OmniPredict: GPT-4o Enhanced Multi-modal Pedestrian Crossing Intention Prediction

Je-Seok Ham
Electronics and Telecommunications Research Institute (ETRI)
Korea Advanced Institute of Science and Technology (KAIST)
jsham@etri.re.kr

Jia Huang
Texas A&M University
jia.huang@tamu.edu

Peng Jiang
Texas A&M University
maskjp@tamu.edu

Jinyoung Moon
Electronics and Telecommunications Research Institute (ETRI)
jymoon@etri.re.kr

Yongjin Kwon
Electronics and Telecommunications Research Institute (ETRI)
scocso@etri.re.kr

Srikanth Saripalli
Texas A&M University
ssaripalli@tamu.edu

Changick Kim*
Korea Advanced Institute of Science and Technology (KAIST)
changick@kaist.ac.kr

Abstract

Pedestrian crossing intention prediction is a crucial component for ensuring safety and responsible navigation in urban autonomous driving systems. Traditional methods, which have relied on vision-based models, struggle to generalize to unseen driving scenarios due to their dependence on training data. Multimodal Large Language Models (MLLMs) offer a novel approach to these challenges through their advanced reasoning capabilities. In this paper, we introduce OmniPredict, the first study to evaluate GPT-4o(mni), a cutting-edge MLLM, for predicting pedestrian crossing intentions. Using the JAAD dataset, our model achieved 67% prediction accuracy in a zero-shot setting, outperforming the performance of existing state-of-the-art MLLM methods by 17.5% without the need for additional data or retraining. By integrating diverse contextual modalities and finely tuned prompts, our approach enhances prediction reliability and reduces uncertainty. This demonstrates that our method contributes to improving prediction performance, thereby advancing safer driving environments.

1 Introduction

As autonomous driving technology advances, extensive research[1, 2, 3, 4] has focused on ensuring pedestrian safety, as they are key participants in urban traffic. Accurate prediction of pedestrian crossing intentions is essential for avoiding accidents and maintaining reliable autonomous driving systems. However, traditional vision-based approaches depend on limited datasets and supervised learning. The mainstream pipeline extracts features from past image frames such as pedestrian trajectories[5], 2D keypoints[1, 6], vehicle speed[1, 3], and semantic maps[7]. Earlier models used Long Short-Term Memory (LSTM)[8] or Recurrent Neural Networks (RNNs)[9] for sequential inputs, while later models applied CNNs[10, 11, 12, 3] and Graph Convolutional Networks (GCNs)[13, 14]

*Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

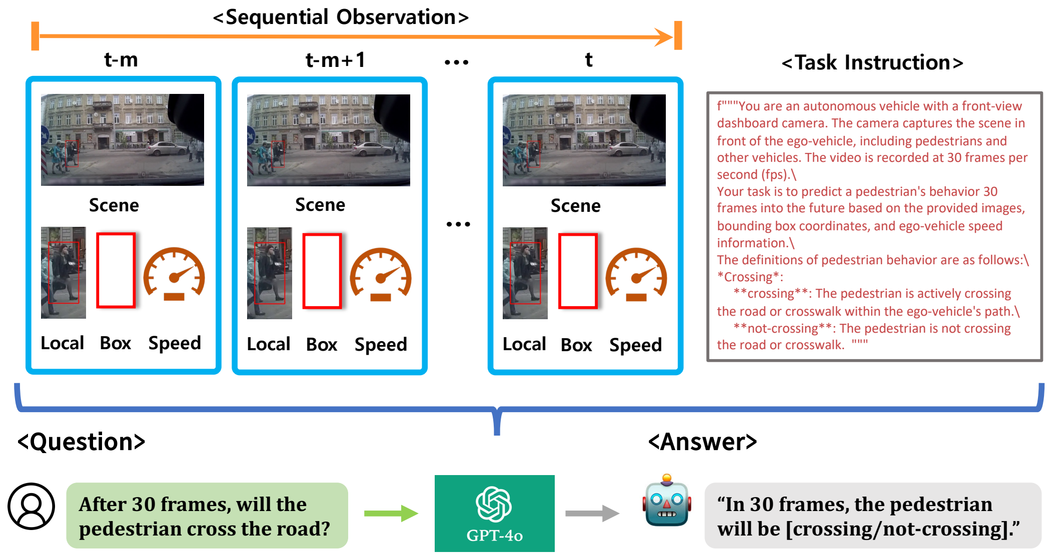


Figure 1: A diagram of the OmniPredict based on the GPT-4o model utilizing multiple modalities. The process involves providing multi-input features (Scene Context Image, Local Context Image, Bounding Box Coordinates, Ego-Vehicle Speed), task instructions, and questions to obtain GPT-4o's predictions on pedestrian crossing intention.

with attention mechanisms. More recently, transformer-based models[7, 1, 5] have been proposed to further enhance prediction accuracy.

