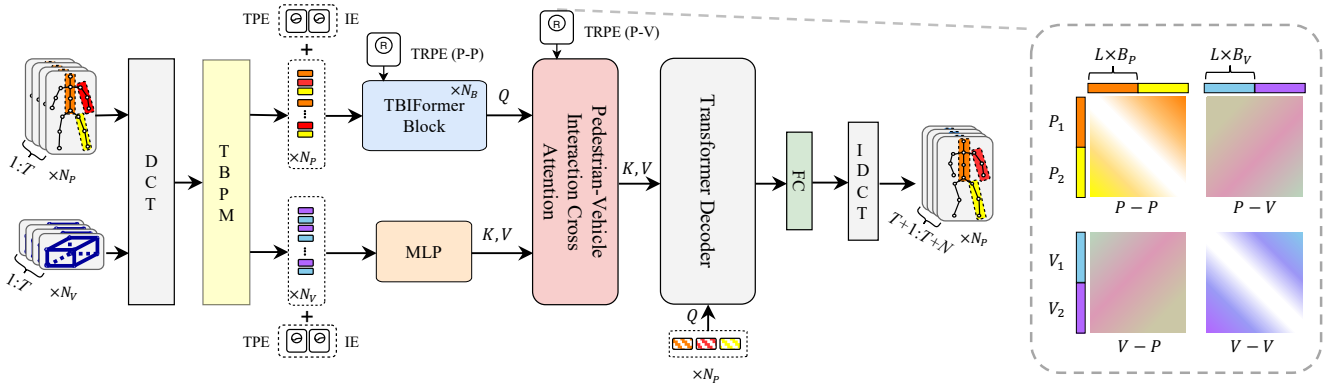

Fig. 4: **Overview of the proposed Vehicle-conditioned pedestrian pose forecasting network.** The model receives 3D pedestrian poses and 3D vehicle bounding box data, transforms them into displacement sequences, and applies Discrete Cosine Transform (DCT) to discard high-frequency components for a more compact representation. Pedestrian and vehicle features are then processed through separate encoders and fused via cross-attention before being decoded into future pedestrian poses.

pedestrian and vehicle information. Finally, two fully connected layers followed by an Inverse Discrete Cosine Transformation (IDCT) [32] generate the future motion trajectory $X_{T+2:T+N+1}$ for each pedestrian.

F. Training Objective

We treat the vehicle motion as a conditioning signal and therefore do not predict its future trajectory. As a result, the training objective is applied only to pedestrian poses.

We optimize the model using a reconstruction loss based on the Mean Per Joint Position Error (MPJPE). For a single training sample, the loss is defined as

$$\mathcal{L}_{\text{rec}} = \frac{1}{J * N} \sum_{t=T+1}^{T+N} \sum_{j=1}^{J} \|\hat{\mathbf{y}}_{t,j} - \mathbf{y}_{t,j}\|_2 \,, \tag{4}$$

where $\hat{\mathbf{y}}_{t,j}$ and $\mathbf{y}_{t,j}$ denote the estimated and ground-truth pose displacements of joint j at time step t, respectively, and J is the number of body joints.

IV. Experiment

A. Implementation Details

We implement our framework in PyTorch, and the experiments are performed on a single Nvidia GeForce RTX 4090 GPU. We train our model for 50 epochs using the Adam optimizer with a batch size of 32, a learning rate of 2×10^{-5}, and a dropout rate of 0.2. All other settings remain consistent with TBIFormer [28], with the model trained for 2 s (50 frames) and evaluated for 1 s (25 frames) prediction.

B. Baselines

Since no prior work directly addresses 3D pose forecasting with vehicle context, we adapt the state-of-the-art multi-agent 3D pose forecasting method TBIFormer [28] as our baseline. Both TBIFormer and our model are trained and evaluated on the one-, two-, and three-pedestrian scenarios extracted from Waymo-3DSkelMo and the Waymo Dataset, following the dataset procedure described in Section II-B. In each scenario, TBIFormer is applied in the pedestrian-only setting, while our model uses pedestrian + vehicle(s). For both methods, we use a 2 s input and predict 1 s ahead, with training and evaluation conducted under varying numbers of vehicles.

C. Metrics

We evaluate our predictions using three widely used metrics, which capture different aspects of pose and trajectory accuracy.

