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

A Novel Benchmarking Paradigm and a Scale- and Motion-Aware Model for Egocentric Pedestrian Trajectory Prediction

Amir Rasouli

Abstract— **In this paper, we present a new paradigm for evaluating egocentric pedestrian trajectory prediction algorithms. Based on various contextual information, we extract driving scenarios for a meaningful and systematic approach to identifying challenges for prediction models. In this regard, we also propose a new metric for more effective ranking within the scenario-based evaluation. We conduct extensive empirical studies of existing models on these scenarios to expose shortcomings and strengths of different approaches. The scenario-based analysis highlights the importance of using multimodal sources of information and challenges caused by inadequate modeling of ego-motion and scale of pedestrians. To this end, we propose a novel egocentric trajectory prediction model that benefits from multimodal sources of data fused in an effective and efficient step-wise hierarchical fashion and two auxiliary tasks designed to learn more robust representation of scene dynamics. We conduct empirical evaluation on common benchmark datasets and show that our model not only achieves state-of-the-art performance, but also significantly improves performance by up to 39% in challenging scenarios, such as high ego-speed, compared to the past arts[1].**

I. Introduction

Pedestrian behavior prediction requires modeling of various contextual information, such as scene dynamics, pedestrian state, etc. [1], [2]. Prediction models are often evaluated by computing their performance on the entire driving datasets. However, due to inherent biases in such datasets, e.g., prevalence of cruising in AD datasets [3] or signalized intersections in pedestrian datasets [4], and high diversity of real-world driving scenarios, such a high-level benchmarking says little about the robustness of the algorithms to challenges arising from different traffic scenes.

To address this problem, we propose a new paradigm for evaluating egocentric pedestrian trajectory prediction models. Our goal is to identify factors for extracting scenarios that expose various challenges for the prediction algorithms (see Figure 1). We propose an effective metric to rank the performance of the models in scenario-based analysis, and present an extensive evaluation of the existing models using the proposed evaluation scheme. Based on our findings from the evaluation, we propose a novel prediction algorithm that achieves state-of-the-art performance on common benchmark datasets. We show the effectiveness of our model by empirical evaluation and ablation studies on common benchmarks.

II. Related Works

A. Egocentric Pedestrian Trajectory Prediction

Pedestrian trajectory prediction is usually done either from a bird's eye view [5]–[16] or egocentric [4], [17]–[29] perspective. The former is applied to scenes recorded from a fixed surveillance camera perspective or the projection of driving scenes into a global coordinate system. Egocentric prediction involves recordings from a moving egocentric camera view, hence the scale of objects in the scenes may change and the observed motion is a combination of ego-motion and other dynamic objects' motions. To model these factors, many works rely on the apparent changes in the scale of the pedestrians, e.g., the size of bounding boxes around them, and implicitly reason about the scene dynamics [20], [21]. Other models use optical flow information [26] or train models based on pedestrian states, e.g., walking or standing [19]. However, without explicit ego-motion information, these models are limited in identifying whether the observed motion is caused by the ego-vehicle, pedestrian, or both.

Fig. 1: Factors, such as observability, state, and scale, impact pedestrian trajectory prediction in an egocentric setting.

Some models use more explicit information, such as ego-speed [4], [30] or angular velocity [31] along with contextual information, such as scene semantics and interactions [30], [31] or intentions [4], often combined in a multimodal framework that also comes with the added challenge of fusing information from different sources. In the trajectory prediction domain, common approaches include early [17] or late fusion [4], bimodal fusion [30], hierarchical fusion [32], and pair-wise cross-modal fusion [31]. However, they either lack the ability to effectively capture the correlation between different data modalities or are computationally expensive. We propose an efficient model that effectively fuses different sources of information in a *step-wise* hierarchical fashion that effectively captures the correlation between different data modalities while minimizing the computational overhead.

B. Benchmarking and Evaluation

There are many egocentric driving datasets, some of which are specifically catered to pedestrian prediction [4], [24], [26], [33]–[35]. These datasets are often collected in urban environments and provide annotations for various road

Huawei Technologies Canada amir.rasouli@huawei.com.
[1]Code is available at github.com/aras62/PIE/tree/master/scenarioEval

structures and contextual factors, e.g., signals, pedestrian behavior, etc. Given that these datasets are collected in natural settings, the recordings are often biased towards simpler driving scenarios, such as driving straight, interacting with pedestrians at signalized intersections, driving at constant speed, etc. Benchmarking on entire datasets with such characteristics, at best, provides a general ranking but reveals very little about common challenges for prediction or models' characteristics. Hence, a more detailed scenario-based evaluation paradigm is needed.