Nevertheless, vision-based methods face challenges in detecting road environments or recognizing and predicting objects not included in the training data, leading to limitations in complex environments. The emergence of Multimodal Large Language Models (MLLMs) like GPT-4V(ision)[15], LLAMA[16], and LLAVA[17] offers promising solutions. These models have demonstrated strong zero-shot recognition abilities and efficiently handle complex multimodal tasks. In particular, MLLMs excel in interpreting visual data for decision-making in driving scenarios.

In this paper, we present **OmniPredict**, the first approach to use GPT-4o[18] (where "o" stands for "omni"), a cutting-edge MLLM, for pedestrian crossing intention prediction. Our method integrates meticulously tuned instruction prompts and diverse contextual modalities such as scene context images, local context images, bounding box coordinates, and ego-vehicle speeds for GPT-4o. We used the widely recognized Joint Attention in Autonomous Driving (JAAD)[19] datasets for evaluation. OmniPredict achieved a 67% accuracy in a zero-shot setting, representing a 17.5% improvement over the performance of GPT4V-PBP[4], the current state-of-the-art MLLM method. Our approach significantly improved performance without the need for additional training or data collection. We compared with domain-specific benchmark models to validate the effectiveness of our method in accurately predicting pedestrian behavior. Qualitative results also revealed that GPT-4o outperforms GPT-4V by providing a deeper understanding of the road environment and interactions between road users, leading to fewer prediction errors and enhancing safety in complex driving conditions.

In conclusion, OmniPredict represents a major step forward by moving beyond traditional vision-based models. Its multi-modal framework enables zero-shot predictions of pedestrian crossing intention in unfamiliar environments, enhancing both the safety and reliability of driving systems.

2 Methodology

GPT-4o represents a significant advancement over previous models like GPT-4V, enabling much more natural human-computer interaction. It excels in visual understanding and functions as a fully comprehensive multimodal model. In our experiments, we used the OpenAI Python API to perform tasks on the JAAD dataset through the GPT-4o API. Specifically, we focused on the JAAD_{beh} dataset, which includes detailed annotations of pedestrian behaviors.

Our tasks used sequences of past frames as inputs to predict pedestrian behavior at a specified future time. In contrast to GPT4V-PBP [4], our method integrates additional contextual data from previous frames, including local context images, bounding box coordinates, and ego-vehicle speed, as illustrated in Fig.1. By combining these elements, we input them into GPT-4o, allowing it to predict pedestrian behavior with richer contextual awareness. The initial prompt informs the model that it functions as an autonomous vehicle, with the visuals coming from a front-facing camera. We designed the model to process information from 16 past frames and make predictions 30 frames into the future, utilizing the same setup that other vision-based benchmark models have used for performance evaluation.

```
f"""You are an autonomous vehicle with a front-view dashboard camera. The camera captures the scene in front of the ego-vehicle, including pedestrians and other vehicles. The video is recorded at 30 frames per second (fps).\
Your task is to predict a pedestrian's behavior 30 frames into the future based on the provided images, bounding box coordinates, and ego-vehicle speed information."
```

Next, we defined pedestrian crossing behavior, as detailed in the following prompt. For crossing, the main criterion was whether the pedestrian's movement was directed towards the ego-vehicle, although it was also important to consider if they were crossing the road or crosswalk.

```
f"""The definitions of pedestrian behavior are as follows:\
*Crossing*:
  **crossing**: The pedestrian is actively crossing the road or crosswalk within the ego-vehicle's path.\
  **not-crossing**: The pedestrian is not crossing the road or crosswalk.\ """
```

We define the following four features to be used as input.

First, the ***Scene Context Image*** feature is defined as follows: $SC_i = \{sc_i^{t-15}, sc_i^{t-14}, sc_i^{t-13}, ..., sc_i^{t_0}\}$, where sc_i refers to the full image (1920×1080 pixels) capturing all agents in the road environment, such as pedestrians, crosswalks, and vehicles. The pedestrian bounding box is marked with a red box, and GPT4V-PBP [4] uses only this full image for predictions.

The ***Local Context Image*** feature is defined as: $LC_i = \{lc_i^{t-15}, lc_i^{t-14}, lc_i^{t-13}, ..., lc_i^{t_0}\}$, where lc_i is cropped to an area 1.5 times the size of the pedestrian's bounding box. This cropped image is then resized to 224×224 pixels to maintain size uniformity across all pedestrians. Similar to the scene context image, the pedestrian's bounding box is highlighted with a red box. GPT-4o utilizes this red box to focus on the pedestrian's movement and body direction when making its predictions.