JPE (Joint Position Error): Evaluates both global and local pose predictions by averaging the L_2 distance of all joints at each predicted timestep:

$$\text{JPE} = \frac{1}{T \cdot N_j} \sum_{t=1}^{T} \sum_{j=1}^{N_j} \|\hat{\mathbf{p}}_t^{(j)} - \mathbf{p}_t^{(j)}\|_2 \tag{5}$$

where T is the number of predicted timesteps, N_j is the number of joints, $\hat{\mathbf{p}}_t^{(j)}$ and $\mathbf{p}_t^{(j)}$ are the predicted and ground-truth positions of joint j at timestep t, respectively.

APE (Aligned Pose Error): Evaluates the forecasted local motion by measuring the average L_2 distance of all joints, after removing global translation to capture pure pose error:

$$\text{APE} = \frac{1}{T \cdot N_j} \sum_{t=1}^{T} \sum_{j=1}^{N_j} \|(\hat{\mathbf{p}}_t^{(j)} - \hat{\mathbf{r}}_t) - (\mathbf{p}_t^{(j)} - \mathbf{r}_t)\|_2 \tag{6}$$

where $\hat{\mathbf{r}}_t$ and $\mathbf{r}_t$ are the predicted and ground-truth root positions at timestep t.

FDE (Final Displacement Error): Measures the accuracy of the forecasted global trajectory by computing the L_2 distance of the root position at the final predicted timestep.

$$\text{FDE} = \|\hat{\mathbf{r}}_T - \mathbf{r}_T\|_2 \tag{7}$$

D. Results

We conduct experiments separately for one-, two-, and three-pedestrian scenarios, comparing prediction performance with and without surrounding vehicles under varying numbers of vehicles. The evaluation metrics at different

TABLE II: Evaluation metrics (in millimeters) for different prediction horizons under varying numbers of vehicles. Results compare one-, two-, and three-pedestrian scenarios, showing performance with and without surrounding vehicle information. Metrics include MPJPE, APE, and FDE at the prediction frames of 0.2s, 0.6s, and 1.0s. Bold face indicates best performance.

# of Veh.	Metric	1 Pedestrian Scene						2 Pedestrians Scene						3 Pedestrians Scene					
		[28] (Ped. only)			Ours			[28] (Ped. only)			Ours			[28] (Ped. only)			Ours		
		0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s
1	MPJPE↓	84	232	316	**78**	**225**	**311**	88	228	324	**79**	**216**	**304**	97	251	361	**82**	**215**	**299**
	APE↓	53	109	118	**49**	**107**	**115**	54	110	116	**52**	**107**	**115**	57	112	**119**	**55**	**111**	**119**
	FDE↓	57	183	273	**54**	**179**	**271**	63	181	283	**55**	**170**	**265**	72	206	323	**55**	**165**	**255**
2	MPJPE↓	88	227	**302**	**82**	**224**	303	86	225	313	**78**	**205**	**275**	94	238	334	**80**	**208**	**283**
	APE↓	56	112	118	**53**	**111**	**117**	54	112	120	**53**	**110**	**117**	59	118	127	**55**	**114**	**123**
	FDE↓	61	178	**259**	**56**	**175**	**259**	60	175	273	**52**	**152**	**229**	66	187	290	**53**	**155**	**236**
3	MPJPE↓	87	220	300	**79**	**213**	**293**	89	231	323	**76**	**208**	**285**	93	237	334	**79**	**201**	**272**
	APE↓	54	114	123	**52**	**113**	**121**	56	115	124	**52**	**112**	**122**	57	117	**125**	**56**	**115**	**125**
	FDE↓	61	168	254	**53**	**161**	**248**	61	178	276	**50**	**155**	**236**	65	184	284	**50**	**147**	**221**
4	MPJPE↓	83	206	**277**	**77**	**205**	285	92	242	348	**76**	**202**	**278**	99	253	364	**78**	**207**	**279**
	APE↓	54	112	121	**52**	**110**	**119**	55	113	122	**51**	**109**	**119**	58	119	131	**54**	**115**	**125**
	FDE↓	55	**152**	**229**	**51**	155	241	67	197	307	**50**	**151**	**230**	71	203	316	**52**	**152**	**229**

prediction horizons are summarized in Table II. From the results, several observations can be made.

