Commun Transp Res

Research Article

VLMPed-CoT: A large vision-language model with a chain-of-thought mechanism for pedestrian crossing intention prediction

Yancheng Ling[1], Zhenlin Qin[1], Leizhen Wang[2], Zhendong Liu[3], Yang Liu[4], Zhenliang Ma[1,5,✉]

Cite this article: Ling Y C, Qin Z L, Wang L Z, et al. *Commun Transp Res* 2026, **6**(1): 9640009. https://doi.org/10.26599/COMMTR.2026.9640009

ABSTRACT: Pedestrian crossing intention prediction is crucial for autonomous driving. While existing models have achieved high accuracy, their generalization and robustness remain limited, hindering their performance in real-world scenarios. To overcome these limitations, we introduce the LVLMPed-CoT, a large vision language model (LVLM) that incorporates a chain-of-thought (CoT) mechanism to enhance pedestrian crossing intention prediction. It takes multimodal data as input and employs data distillation along with a two stage fine-tuning strategy to elicit the implicit CoT capability of a lightweight vision-language model for enhanced perception, reasoning, and prediction. The unified LVLMPed-CoT is trained on a joint open-source dataset (JAAD and PIE) and achieves superior or comparable performance to state-of-the-art models on both large-scale public datasets. The ablation study validates the contribution of the CoT prompt design and the two-stage fine-tuning strategy to the model's performance. Further analysis investigates the impact of input data sequence length and image quality on both accuracy and inference time, as well as the interpretability of the enhanced CoT reasoning ability achieved through fine-tuning.

KEYWORDS: pedestrian crossing intention prediction; large vision language model (LVLM); chain of thought (CoT); data distillation; two-stage fine-tuning strategy

1 Introduction

Pedestrians are important participants in traffic and are also the most vulnerable to injury on the road (Yang et al., 2026). With the advancement of autonomous technology, pedestrian safety is receiving increasing attention (Ling et al., 2023a; Liu et al., 2023; Sharma et al., 2025a; Soleimani and Saria, 2025; Xu and Zheng, 2024; Zhou et al., 2024). Predicting pedestrian crossing intentions plays a crucial role by enabling the early identification of crossing behavior (Bai et al., 2025; Sakib et al., 2024; Sharma et al., 2025b; Zhou et al., 2025). This allows vehicles to respond automatically and avoid potential conflicts with pedestrians, thereby reducing traffic accidents and enhancing overall pedestrian safety. However, pedestrian crossing intention prediction is challenging, as it requires comprehensive considered spatiotemporal information from both pedestrian movement and the surrounding traffic environment, as well as effective modeling of their interactions to enhance predictive performance (Gao et al., 2025).

Early methods used recurrent relational networks (RNNs) to extract pedestrian and environmental features, with attention for multimodal fusion (Gesnouin et al., 2021; Kotseruba et al., 2021; Yang et al., 2022). However, their performance is limited due to RNNs' weak ability to model long spatiotemporal sequences. In comparison, spatiotemporal graph convolutional network (ST-GCN) and transformer-based models demonstrate significantly stronger capabilities in extracting spatiotemporal information from multimodal data and have achieved state-of-the-art performance (Cadena et al., 2022; Chen et al., 2025b; Ling et al., 2024b; Yang et al., 2024; Zhou et al., 2023). However, traditional deep learning-based models still lack overall stability. For example, state-of-the-art models may achieve high accuracy but suffer from low recall performance (e.g., $JAAD_{all}$ in Table 1) (Chen et al., 2025b; Ling et al., 2024b; Ling and Ma, 2024a; Yang et al., 2024), indicating an overreliance on training data and limited generalization and robustness. Additionally, no unified model has been proposed that performs consistently well across different datasets for the same task. Moreover, these methods operate as black boxes, making it difficult to explain or trace issues when they arise for real-world deployment. This lack of transparency hinders trust, safety, and the ability to effectively optimize or adjust the model (Nazari et al., 2025). These limitations hinder the real-world deployment and application of such models (Chen et al., 2025a).