The ***Bounding Box Coordinates*** feature is defined as: $B_i = \{b_i^{t-15}, b_i^{t-14}, b_i^{t-13}, ..., b_i^{t_0}\}$, where $b_i = [x_{tl}, y_{tl}, x_{br}, y_{br}] \in \mathbb{R}^4$ represents a 2D bounding box, defined by the coordinates of the top-left (x_{tl}, y_{tl}) and bottom-right (x_{br}, y_{br}) corners of each pedestrian.

The ***Ego-Vehicle Speed*** feature is defined as: $ES_i = \{es_i^{t-15}, es_i^{t-14}, es_i^{t-13}, ..., es_i^{t_0}\}$, where $es_i \in \{moving\ slow, decelerating, stopped, accelerating, moving\ fast\}$ categorizes the ego-vehicle speed into five levels for each frame. GPT-4o uses this to infer pedestrian behavior, such as the likelihood of crossing based on the speed of ego-vehicle.

Based on these multi-input features, GPT-4o provides the following prediction results for pedestrian behavior 30 frames ahead: "In 30 frames, the pedestrian will be [crossing/not-crossing]."

3 Experimental Results

3.1 Dataset

Joint Attention in Autonomous Driving (JAAD) dataset: The JAAD dataset was collected using front-facing cameras mounted on vehicles, capturing from various locations. Video clips were divided into 188 for training, 32 for validation, and 126 for testing. The JAADall includes all pedestrians, regardless of whether they are involved in crossing. In contrast, the JAADbeh focuses on pedestrians who are either crossing or showing an intention to cross, featuring labeled behavioral information.

3.2 Quantitative Results

In Table 1, the performance of traditional domain-specific models and GPT-based approaches is evaluated across five key metrics. The models range from MultiRNN [20] (2018) to the latest OmniPredict (2024), all tasked with predicting pedestrian crossing intent 30 frames in advance. The

Table 1: Performance comparison with state-of-the-art methods summarized using Accuracy (ACC), Area Under the Curve (AUC), F1 Score (F1), Precision (P), and Recall (R). The best results are indicated in bold, while the second-best results are underlined.

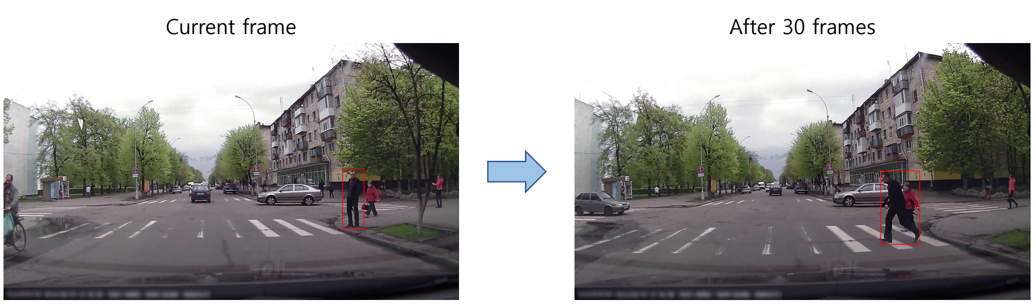
Models	Year	Model Variants	Use Frames	JAAD-beh				
				ACC↑	AUC↑	F1 ↑	P ↑	R ↑
MultiRNN [20]	2018	GRU	16	0.61	0.50	<u>0.74</u>	0.64	<u>0.86</u>
SFRNN [9]	2020	GRU	16	0.51	0.45	0.63	0.61	0.64
SingleRNN [21]	2020	GRU	16	0.58	0.54	0.67	0.67	0.68
PCPA [22]	2021	RNN+Attention	16	0.58	0.50	0.71	\	\
IntFormer [23]	2021	Transformer	16	0.59	0.54	0.69	\	\
ST CrossingPose [24]	2022	Graph CNN	16	0.63	0.56	<u>0.74</u>	0.66	0.83
FFSTP [25]	2022	GRU+Attention	16	0.62	0.54	<u>0.74</u>	0.65	0.85
PIT-Block(a) [1]	2022	Transformer	16	**0.70**	**0.65**	**0.81**	0.71	**0.93**
GPT4V-PBP [4]	2023	MLLM	10	0.57	<u>0.61</u>	0.65	**0.82**	0.54
GPT4V-PBP Skip [4]	2023	MLLM	10	0.55	0.59	0.64	<u>0.81</u>	0.53
OmniPredict	2024	MLLM	16	<u>0.67</u>	**0.65**	0.65	0.66	0.65

table compares models based on their year of release and the number of preceding frames they use. While the PIT-Block(a) [1] model leads in Accuracy of 0.70, AUC of 0.65, F1 of 0.81, and Recall of 0.93, this can be attributed to its inclusion of pedestrian 2D keypoints, a feature not used in our approach. Even without additional training, GPT-based MLLM models, including OmniPredict, show competitive results. Notably, OmniPredict achieves the highest AUC of 0.65 and ranks second in Accuracy of 0.67. It significantly outperforms GPT4V-PBP in both Accuracy of 0.67 compared to 0.57 and a recall of 0.65 compared to 0.54, indicating stronger performance in identifying relevant instances. OmniPredict offers a balanced trade-off between Precision and Recall, making it a more robust model for general classification purposes.

