PSI: A Benchmark for Human Interpretation and Response in Traffic Interactions - Supplementary

Taotao Jing[1] **Tina Chen**[2] **Renran Tian**[3]* **Yaobin Chen**[2] **Joshua Domeyer**[4]
Heishiro Toyoda[4] **Rini Sherony**[4] **Zhengming Ding**[1]
[1]Tulane University [2]Purdue University
[3]North Carolina State University [4]Toyota Motor North America
{tjing, zding1}@tulane.edu cchen.tina@gmail.com rtian2@ncsu.edu chen62@purdue.edu
{joshua.domeyer, rini.sherony}@toyota.com heishiro.toyoda.tmc@tri.global

1 The Proposed Model

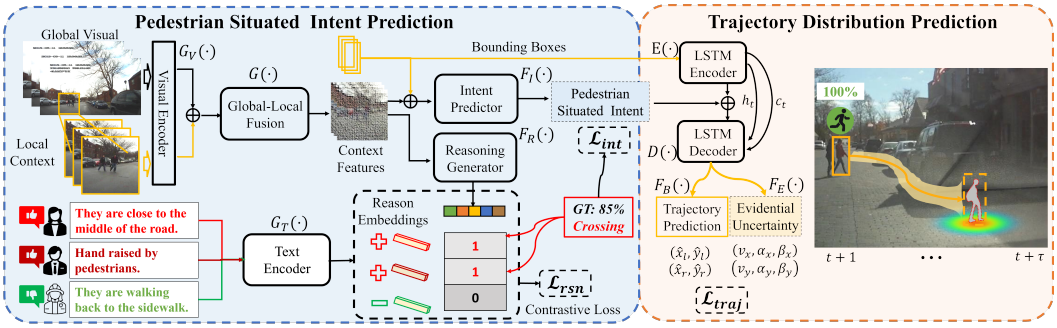

Figure 1: Illustration of the proposed framework, where video signals and language-based annotation will be used to predict the pedestrian trajectory and intention associated with uncertainty.

Given the proposed dataset, we will perform three tasks including pedestrian intent prediction, corresponding reasoning estimation, and pedestrian trajectory prediction, to demonstrate the potential tasks that can be completed with the dataset and the use cases of the novel labels.

The proposed **eP2P** is illustrated in Figure 1, which consists of two modules, *Pedestrian Situated Intent Prediction* and *Trajectory Distribution Prediction*. The intent prediction module consists of a visual encoder $G_V(\cdot)$ and a text encoder $G_T(\cdot)$ to extract the corresponding features from the input video and textual reasoning annotations with, as well as a global-local context feature fusion network $G(\cdot)$ to aggregate the global information of the whole scene and local cues from the region near the target pedestrian. Besides, an intent predictor $F_I(\cdot)$ and a reasoning generator $F_R(\cdot)$ are deployed to predict the crossing intent and the corresponding reason for the prediction, respectively. For the trajectory prediction module, an encoder $E(\cdot)$ and decoder $D(\cdot)$ are used to integrate the spatial-temporal information of the observed moving location of the pedestrian and the crossing intent estimation. Subsequently, the trajectory predictor $F_B(\cdot)$ predicts the future locations of the target pedestrian, while the uncertainty predictor $F_E(\cdot)$ estimates hyper-parameters for the model uncertainty.

To better utilize the PSI benchmark, the lower-level visual annotations and the cognitive annotations are aggregated to predict pedestrian intention as well as the future pedestrian positions with human reasoning descriptions and visual scene understanding. Considering the strong relationship between intention and behaviors

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Mathematically, the goal of the proposed framework **eP2P** is to estimate the conditional distribution $p(\mathcal{P}_{t+1}^{t+\tau}|\mathcal{O}_{t-m}^{t})$ for each pedestrian i, where $1 \leq i \leq N$ and N is the total number of pedestrians in the dataset. Specifically, $\mathcal{O}_{t-m}^{t}$ consists of the observed sequence of m frames ($\mathbf{v}^j$) and the pedestrian's trajectory ($\mathbf{l}^j$) in the time period $j \in (t-m, t]$. Besides, $\mathcal{P}^j$ includes the target pedestrian's situated intent $c^t \in \{0, 1\}$ estimated at time t, where 0 denotes "not crossing" and 1 is "crossing", and bounding boxes $\mathbf{l}^j = [(x_l, y_l), (x_r, y_r)|_j]$ in the future time steps $j \in [t+1, t+\tau]$.