Contributions of our paper are: 1) We propose a new scenario-based paradigm for evaluating egocentric pedestrian trajectory prediction models and a new metric for better ranking of the models under each scenario; 2) We conduct extensive evaluation of existing methods on the scenarios and provide insights into common challenges for the algorithms; 3) Based on our findings, we propose a novel state-of-the-art model that takes advantage of multimodal data input and auxiliary tasks to learn more robust representation of the scene dynamics and small-scale samples, two of common challenges for egocentric prediction models; 4) Lastly, we present experimental results on the proposed model on common benchmark datasets followed by ablation studies.

III. Scenario-based Benchmark

A. Problem Formulation

We formulate egocentric trajectory prediction as an optimization problem, the goal of which is to learn a distribution $p(L|C)$ where $L = \{l_i^{t+1:t+\tau}\}$ is the trajectory of a pedestrian i, $1 < i < n$, and $C = \{c_i^{t-o+1:t}\}$ denotes observation context. Here, o and τ correspond to observation and prediction horizons, respectively. In this setting, each point is in the form of $[x_1, y_1, x_2, y_2]$ which captures the coordinates of bounding box around the pedestrian.

B. Scenario Extraction

The first step for effective benchmarking is to divide the evaluation dataset into meaningful subsets that highlight different aspects of the models. Different from the bird's eye view perspective, in the egocentric setting, object coordinates are in the image plane and change depending on motion of both the ego-camera and pedestrians.

1) Factors: To characterize the subsets of the data, we look at both pedestrians and ego-camera factors. **Pedestrian** factors are as follows: *Scale*, which reflects the proximity of pedestrians to the ego-camera. We use the height of bounding boxes as a measure of scale, since width of boxes can fluctuate due to pedestrian gait. *State*, which refers to whether the pedestrian is walking or standing. Predicting the future trajectory of a walking pedestrian in the presence of ego-motion can be challenging because there are two sources of motion combined. **Ego-camera** is described in terms of ego-*speed* (km/h) and ego-*action* (going straight/turning).

2) Changing behavior: One of the key challenges in the context of prediction is when the behavior of agents is different within the observation and prediction horizons. For instance, the pedestrian might be standing during observation period and starts walking during prediction period and vice versa. Similarly, the speed of the ego-camera can vary across observation and prediction horizons. Such changing behavior can potentially pose a challenge for prediction models.

C. Trajectory Metrics

Distance-based metrics are most common for egocentric prediction [19]–[22], where the error is calculated as the mean square error (MSE) of bounding boxes or their center coordinates. Although effective, these metrics are not adequate for scenario-based analysis since bounding box scales change depending on pedestrians' proximity to the ego-camera (see Figure 1), causing the larger scale boxes' error to dominate the overall metric value. To reflect the real-world changes, error term should be measured relative to the scale of the pedestrian. This is due to the perspective effect in the image plane, causing the same pixel difference values to correspond to very different real-world values depending on the proximity of the pedestrian to the ego-camera.

Scaled Distance Error is a new metric proposed for our scenario-based evaluation. We compute the metric by scaling the pixel error values by the average area of bounding boxes, computed based on their widths and heights.

Boundary issue. The full area of bounding boxes is not always visible due to occlusion or as a result of pedestrians entering/exiting the scene, as shown in the leftmost sample in Figure 1. Hence, simply computing areas based on the visible boxes results in incorrect scaling of certain samples. To address this issue, we first measure the average aspect ratio of height and width of the fully visible bounding boxes in the dataset. We then use this ratio to identify the samples that are partially observable and then adjust their ratio accordingly.

IV. The Benchmark

Dataset. We use the publicly available *Pedestrian Intention Estimation (PIE)* [4] dataset that contains a diverse set of labels for pedestrians and ego-vehicle. The data consists of 6 hours of recording from a monocular camera inside a moving vehicle. Following the common protocol for egocentric prediction [4], [20], [21], we extract sequences with 50% overlap divided into 0.5s for observation and 1.5s prediction. The evaluations are done using the default test set.

Scenarios. We extract scenarios based on the factors described in Sec. III-B. Using the height of pedestrians, we divide the data by scale into categories with balanced number of samples. For the state, we use walking and standing labels and consider the majority voting to characterize the observation/prediction sequences as either walking or standing. For ego-camera, we consider the vehicle's speed to characterize ego-motion, yaw angle with the threshold of 5^o change to determine turning action, and speed threshold of 5 km/h to determine whether the average speed of the ego-vehicle changes from observation to prediction.

Metrics. We report on the common metrics for egocentric trajectory prediction [4], [20]. For the sake of brevity, we only report on MSE error over bounding boxes (B_{MSE}) and its proposed scaled version denoted as sB_{MSE} averaged

TABLE I: Results on single-factor scenarios of pedestrians with different scales and states. The third row shows the number of train/test samples (x10^3) and results are reported as B_{mse}/sB_{mse}. For all values, lower is better and **best** and second best values are highlighted.