Firstly, the inclusion of vehicles consistently improves overall prediction accuracy across all metrics. For example, in the one-pedestrian scenario with a single vehicle, the MPJPE decreases from 84–316 mm (pedestrian only) to 78–311 mm (ours), corresponding to an improvement of around 1.5%–7%. Similar improvements are observed for APE and FDE, with reductions of around 0.7%–7.5% overall. This trend holds across different numbers of vehicles and all pedestrian scenarios, with overall improvements ranging from 1.5%-21% for MPJPE, 1.8%-12% for APE, and 0.7%-15% for FDE, depending on the number of vehicles in the scenes. However, in the one-pedestrian scenario with four vehicles, the prediction performance at 1.0s slightly decreases. Since the four-vehicle outperforms in the two- and three-pedestrian scenarios, this is unlikely due to limited capacity for multiple contextual agents and is more likely because some vehicles are distant or less relevant, which introduces noise to the training scenarios.

Secondly, comparing the one-, two-, and three-pedestrian scenarios, the overall improvements generally increase with the number of pedestrians. For example, in the one-pedestrian scenario with a single vehicle, the average MPJPE across 0.2s, 0.6s, and 1.0s decreases from 211 mm (pedestrian only) to 205 mm (ours), corresponding to an improvement of about 2.9%. In the two-pedestrian scenario with one vehicle, the average MPJPE decreases from 213 mm to 200 mm ($\sim 6.4\%$ improvement), and in the three-pedestrian scenario with one vehicle, it decreases from 236 mm to 199 mm ($\sim 15.9\%$ improvement). This trend is likely because multi-pedestrian scenes involve more complex interactions, allowing vehicle information to provide stronger contextual cues, whereas in single-pedestrian scenes, the interaction context from vehicles is more limited.

Thirdly, we examine the effect of increasing vehicle count within each scenario by focusing on relative improvements, as absolute metrics are not directly comparable due to differences in the underlying training data for each vehicle count. We observe that, overall, relative improvements increase with the number of vehicles. For instance, in the two-pedestrian scenario, increasing the number of surrounding vehicles leads to steadily larger improvements in overall MPJPE, with the improvement growing from approximately 6% with one vehicle to around 18% with four vehicles. A similar trend is observed in the three-pedestrian scenario, where additional vehicles also result in substantial performance gains. In contrast, in the one-pedestrian scenario, the improvements are relatively small and even show slight decreases when adding more than three vehicles, indicating that the contribution of vehicle information is limited when only a single pedestrian is present.

Overall, the results confirm that vehicle information plays a positive role in improving pedestrian trajectory prediction. Compared with the baseline, our model more effectively leverages such contextual cues, leading to consistent performance gains.

TABLE III: Ablation studies on different components of Our model. Our full method and its variants are evaluated on the 2 pedestrians and 3 vehicles subdataset (averaged over 0.2s, 0.4s, 0.6s, 0.8s, 1.0s). Bold face indicates best performance.

Method	MPJPE↓	APE↓	FDE↓
TBIFormer	218.8	103.0	173.4
+ Vehicle Center	224.4	105.2	177.8
+ Veh Branch w/o TRPE	195.4	100.4	150.2
+ Veh Branch w TRPE			
1 group	195.4	100.2	150.0
2 group	194.8	100.2	150.6
4 group	195.4	100.2	150.0
6 group	195.6	100.4	150.6
8 group	195.4	100.0	150.2
12 group	194.8	100.0	149.8
Full Model (with TRPE)	**194.8**	**100.0**	**149.8**