The emergence of large language models (LLMs), such as ChatGPT (Brown et al., 2020) and Gemini (Team et al., 2023), offers promising solutions. These models have demonstrated exceptional capabilities in downstream tasks and have proven

[1] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm 11428, Sweden. [2] Department of Data Science and Artificial Intelligence, Monash University, Clayton 3800, Australia. [3] Department of Engineering Mechanics, KTH Royal Institute of Technology, Stockholm 11428, Sweden. [4] State key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [5] Digital Futures, KTH Royal Institute of Technology, Stockholm 10044, Sweden.

✉ Corresponding author. E-mail: zhema@kth.se

Received: October 15, 2025; Revised: December 13, 2025; Accepted: January 4, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

effective in handling complex multimodal scenarios, enabled by pretraining on large-scale datasets. Recent studies have explored the use of multimodal LLMs, such as ChatGPT-4o, for pedestrian crossing intention prediction by carefully designing prompts in zero-shot scenarios (Ham et al., 2026; Huang et al., 2024). However, their performance still falls short compared to state-of-the-art deep learning methods. While these models have strong general capabilities, challenges remain in pedestrian crossing intention prediction, particularly in perception, reasoning, and decision-making. For example, the model should focus on pedestrian movement over clothing color for feature perception and learn decision rules such as higher vehicle speeds reducing crossing intention. Moreover, current research relies on large-scale language models ($\geq$ 100B) that are time- and resource-intensive, limiting their practicality for real-world deployment.

To address these issues, we propose a lightweight, open-source large vision-language model (LVLM) with a chain-of-thought (CoT) (Wei et al., 2022) reasoning mechanism for pedestrian crossing intention prediction (LVLMPed-CoT). As shown in Fig. 1, we use a single scene image to capture the traffic environment, in contrast to previous studies that rely on scene image sequences, thereby reducing both resource consumption and inference time. To extract pedestrian-specific information, we use a sequence of cropped pedestrian images to capture pose features and a sequence of bounding boxes to represent motion and spatial position. Additionally, ego-vehicle velocity is incorporated to characterize vehicle dynamics, such as speed and motion behavior (Ling et al., 2024b). For model perception and reasoning, we propose a CoT prompting strategy to stimulate the LVLM's ability to perform autonomous, stage-by-stage reasoning for pedestrian crossing intention prediction. To reduce resource consumption and inference time, we adopt a lightweight LVLM (e.g., Qwen2.5-VL-3B). To ensure strong reasoning capability, we leverage a large-scale model (e.g., Gemini-2.5) as a teacher to distill CoT reasoning data and introduce a two-stage fine-tuning framework to enhance both explicit CoT reasoning (Yao et al., 2025) and implicit inference abilities (Deng et al., 2023, 2024) of the lightweight LVLM. Furthermore, unlike previous studies that trained separate models on different datasets (e.g., the pedestrian intention estimation (PIE) dataset and the joint attention in autonomous driving (JAAD) dataset), we propose a unified model trained on the joint dataset and evaluated across multiple benchmarks to improve and validate its generalization performance. The main contributions of the study are as follows:

1) We introduce LVLMPed-CoT, a lightweight LVLM-based model that incorporates a CoT reasoning mechanism to improve pedestrian crossing intention prediction.

2) We propose a CoT prompting approach to stimulate the model's automatic reasoning capabilities, along with data distillation and a two-stage fine-tuning strategy to enhance both explicit CoT reasoning and implicit inference abilities of the lightweight LVLM.

3) We validate the model performance on two large-scale public datasets, JAAD and PIE. The unified model achieves comparable or superior results across both datasets compared to state-of-the-art approaches.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 presents the proposed methodology. Section 4 evaluates the model performance, conducts ablation studies, and provides further analysis. The final Section 5 concludes the study and outlines directions for future research.

2 Related work

This review focuses on traditional deep learning-based models and MLLM-based models for pedestrian crossing intention prediction, as well as on both explicit and implicit CoT reasoning approaches.