	Pedestrian Scale (pixels)							**Pedestrian State**	
	0-50	*50-80*	*80-100*	*100-150*	*150-200*	*200-300*	*300+*	*Walking*	*Standing*
	5.2/5.3	*10.8/8.6*	*4.9/4.4*	*8.2/5.7*	*5.4/4.8*	*6.5/4.6*	*3.4/2.9*	*22.4/18.3*	*20.0/16.1*
PIE$_{traj}$	187/0.308	170/0.105	373/0.122	595/0.098	709/0.058	980/0.041	2064/0.036	677/0.048	441/0.060
B-LSTM	225/0.371	215/0.132	464/0.152	840/0.138	886/0.073	1170/0.049	2492/0.044	833/0.059	582/0.080
PIE$_{full}$	195/0.322	163/0.101	313/0.102	447/0.074	513/0.042	628/0.026	1838/0.032	559/0.040	337/0.046
BiTraP	134/0.220	158/0.097	361/0.118	538/0.088	661/0.054	852/0.036	1633/0.029	579/0.041	400/0.055
PedFormer	**76/0.125**	**72/0.044**	**124/0.041**	**250/0.041**	**295/0.024**	**522/0.022**	**1363/0.024**	**394/0.028**	**153/0.021**

TABLE II: Results on single-factor scenarios for different values of ego-speed. The third row shows the number of train/test samples (x10^3) and results are reported as B_{mse}/sB_{mse}. For all values, lower is better and **best** and second-best values are highlighted.

	Ego-speed (km/h)					
	0	*0-5*	*5-10*	*10-20*	*20-30*	*30+*
	20.2/19.7	*6.8/4.8*	*5.1/2.4*	*6.7/4.0*	*3.6/3.5*	*2.0/1.8*
PIE$_{traj}$	284/0.020	800/0.071	1495/0.136	1153/0.189	709/0.208	649/0.320
B-LSTM	367/0.026	884/0.079	1849/0.168	1405/0.231	954/0.280	969/0.477
PIE$_{full}$	259/0.019	659/0.059	939/0.085	976/0.160	563/0.165	317/0.156
BiTraP	269/0.019	704/0.063	1221/0.111	1018/0.167	597/0.175	386/0.190
PedFormer	**225/0.016**	**390/0.035**	**487/0.044**	**467/0.077**	**272/0.080**	**225/0.111**

over the entire prediction horizon. The remaining metrics are released with code.

Models. Our goal is to highlight the differences between models that rely on different architectures, learning methods and contextual information. For this purpose, we report on the following models: **PIE**$_{traj}$ [4] which is a basic recurrent encoder-decoder architecture that only uses bounding box coordinates; **PIE**$_{full}$ which is an extension of PIE$_{traj}$ that also incorporates pedestrian intention and ego-speed; **B-LSTM** [28], a recurrent architecture with Bayesian weight learning method; **BiTraP** [21], a conditional variational (CVAE) model that uses pedestrian goals and a bidirectional decoder for prediction; and **PedFormer** [31], a multitasking framework with hybrid Transformer-recurrent architecture[2].

A. Single-factor Scenarios

Here, we extract scenarios only based on a single factor, namely the scale of pedestrians, their state, and ego-speed. For pedestrian states, we only report on the cases where the state of the pedestrians is the same across observation and prediction for all models. The results are shown in Table I.

When considering the absolute error, categories corresponding to larger scales appear worse. However, once the error is scaled, we can see that relatively speaking, prediction of smaller scale pedestrians are more challenging. Although performance degradation happens for all models, the rate of change varies and results in different ranking across different scenarios for PIE$_{full}$ and BiTraP.

As expected, pedestrian state also impacts the performance. Intuitively, one would expect scaled error to be lower when the pedestrian is standing. However, this is not the case for most models due to other confounding factors that are not considered under state scenarios. This means that further breakdown of scenarios with additional factors is needed to get a better understanding of the performance variation.

[2]Since trajectory samples are extracted over the entire tracks (not up to crossing events), we drop the action prediction branch from PedFormer and only predict trajectories and grid locations.

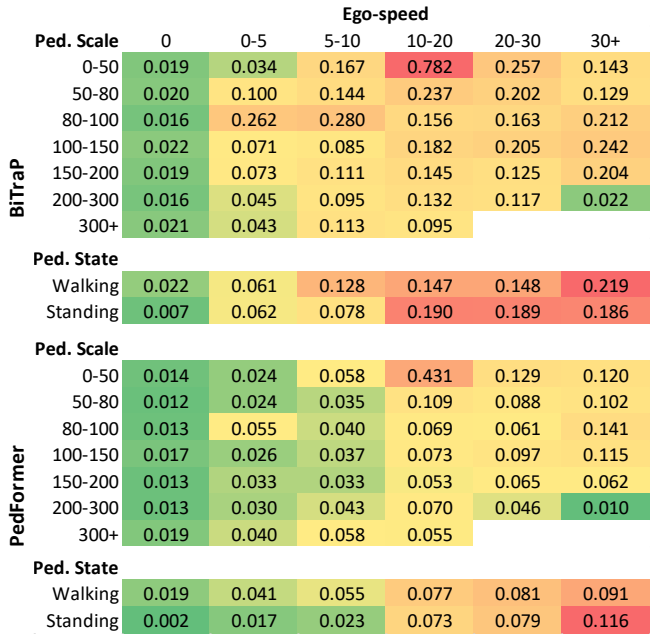

Fig. 2: Two-factor scenario-based evaluation showing pedestrian scale and state vs ego-speed using sB_{MSE}. Generally, increase in ego-speed has a negative impact on performance, however, with different intensity depending on the pedestrian's scale and state.