3.3 Qualitative Results

To understand the reasons behind GPT-4o's enhanced performance compared to GPT-4V using the same input features, we conducted a qualitative analysis. In this evaluation, we observed that GPT-4o better interprets complex traffic scenarios and pedestrian interactions, demonstrating a higher ability to generate accurate and reliable predictions. In the following example, we compare the conversations of GPT-4V and GPT-4o regarding a particular traffic situation. The key phrases are emphasized in blue text, while the comparison of performance between GPT-4V and GPT-4o is highlighted in red. In the "Pedestrian Crossing Intention Prediction" scenario, when asked whether a pedestrian standing near the curb is likely to cross the road, GPT-4o delivers a more definitive response, predicting that the pedestrian will cross. In contrast, GPT-4V provides a less conclusive answer, considering both possibilities: that the pedestrian might cross or remain in place.

Pedestrian Crossing Intention Prediction



Do you think the man in the red box will cross the crosswalk in 30 frames, or will he not?

GPT-4V: From the image, it appears that the pedestrian in the red box is looking towards the ego-vehicle, which suggests that he might be assessing the traffic before crossing. His body posture and direction of gaze indicate that he might be intending to cross the road. However, it's also possible that he is simply waiting at the corner or deciding his next move based on traffic conditions.

GPT-4o: It appears that the pedestrian in the red box is trying to cross the road that the ego-vehicle is on. The pedestrian is positioned at the edge of the crosswalk, facing the road, and seems to be looking in the direction of oncoming traffic, which typically indicates an intention to cross.

4 Conclusion

This paper introduced OmniPredict, employing GPT-4o with multi-contextual inputs to improve pedestrian crossing intention prediction in urban driving. By integrating multiple modalities and finely tuned instruction text prompts, we increased pedestrian crossing intent prediction accuracy on the JAAD dataset by 17.5%. Our zero-shot approach eliminates the need for further training, making it efficient and sustainable. We compared its performance against traditional domain-specific models. Qualitative comparisons showed that GPT-4o demonstrates superior adaptability to unfamiliar environments compared to GPT-4V. Our results demonstrate the potential of zero-shot predictions to improve road safety and become integral to future traffic management systems.

Acknowledgement

This work was supported by IITP grant funded by the Korea government(MSIT) (No.2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

References

[1] Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[2] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *International Conference on Computer Vision (ICCV)*, 2019.

[3] Je-Seok Ham, Dae Hoe Kim, NamKyo Jung, and Jinyoung 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*, pages 3666–3675, 2023.

[4] Jia Huang, Peng Jiang, Alvika Gautam, and Srikanth Saripalli. Gpt-4v takes the wheel: Evaluating promise and challenges for pedestrian behavior prediction. *arXiv preprint arXiv:2311.14786*, 2023.

[5] Jia Huang, Alvika Gautam, and Srikanth Saripalli. Learning Pedestrian Actions to Ensure Safe Autonomous Driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8, 2023.

[6] Shile Zhang, Mohamed Abdel-Aty, Yina Wu, and Ou Zheng. Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):2331–2339, 2021.

[7] Amir Rasouli and Iuliia Kotseruba. Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9844–9851. IEEE, 2023.

[8] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks. *IEEE Transactions on Intelligent Vehicles*, 3(4):414–424, 2018.

[9] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*, 2020.

[10] Haziq Razali, Taylor Mordan, and Alexandre Alahi. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation Research Part C: Emerging Technologies*, 130:103259, 09 2021.

[11] Ankur Singh and Upendra Suddamalla. Multi-input fusion for practical pedestrian intention prediction. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2304–2311, 2021.

[12] Je-Seok Ham, Kangmin Bae, and Jinyoung Moon. Mcip: Multi-stream network for pedestrian crossing intention prediction. In *European Conference on Computer Vision*, pages 663–679. Springer, 2022.

[13] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[14] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21050–21061, 2022.

[15] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[18] Hello gpt-4o, 2024.

[19] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017.

[20] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4194–4202, 2018.

[21] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693. IEEE, 2020.

[22] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1258–1268, 2021.

[23] Javier Lorenzo, Ignacio Parra, and MA Sotelo. Intformer: Predicting pedestrian intention with the aid of the transformer architecture. *arXiv preprint arXiv:2105.08647*, 2021.

[24] Xingchen Zhang, Panagiotis Angeloudis, and Yiannis Demiris. St crossingpose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):20773–20782, 2022.

[25] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith A Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 7(2):221–230, 2022.