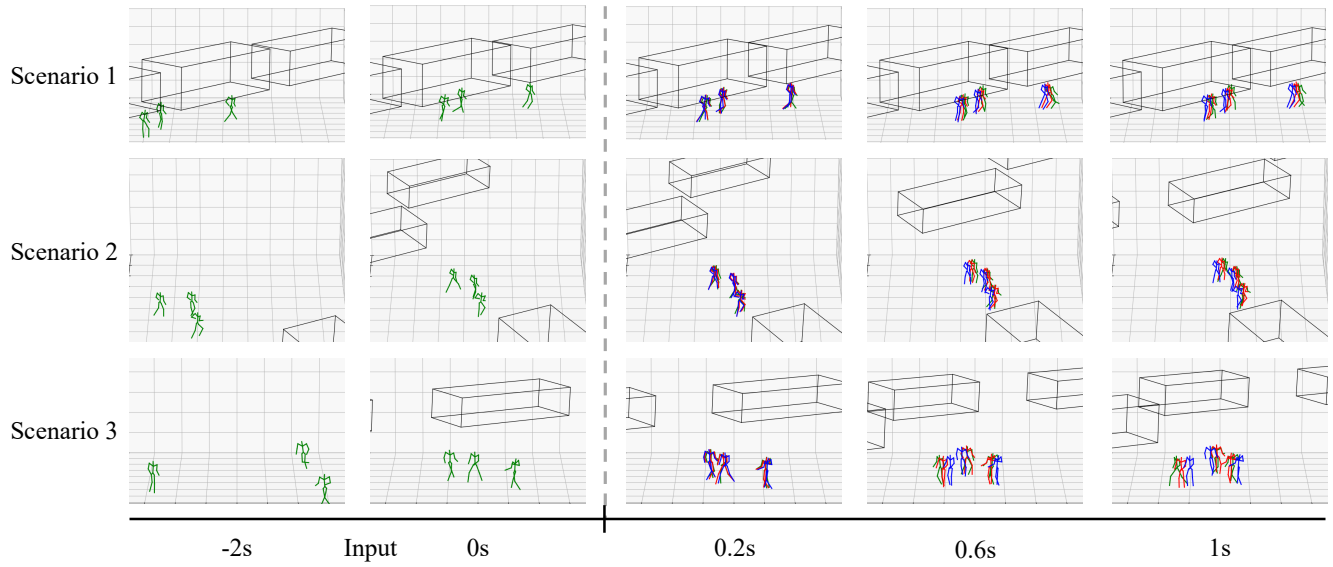

Fig. 5: Qualitative comparison between TBIFormer [28] and our model across three scenarios involving three pedestrians and four vehicles (either stationary or moving). Ground truth trajectories are shown in green, TBIFormer predictions in blue, and ours in red. Black boxes denote surrounding vehicles used by our model as input, whereas TBIFormer does not use vehicle information.

E. Ablation Studies

Table III shows ablation studies on the two-pedestrian, three-vehicle subset. Simply treating the vehicle center as a pseudo pedestrian body part and applying the original TBIFormer [28] increases MPJPE, APE, and FDE by roughly 2.5% each, indicating that naively treating the vehicle as a pedestrian provides insufficient cues for accurate pedestrian–vehicle interaction modeling.

Introducing a dedicated vehicle branch with conventional cross-attention but without TRPE in it substantially reduces errors, with MPJPE and FDE decreasing by about 10.7% and 13.4%, respectively, demonstrating that explicit vehicle modeling effectively improves performance. Adding TRPE within the cross-attention, thereby forming PVI-CA, brings marginal but consistent improvements under different grouping configurations of the vehicle bounding box corners. Here, *1 group* treats all 8 corners as a single group; *2 groups* split the corners into front and back faces; *4 groups* represent the four vertical edges of the bounding box; *6 groups* treat each face of the bounding box as a group; *8 groups* treat each corner individually; and *12 groups* consider each bounding box edge as a separate group. It can be seen that the performance differences across different groupings are small, and our full model adopts the *12 groups* setting, as it best captures the spatial geometry and scale of surrounding vehicles without much extra computation, though all group configurations are supported.

F. Visualization

We visualize three scenarios involving three pedestrians and four vehicles (either stationary or moving), and compare the predictions of TBIFormer [28] without vehicles and our model with vehicles. From the qualitative results in Fig. 5, we observe that both TBIFormer and our model perform well for short-duration predictions (0.2s), while our model shows clear advantages for longer-duration predictions (0.6s and 1s). TBIFormer's predictions tend to lead the ground truth by a noticeable margin, whereas our predictions more accurately follow the true trajectories. In particular, in the 1.0s prediction, TBIFormer predicts more conservative, delayed motion trends. In contrast, our model better respects the positions of surrounding vehicles and predicts more accurate motions that align more closely with the ground truth, highlighting the importance of incorporating vehicle information and explicitly modeling pedestrian–vehicle interactions.

V. Conclusion

In this work, we address the critical problem of 3D pedestrian pose forecasting in autonomous driving scenarios, emphasizing the often-overlooked influence of interactions between 3D pedestrian poses and 3D vehicles on pedestrian behavior. We enhance the Waymo-3DSkelMo dataset by incorporating 3D vehicle bounding boxes and a sampling strategy, enabling realistic modeling of pedestrian–vehicle interactions. Building on the TBIFormer architecture, we propose a Vehicle-conditioned 3D pose forecasting network, where pedestrian predictions are conditioned on not only their historical motion but also the surrounding vehicle context. Extensive experiments demonstrate that incorporating vehicle information significantly improves the accuracy of predicted pedestrian poses and validates different approaches for modeling pedestrian–vehicle interactions. Although we focus exclusively on vehicles and omit other contextual cues, such as static scene structure and other road users (e.g., cyclists), our findings indicate that vehicle context provides

impactful information beyond motion history alone, and we believe integrating richer scene context is a promising direction for future work. Our work highlights the importance of multi-agent context, particularly vehicles, in accurate pedestrian motion prediction and provides a foundation for safer autonomous driving systems.