A similar trend of changes in performance can be seen in different speed categories, as shown in Table II. Here, we can see that at speed 0, where no motion is present, the top three models perform within a similar range. As the ego-speed increases, the gap between the models with ego-motion modeling and others increases significantly. It should be noted that the way ego-motion is modeled is also important. For instance, PIE$_{full}$, which only uses speed to model ego-motion, performs similar to BiTraP that does not use any ego-motion. Whereas, PedFormer achieves much better results by using angular velocity in addition to speed.

B. Two-factor Scenarios

As mentioned earlier, single-factor analysis is not always sufficient. Hence, we extract scenarios based on combinations of two factors, namely scale and ego-speed, as well as state and ego-speed. Following the previous experiment, since our focus is mainly on scale changes and due to space limitations, we report the results only on sB_{MSE} metric.

The results are illustrated in Figure 2. As expected, the performance of both models degrades as the speed increases (row-wise) and improves as scale increases (column-wise). The rate of degradation, however, is different. For BiTraP,

TABLE III: Benchmark results on challenging scenarios. W and S refer to walking and standing respectively and the transition of one state to other from observation $_o$ to prediction $_p$. The third row indicates the number of train/test samples (x10^3) and results are reported as B_{mse}/sB_{mse}. For all values, lower is better and **best** and second-best values are highlighted.

	Pedestrian State		Ego-motion		Ego-action	
	W_o-S_p	S_o-W_p	Constant	Change	Straight	Turn
	1.2/1.0	*1.0/0.8*	*41.5/34.2*	*3.0/2.0*	*41.3/34.7*	*3.2/1.5*
PIE$_{traj}$	536/0.064	1553/0.108	474/0.043	2576/0.236	436/0.040	4166/0.430
B-LSTM	513/0.061	1847/0.128	593/0.054	3213/0.295	542/0.049	5274/0.545
PIE$_{full}$	376/0.045	1167/0.081	402/0.037	1636/0.150	338/0.031	3547/0.366
BiTraP	499/0.059	1142/0.079	430/0.039	1899/0.174	383/0.035	3492/0.361
PedFormer	**252/0.030**	**970/0.067**	**259/0.024**	**924/0.085**	**248/0.023**	**1422/0.147**

which does not explicitly model ego-motion, the performance drops significantly even when the ego-speed is low.

Similar degradation is observable in the state vs. speed case. An interesting finding is that the ranking between walking and standing, within each speed category, differs across models. For PedFormer, walking samples are more challenging, as one would expect, all the way up to 30+ speed category where the ranking flips. For BiTraP, on the other hand, the ranking switches once at 10-20 category and again at 30+. One potential reason can be the relative motion of walking pedestrians and ego-vehicle. As the speed gets higher, pedestrian motion may appear smaller or even constant in the image plane. Consequently, in the case of PedFormer, performance degradation is reduced as the model relies more on the ego-motion. This also applies to BiTraP but only up to a degree, hence its ranking fluctuates more.

In two-factor evaluation, we see some anomalies that do not follow the overall trends. For instance, for both BiTraP and PedFormer, there is a significant performance drop at scale 0-50 and speed 10-20. Such drastic changes can be due to additional factors in those categories, e.g., pedestrian state, ego-vehicle action, or abrupt changes between observation and prediction. To shed more light on it, one can use additional factors to extract finer scenario categories. However, it may lead to a longtail effect, i.e., small number of training samples in each scenario category. For instance, in the two-factor scenarios, there are no instances of high-speed driving and large scales, and in some categories the number of samples is below 100. It should also be noted that some anomalies are model-specific, e.g., BiTraP unexpectedly does very well at scale 0-50 and speed 0-5.

C. Challenging Scenarios

Table III summarizes the results of experiments on challenging scenarios. As discussed earlier, changes in the pedestrian state and the ego-motion can have a significant negative impact on prediction. One interesting observation here is the large difference between $W_o - S_p$ and $S_o - W_p$ cases. The reason for this can be that generally in $W_o - S_p$ cases, the pedestrian is slowing down, indicating the possibility of standing in the near future. In $S_o - W_p$ cases, however, there is no motion history to hint to the possibility of walking in the future. Thus, the models need to better understand the context (besides dynamics) to infer pedestrians' future state.