1.1 Pedestrian Intention and Reasoning Prediction

The crossing intention prediction is addressed as a binary-classification problem. In order to leverage the rich knowledge of both the global environment (e.g., road conditions, traffic elements, etc.) and local appearance (e.g., pose, motion, etc.) of the target pedestrian, the local region surrounding the target pedestrian is cropped out and inputted to a Transformer visual encoder $G_V(\cdot)$, in addition to the complete observed image, to extract features. The extracted features of the two types of input images contain the local and global knowledge, respectively, which are then fused through the global-local knowledge fusion module $G(\cdot)$ to obtain the visual contextual features. With the visual contextual features and the location coordinates of the pedestrian as input, the situated intent estimation and the corresponding reasoning for the estimation are predicted by $F_I(\cdot)$ and $F_R(\cdot)$, respectively. The predicted intent is supervised by the ground-truth crossing intent as:

$$\mathcal{L}_{int} = \frac{1}{N} \sum\nolimits_{i=1}^{N} (1 - d_i) \cdot \mathbf{BCE}(\bar{c}_i, \hat{c}_i), \tag{1}$$

where $\mathbf{BCE}(\cdot, \cdot)$ defines the binary cross-entropy loss, $\hat{c}_i$ is the predicted crossing intention for pedestrian i, and $\bar{c}_i$ is the ground-truth. Specifically, we accept the intention from {"crossing", "not crossing" } with the majority agreement by all annotators as the ground-truth $\bar{c}_i$, and the ratio of observers' difference with $\bar{c}_i$ as the disagreement score $d_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{1}_{c_i^k \neq \bar{c}_i}, d_i \in [0, 1]$, where n_i is the total number annotators for the current case, c_i^k is the crossing intention annotated by the specific annotator $1 \leq k \leq n_i$. Utilizing the disagreement score to reweigh the learning objective can mitigate the distraction caused by uncertain situations, in which even human drivers struggle to make decisions effortlessly.

Moreover, we notice that human drivers used to estimate pedestrians' behavior through visual observation and some common social knowledge. Thus, we aim to bridge the gap between visual observation and the human reasoning process. Specifically, based on global-local contextual visual observation, another reasoning generator module $F_R(\cdot)$ is deployed to generate reasoning for the crossing intent prediction. This process is guided by ground-truth reasoning annotations. To achieve this, the ground-truth reasoning annotations from annotators are encoded by a Transformer text encoder $G_T(\cdot)$. The output embeddings are then used to supervise reasoning prediction via contrastive loss. Given the intent and reasoning annotations from multiple annotators, the agreement on intent is obtained through majority voting. Intent annotations that match the agreement are considered positive cases, while those differing from the agreement are deemed negative cases. We adhere to state-of-the-art visual-language model training strategies to align ground-truth and predicted reasoning embeddings

1.2 Trajectory Prediction

The eventual goal is predicting the future trajectory of the target pedestrian in the following $t+1 \sim t+\tau$ time steps based on the past $t-m \sim t$ observations. Instead of predicting a deterministic trajectory as

Moreover, we define the position of a target pedestrian at every time step as the center of the bounding box, which is denoted as $\mathbf{l}_c = (l_x, l_y)$, and $l_x = \frac{x_l + x_r}{2}, l_y = \frac{y_l + y_r}{2}$. To model the prediction uncertainty, we assume $\mathbf{l}_c$ is drawn from a bivariate Gaussian distribution $\mathbf{l}_c \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_x; \mu_y)$ is the mean and $\boldsymbol{\Sigma} = \mathbf{diag}(\sigma_x^2, \sigma_y^2)$ is the variance. Since only a single sample is provided in the training data while the ground-truth distribution of the target trajectory is unknown and unobservable, we place evidential priors to model the prediction uncertainty. Specifically, for each prediction of $l_x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $l_y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ of the pedestrian center position, we follow

To simultaneously maximize the model evidence in support of the observations in the training data and inflating uncertainty when the prediction is wrong, the model is optimized with the learning

objective as:

$$\mathcal{L}_{evi}(\boldsymbol{w}) = \mathcal{L}_{NLL}(\boldsymbol{w}) + \mathcal{L}_{R}(\boldsymbol{w}), \tag{2}$$

where $\mathcal{L}_{NLL}(\boldsymbol{w}) = \frac{1}{2}\log(\frac{\pi}{v}) - \alpha\log(\Omega) + (\alpha+\frac{1}{2})\log((l-\gamma)^2 v + \Omega) + \log(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})})$ is the negative logarithm of model evidence, $\mathcal{L}_{R}(\boldsymbol{w}) = |l-\gamma| \cdot (2v+\alpha)$ is the evidence regularizer which imposes an incorrect evidence penalty to minimize evidence on incorrect predictions, in which $\Omega = 2\beta(1+v)$, and l is short for l_x/l_y for x and y dimensions, respectively. We apply Eq. 2 to estimate x and y independently and define the pedestrian trajectory prediction objective with evidential uncertainty as:

$$\mathcal{L}_{traj} = \frac{1}{2N\tau}\sum_{i=1}^{N}\sum_{j=t+1}^{t+\tau}\Big(\mathcal{L}_{evi}(\boldsymbol{w}_x^{ij}) + \mathcal{L}_{evi}(\boldsymbol{w}_y^{ij})\Big), \tag{3}$$

where $\boldsymbol{w}_{x/y} = \{\gamma_{x/y}, v_{x/y}, \alpha_{x/y}, \beta_{x/y}\}$ are the estimated parameters for the NIG distribution on x and y dimensions, respectively. $\{v_{x/y}, \alpha_{x/y}, \beta_{x/y}\}$ are estimated by a neural network $F_E(\cdot)$, and $\gamma_{x/y} = \hat{l}_{x/y}$ is calcualted based on the bounding boxes predicted by $F_B(\cdot)$, where $(\hat{l}_x, \hat{l}_y)$ is the center of the predicted bounding box.

1.3 Overall Objective

To sum up, we integrate all objectives to formulate our overall loss function for the proposed model as $\mathcal{L} = \mathcal{L}_{int} + \mathcal{L}_{rsn} + \mathcal{L}_{bbox} + \mathcal{L}_{traj}$.

2 Comparison with Existing Datasets

Several recent datasets have advanced behavior understanding in autonomous driving, but key gaps remain. JRDB-Act [2] extends JRDB [7] with fine-grained activity labels for 3D behavior recognition, but focuses on classification and detection without modeling reasoning.

Large-scale datasets like Waymo Open Motion v2 [3] and nuPlan [1] support long-horizon planning but do not capture subjective or interpretable human intent. Language-based datasets such as Reason2Drive [9], DriveLMM [5] and DriveLM [11] explore reasoning in driving scenarios, but often rely on synthetic data and structured tasks, targeting lower-level inferences.

In contrast, PSI focuses on the socially complex challenge of pedestrian intent understanding, offering real-world video with 10–24 annotators per scenario to capture inter- and intra-human variability. By modeling multi-human reasoning chains, disagreement, and contextual cues, PSI fills a critical gap in the current landscape. PSI provides real-world video data with rich, free-text reasoning from 10–24 annotators per scenario. It captures inter- and intra-human variability, enabling research into ambiguity, consensus, and social cognition—critical for human-aligned autonomous systems.

More of recent advances in explainable autonomous driving, particularly those integrating vision-language reasoning and human interaction understanding. For example, 3D-LLM-Autonomous-Driving (OmniDrive) [12] enables multimodal reasoning and counterfactual queries in 3D driving scenes, while Awesome-LLM4AD [13] curates a broad collection of LLM-based research across perception, planning, and decision-making. The HRI Advice Dataset (HAD) [6] focuses on natural language guidance for AVs, enhancing transparency in control. nuScenes-QA [10] extends nuScenes with over 460K visual QA pairs for evaluating perception and reasoning in complex urban environments. Additionally, works like PedVLM [8] and [4] explore vision-language models for pedestrian behavior prediction, highlighting the potential of language-grounded reasoning.

In contrast, our dataset uniquely captures both driver and pedestrian intentions, includes explicit reasoning annotations, and models annotator disagreement—offering a rich lens into the subjective and interactive nature of road user behavior. This enables more nuanced modeling of intent and decision-making, especially in ambiguous or contested scenarios.

3 Distrubtion of the Annotators

Our annotator pool includes 74 individuals (44 male, 30 female), aged 19–77, evenly distributed across three age groups (19–30, 31–54, 55+). All hold valid U.S. driver's licenses and represent a

range of driving experience: 25% drive 5,000 miles/year, 40% 10,000 miles/year, and the rest over 15,000 miles/year.

This diversity in age, gender, and driving experience provides a broad range of perspectives in pedestrian intent estimation and reasoning. However, we acknowledge the U.S.-centric nature of the dataset, which may limit generalizability to other cultural contexts.

References

[1] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[2] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20983–20992, 2022.

[3] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset. *arXiv preprint arXiv:2104.10133*, 2021.

[4] Jia Huang, Peng Jiang, Alvika Gautam, and Srikanth Saripalli. Gpt-4v takes the wheel: Promises and challenges for pedestrian behavior prediction. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 134–142, 2024.

[5] Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *arXiv preprint arXiv:2503.10621*, 2025.

[6] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10591–10599, 2019.

[7] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6748–6765, 2021.

[8] Farzeen Munir, Shoaib Azam, Tsvetomila Mihaylova, Ville Kyrki, and Tomasz Piotr Kucner. Pedestrian vision language model for intentions prediction. *IEEE Open Journal of Intelligent Transportation Systems*, 2025.

[9] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024.

[10] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024.

[11] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024.

[12] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22442–22452, 2025.

[13] Z Yang, X Jia, H Li, and J Yan. Llm4drive: A survey of large language models for autonomous driving. arxiv 2023. *arXiv preprint arXiv:2311.01043*.