ACKNOWLEGEMENT

Guangxun Zhu is supported by the funding from the China Scholarship Council (CSC).

REFERENCES

[1] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.

[2] J.-S. Ham, D. H. Kim, N. Jung, and J. Moon, "Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 3666–3675.

[3] L. Crosato, K. Tian, H. P. H. Shum, E. S. L. Ho, Y. Wang, and C. Wei, "Social interaction-aware dynamical models and decision-making for autonomous vehicles," *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2300575, 2024.

[4] C. Wong, B. Xia, Z. Zou, Y. Wang, and X. You, "Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19 005–19 015.

[5] K. Chen, X. Zhao, Y. Huang, G. Fang, X. Song, R. Wang, and Z. Wang, "Socialmoif: Multi-order intention fusion for pedestrian trajectory prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 465–22 475.

[6] P. Yao, Y. Zhu, H. Bi, T. Mao, and Z. Wang, "Trajclip: Pedestrian trajectory prediction method using contrastive learning and idempotent networks," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 023–77 037, 2024.

[7] 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.

[8] 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.

[9] A. Rasouli and I. Kotseruba, "Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9844–9851.

[10] M. Azarmi, M. Rezaei, and H. Wang, "Pip-net: Pedestrian intention prediction in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 9824–9837, 2025.

[11] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 186–11 196.

[12] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 206–213.

[13] ——, "Agreeing to cross: How drivers and pedestrians communicate," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE Press, 2017, p. 264–269. [Online]. Available: https://doi.org/10.1109/IVS.2017.7995730

[14] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *International Conference on Computer Vision (ICCV)*, 2019.

[15] F. Munir, S. Azam, T. Mihaylova, V. Kyrki, and T. P. Kucner, "Pedestrian vision language model for intentions prediction," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 6, pp. 393–406, 2025.

[16] J. Jeong, S. Lee, D. Park, G. Lee, and K.-J. Yoon, "Multi-modal knowledge distillation-based human trajectory forecasting," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[17] S. Saadatnejad, Y. Gao, K. Messaoud, and A. Alahi, "Social-transmotion: Promptable human trajectory prediction," in *International Conference on Learning Representations (ICLR)*, 2024.

[18] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, " Single-Shot Multi-person 3D Pose Estimation from Monocular RGB ," in *2018 International Conference on 3D Vision (3DV)*. Los Alamitos, CA, USA: IEEE Computer Society, Sept. 2018, pp. 120–130. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/3DV.2018.00024

[19] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[20] CMU, "Cmu graphics lab motion capture database," 2003. [Online]. Available: http://mocap.cs.cmu.edu/

[21] J. Jeong, D. Park, and K.-J. Yoon, "Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 975–16 984.

[22] L. Crosato, C. Wei, E. S. L. Ho, H. P. H. Shum, and Y. Sun, "A virtual reality framework for human-driver interaction research: Safe and cost-effective data collection," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 167–174. [Online]. Available: https://doi.org/10.1145/3610977.3634923

[23] G. Zhu, S. Fan, H. Dai, and E. S. L. Ho, "Waymo-3dskelmo: A multi-agent 3d skeletal motion dataset for pedestrian interaction modeling in autonomous driving," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025.

[24] "Waymo open dataset: An autonomous driving dataset," 2019.

[25] B. Fan, W. Zheng, J. Feng, and J. Zhou, "Lidar-hmr: 3d human mesh recovery from lidar," *arXiv preprint arXiv:2311.11971*, 2023.

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[27] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou, "Nemf: Neural motion fields for kinematic animation," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 4244–4256.

[28] X. Peng, S. Mao, and Z. Wu, "Trajectory-aware body interaction transformer for multi-person pose forecasting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 121–17 130.

[29] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[30] N. van der Aa, X. Luo, G. Giezeman, R. Tan, and R. Veltkamp, "Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1264–1269.

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

[32] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 2006.