Lastly, we can see a drastic difference between ego-vehicle moving straight vs. turning. Turn actions often generate irregular motion patterns, which are difficult to estimate without a clear understanding of the road structure or the state of other agents. Including information, such as angular velocity, can help improve the results (as in PedFormer) but is still not sufficient for an accurate estimation.

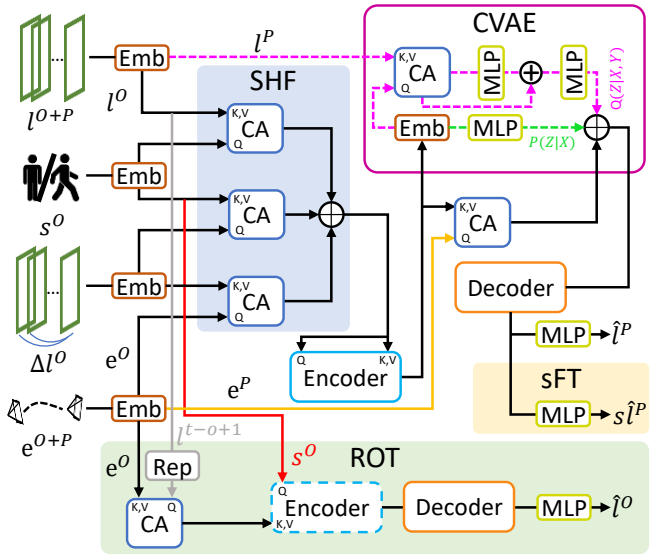

Fig. 3: Overview of the proposed approach. The model receives four inputs, which are combined via the step-wise hierarchical fusion (SHF) module. SHF's output is fed to the encoder, followed by the CVAE module and decoder to predict both future trajectories $\hat{l}^P$ and the auxiliary scaled future trajectory (sFT) $s\hat{l}^P$. For reconstruction of observed trajectories (ROT), ego-motion e^O and the first pedestrian location l^{t-o+1} (repeated to the size of observation) are fused and fed into the encoder, in addition to state s^O, followed by the decoder to produce constructed observation $\hat{l}^O$. Pink and green dashed lines indicate the flow during training and test time, respectively. Here, $O = t-o+1:t$ and $P = t+1:t+\tau$.

V. Model

Based on the findings from the scenario-based analysis, here we propose a novel model for egocentric prediction. In this model, we emphasize modeling dynamic factors to improve the overall performance, as illustrated in Figure 3. The proposed model is a fully attention-based architecture divided into three main modules: a multimodal encoder, decoders, and a reconstructor for observed trajectory.

A. Multimodal Encoder

For capturing scene dynamics, the model relies on a set of multimodal inputs, namely pedestrian normalized locations, states (actions) as an indicator of whether pedestrian is moving, velocity computed based on the changes in location of pedestrian at time t, and the ego-motion, which includes GPS coordinates, acceleration, speed, and angular velocity.

As demonstrated by past arts [30]–[32], the manner in which the multimodal data is processed is important. In transformer-based models, besides early or late fusion techniques, hierarchical [32] and cross-modal fusion methods have been used [31]. In the hierarchical method, different modalities are processed using different attention heads, the output of which are combined via a cross-attention operation. In cross-modal fusion, pair-wise cross-attention units are used to better capture the cross-correlation between different data modalities. However, this approach is computationally

expensive as it requires $m(m-1)$ attention modules where m is the number of data modalities. Here, we propose a **step-wise hierarchical fusion (SHF)** approach in which different data inputs are gradually fused one at the time by cross-attention (CA) modulation. In this way only $m-1$ attention units are needed while the relation between different data types are captured effectively. The outputs of the CA units are concatenated and fed into the encoder transformer.

To model uncertainty, we use a conditional variational autoencoder (CVAE), which generates a latent distribution $p(z|x)$ by learning its similarity to a conditional distribution $q(z|x,y)$ at training time. The combination of the k samples drawn from the latent distribution and the input encodings form the input into the trajectory decoder.

B. Decoder

The decoder module has a similar architecture to the encoder, with additional masking operation for prediction. Assuming that the future ego-motion is based on the planned behavior at the time t, the encodings are fused with the future ego-motion features before being fed to the decoder. The final output of the decoder is fed into multilayer perceptron layers (MLPs) to infer future trajectories (FT).

1) Auxiliary scaled future trajectory (sFT): As in Sec. IV-A, computing absolute error would bias the error term towards larger scale samples. To reduce bias against smaller scale samples, we add an auxiliary task for predicting scaled bounding box coordinates.

C. Reconstruction of Observed Trajectory (ROT)

A key challenge for the egocentric prediction is to separate observed motion resulting from the agent or ego-motion. For this, we add a reconstruction task primarily based on ego-motion with a reference to bounding box coordinate at time $t-o+1$ and pedestrian state, i.e., walking or standing. The state is used as a signal to indicate whether pedestrian motion is contributing to the observed motion in the image plane. We combine this information via cross-modal attention modulation, where ego-motion and bounding box coordinates serve as a value and the state as a query. The output is fed to the same encoder, followed by a separate decoder to generate observed bounding boxes. The reason for reconstructing observation instead of predicting future is twofold: first, observation is shorter and therefore the propagation of error is less, and second, reconstructing observation without a full context can serve as an additional supervision signal which is different to future prediction objective.

D. Objective Function

We use a separate loss for each regression task, namely L_{FT}, L_{sFT} and L_{ROT} which correspond to losses for future trajectories, scaled future trajectories, and reconstructed observed trajectories. An additional KL-divergence (KLD) loss is added for learning the latent distribution. For regression losses, we use LogCosh and minimize the losses in a "best of many" fashion, e.g., $min_k \sum f(y^{t:t+\tau}, \hat{Y}_k^{t:t+\tau})$. The final loss is a weighted combination of losses as follows:

$$L = L_{FT} + \alpha L_{sFT} + \beta L_{ROT} + \gamma KLD,$$

ENCORE-D	Ego-speed					
Ped. Scale	0	0-5	5-10	10-20	20-30	30+
0-50	0.014	0.023	0.052	0.280	0.119	0.079
50-80	0.012	0.023	0.029	0.077	0.078	0.065
80-100	0.012	0.035	0.035	0.057	0.059	0.090
100-150	0.015	0.023	0.026	0.053	0.068	0.072
150-200	0.011	0.029	0.028	0.049	0.047	0.044
200-300	0.012	0.020	0.037	0.081	0.042	0.010
300+	0.017	0.036	0.039	0.063		
Ped. State						
Walking	0.017	0.033	0.042	0.090	0.068	0.075
Standing	0.002	0.014	0.016	0.047	0.060	0.071

Fig. 4: Two-factor scenario-based evaluation showing pedestrian scale and state vs. ego-speed using sB_{MSE}. Although similar performance degradation as other models across different ego-speed values is observed, ENCORE-D is affected significantly less.

where weights α, β, and γ are set empirically.

VI. Evaluation

Implementation. We set all input embedding layers to 64 and rest to 128. CVAE MLPs are two layers with 256 and 128, respectively. We use two attention heads for all attention modules in CAs and transformers. We used a Butterworth low-pass filter with order of 4 and frequency of 5 for acceleration and angular velocity data. Loss weights are set empirically to 10, 0.5, and 0.1 for α, β, and γ respectively. For adjustment, we use width/height ratio of 0.34. The model is trained for 105 epochs, with a batch size of 64, learning rate of 4×10^{-4}, and Adam optimizer. We refer to our model as **ENCORE** (**E**goce**N**tric predi**C**ti**O**n with **RE**construction).

A. Scenario-based Analysis

To be comparable to the previous models, we report the results on the deterministic version of our model (ENCORE-D) in this section.

Single-factor. In designing ENCORE, we pursued three objectives: better learning of smaller scale samples, distinguishing between sources of motion, and modeling ego-motion. As the results in Table IV suggest, these objectives have been achieved. We can observe significant improvement in smaller scale scenarios, by up 27%. In the case of state, there is a significant drop, 29% in the error of *standing* scenarios due to the use of explicit state encoding. Lastly, we can see that as the overall speed of the vehicle increases, so does the performance gap between ENCORE and PedFormer, reaching the maximum of 36% at 30+ km/h. On other metrics, such as CF_{MSE}, which measure final trajectory error, even more improvement can be achieved — up to 37% and 40% in scale and speed scenarios, respectively.

Two-factor. As shown in Figure 4, ENCORE, performs better across different scenarios with smoother transition across speed dimension. More notable improvements are achieved on smaller scales, 0-100, by up to 37% at low speed 0-5, and 35% at mid-range speed of 10-20. There is also significant improvement in state cases, in particular Standing scenarios. The improvement ratio increases with the ego-speed, peaking at 39%.

Challenging Scenarios. A similar pattern of improvement can be observed across different challenging scenarios, with

TABLE IV: Comparison to SOTA on single factor scenarios of pedestrians with different scales and states using B_{MSE}/sB_{MSE} metrics. For all values, lower is better and improvements at the end are computed with respect to PedFormer.

Scenario	Pedestrian Scale (pixels)							Pedestrian State		Ego-speed (km/h)					
Sub-scenario	**0-50**	**50-80**	**80-100**	**100-150**	**150-200**	**200-300**	**300+**	**Walking**	**Standing**	**0**	**0-5**	**5-10**	**10-20**	**20-30**	**30+**
BiTraP	134/0.220	158/0.097	361/0.118	538/0.088	661/0.054	852/0.036	1633/0.029	579/0.041	400/0.055	269/0.019	704/0.063	1221/0.111	1018/0.167	597/0.175	386/0.190
PedFormer	76/0.125	72/0.044	124/0.041	250/0.041	295/0.024	522/0.022	1363/0.024	394/0.028	153/0.021	225/0.016	390/0.035	487/0.044	467/0.077	272/0.080	225/0.111
ENCORE-D (ours)	**55/0.091**	**58/0.036**	**99/0.032**	**189/0.031**	**257/0.021**	**485/0.020**	**1165/0.020**	**351/0.025**	**108/0.015**	**200/0.014**	**320/0.028**	**377/0.034**	**419/0.069**	**217/0.064**	**145/0.071**
Improvement	27%	20%	21%	24%	13%	7%	15%	11%	29%	11%	18%	23%	10%	20%	36%

TABLE V: Comparison to SOTA on challenging scenarios using B_{MSE}/sB_{MSE} metrics. For all values, lower is better.

	Pedestrian State		**Ego-motion**		**Ego-action**	
	W_o-S_p	S_o-W_p	**Constant**	**Change**	**Straight**	**Turn**
BiTraP	499/0.059	1142/0.079	430/0.039	1899/0.174	383/0.035	3492/0.361
PedFormer	252/0.030	970/0.067	259/0.024	924/0.085	248/0.023	1422/0.147
ENCORE-D (ours)	**222/0.026**	**877/0.061**	**225/0.021**	**707/0.065**	**217/0.020**	**1055/0.109**

TABLE VI: Comparison to SOTA on the PIE dataset. For all values, lower is better and **best** and second-best values are highlighted.

	B_{MSE}			C_{MSE}	CF_{MSE}	sB_{MSE}
	0.5s	**1s**	**1.5s**	**1.5s**	**1.5s**	**1.5s**
B-LSTM [28]	101	296	855	811	3259	0.078
FOL-X [26]	47	183	584	546	2303	0.053
PIE$_{traj}$ [4]	58	200	636	596	2477	0.058
PIE$_{full}$ [4]	-	-	556	520	2162	0.051
BiTraP-D [21]	41	161	511	481	1949	0.047
PEV [22]	42	153	453	418	1683	0.041
SGNet-ED [20]	**34**	133	442	413	1761	0.040
BiPed [30]	37	119	320	291	1104	0.029
PedFormer [31]	38	118	295	265	943	0.027
ENCORE-D (ours)	37	**102**	**251**	**222**	**806**	**0.023**
BiTrap-NP(20) [21]	23	48	102	81	261	0.009
SGNet-ED(20) [20]	16	39	88	66	206	0.008
ABC+(20) [19]	16	38	87	65	191	0.008
ENCORE(20) (ours)	**15**	**33**	**70**	**49**	**155**	**0.006**

TABLE VII: Comparison to SOTA on the JAAD dataset. For all values, lower is better and **best** and second-best values are highlighted.

	B_{MSE}			C_{MSE}	CF_{MSE}	sB_{MSE}
	0.5s	**1s**	**1.5s**	**1.5s**	**1.5s**	**1.5s**
B-LSTM [28]	159	539	1535	1447	5617	0.122
FOL-X [26]	147	484	1374	1290	4924	0.110
PIE$_{traj}$ [4]	110	399	1280	1183	4780	0.102
PIE$_{full}$ [4]	-	-	1208	1154	4717	0.096
BiTraP-D [21]	93	378	1206	1105	4565	0.096
PEV [22]	97	373	1158	1042	4471	0.092
BiPed [20]	85	362	1202	1147	4759	0.096
PedFormer [31]	93	364	1134	1080	4364	0.090
SGNet-ED [20]	**82**	328	1049	996	4076	0.084
ENCORE-NR-D (ours)	83	**319**	**980**	**930**	**3766**	**0.078**
BiTrap-NP(20) [21]	38	94	222	177	565	0.018
SGNet-ED(20) [20]	37	86	197	146	443	0.016
ABC+(20) [19]	40	89	**189**	**145**	**409**	**0.015**
ENCORE-NR(20) (ours)	**32**	**85**	210	167	554	0.017

TABLE VIII: Ablation study results on proposed modules.

HSF	**sFT**	**POFT**	**ROT**	B_{MSE}	C_{MSE}	CF_{MSE}	sB_{MSE}
				90	66	248	0.008
✓				80	58	232	0.007
✓	✓			77	56	198	0.007
✓	✓	✓		79	56	222	0.007
✓	✓		✓	**70**	**49**	**155**	**0.006**

more improvement, up to 26%, achieved on ego-motion and action. This indicates that ENCORE models the ego-motion more effectively compared to the past arts. The improvement on pedestrian state is more modest, pointing to the need for more contextual information, e.g., pedestrian pose, signal state, group dynamics, etc. which can be used to deduce future changes in state.

B. Comparison to SOTA

We evaluate the proposed approach on PIE [4] (introduced earlier) and JAAD [33], a commonly used dataset for egocentric trajectory prediction. Note that JAAD does not contain any ego-motion information, hence a modified version of our model with no reconstruction (NR) is evaluated. In addition to the models introduced earlier, we report the results on the following models: FOL-X [26], PEV [22], BiPed [30] (from which, similar to PedFormer, we remove the action prediction task), SGNet-ED [20], and ABC+ [19]. As for metrics, we report on B_{mse}, sB_{mse}, C_{mse}, and CF_{mse}.

Tables VI and VII summarize the results of our model on both datasets. On PIE, both versions of our model significantly improve upon the past arts on most metrics, ENCORE-D improves performance by up to 16% compared to PedFormer and ENCORE by up to 25% compared to ABC+. On JAAD, improvement of up to 8% is achieved on the deterministic model, while the non-deterministic model improves on two metrics (by up to 13%) and achieves comparable results to the past arts on the rest. One reason for a smaller gap is that the majority of JAAD samples are small-scale, therefore the adjustment offers limited improvement and its effect is lowered in non-deterministic modeling.

Ablation Study. We examine the effectiveness of the proposed modules, namely step-wise hierarchical fusion (SHF) (compared to the approach in [31]), scaled future trajectories (sFT), and reconstruction of observed trajectory (ROT). We also report on using partial observation for future trajectories (POFT) which is similar to ROT but generates future trajectories instead of reconstructing the observation. As shown in Table VIII, SHF results in moderate improvement across all metrics by providing a better fusion of multimodal inputs. sFT, as an auxiliary task, further improves the performance (especially on final displacement error) by minimizing the error propagation of partially observable boxes as well as reducing error bias during training. Lastly, the reconstruction module improves upon on all metrics, with more significant impact on the final error metric. This is due to the regularization effect that reconstruction has to minimize error propagation, something that is often achieved with explicit goal setting in the past arts [20], [21], [31]. As expected, predicting future trajectories based on partial observation, instead of reconstructing input, does not provide any benefits.

VII. Conclusion

We proposed a new approach to evaluating egocentric pedestrian trajectory prediction models based on various contextual factors extracted from data. We highlighted a number of challenges stemming from a diverse set of factors related to pedestrians and the ego-agent. Based on our findings, we proposed a novel model, ENCORE, which significantly improves upon the past arts. Through scenario-based analysis using the newly proposed framework and ablation studies, we showed how effective our approach is in resolving common challenges in egocentric prediction.

REFERENCES

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *Intelligent Vehicles Symposium (IV)*, 2017.

[2] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[3] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, 2019.

[4] 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 *ICCV*, 2019.

[5] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *CVPR*, 2021.

[6] P. Dendorfer, S. Elflein, and L. Leal-Taixe, "Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *ICCV*, 2021.

[7] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human trajectory prediction via conditional 3d attention," in *CVPR*, 2021.

[8] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *CVPR*, 2020.

[9] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *CVPR*, 2020.

[10] J. Sun, Q. Jiang, and C. Lu, "Recursive social behavior graph for trajectory prediction," in *CVPR*, 2020.

[11] H. Sun, Z. Zhao, and Z. He, "Reciprocal learning networks for human trajectory prediction," in *CVPR*, 2020.

[12] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *ECCV*, 2020.

[13] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *ICCV*, 2019.

[14] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *CVPR*, 2019.

[15] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *CVPR*, 2019.

[16] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[17] H. Damirchi, M. Greenspan, and A. Etemad, "Context-aware pedestrian trajectory prediction with multimodal transformer," *arXiv:2307.03786*, 2023.

[18] J. Li, X. Shi, F. Chen, J. Stroud, Z. Zhang, T. Lan, J. Mao, J. Kang, K. S. Refaat, W. Yang, *et al.*, "Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints," in *ICRA*, 2023.

[19] M. Halawa, O. Hellwich, and P. Bideau, "Action-based contrastive learning for trajectory prediction," in *ECCV*, 2022.

[20] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *RAL*, vol. 7, no. 2, pp. 2716–2723, 2022.

[21] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation," *RAL*, vol. 6, no. 2, pp. 1463–1470, 2021.

[22] L. Neumann and A. Vedaldi, "Pedestrian and ego-vehicle trajectory prediction from monocular camera," in *CVPR*, 2021.

[23] O. Makansi, O. Cicek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *CVPR*, 2020.

[24] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," in *CVPR*, 2020.

[25] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *CVPR*, 2018.

[26] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *ICRA*, 2019.

[27] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IROS*, 2019.

[28] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018.

[29] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *CVPR*, 2019.

[30] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *ICCV*, 2021.

[31] A. Rasouli and I. Kotseruba, "Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning," in *ICRA*, 2023.

[32] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *ICRA*, 2023.

[33] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017.

[34] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, "Loki: Long term and key intentions for trajectory prediction," in *ICCV*, 2021.

[35] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *RAL*, vol. 5, no. 2, pp. 3485–3492, 2020.