2.1 Pedestrian crossing intention prediction based on traditional deep learning models

Early models predict pedestrian crossing intention from a single image (Rasouli et al., 2017), typically using VGG16 (Simonyan and Zisserman, 2014) or ResNet50 (He et al., 2016) on the last frame to decide whether the pedestrian will cross in the next 1–2 s, but the lack of temporal information limits their ability to model motion. To capture sequence dynamics, later works employed 3-dimensional convolutional neural networks (3D CNNs) on image sequences (Carreira and Zisserman, 2017; Tran et al., 2016; Zhang et al., 2024), achieving better performance by modeling

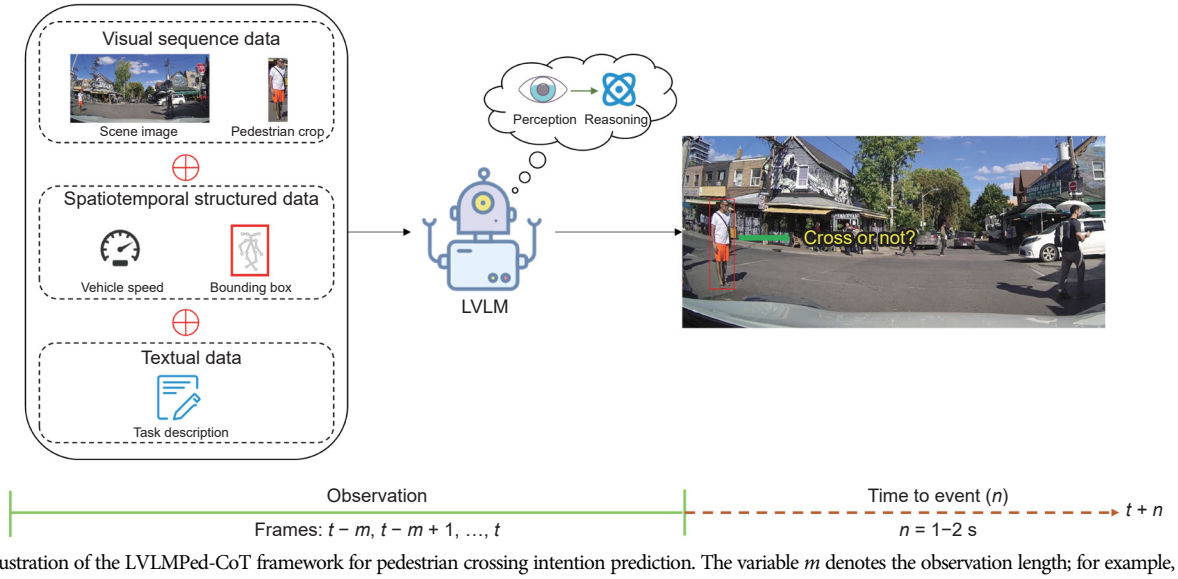


Fig. 1 Illustration of the LVLMPed-CoT framework for pedestrian crossing intention prediction. The variable m denotes the observation length; for example, m = 16 corresponds to an observation window of 16 frames. Within this window, the scene image is first processed by an object detector to obtain the cropped pedestrian images and their bounding boxes, and the vehicle speed is recorded at the same time step; all modalities at each time step are extracted from the same frame (from $t - m + 1$ to t), ensuring temporal synchronization across the multimodal inputs.

richer spatiotemporal dependencies yet still relying on a single modality.

Subsequently, RNN-based models such as gated recurrent unit (GRU) and long short-term memory (LSTM) network fuses multiple modalities (e.g., scene images, pedestrian crops, vehicle speed, and bounding boxes) for pedestrian crossing prediction (Kotseruba et al., 2020, 2021; Rasouli et al., 2020; Yang et al., 2022). For instance, Kotseruba et al. (2021) used an LSTM/GRU backbone to model skeletons, bounding boxes, and speed sequences combined with CNN features from local context images via attention, while Yang et al. (2022) further introduced global context image sequences to capture traffic scene information. However, the limited capacity of RNNs to model long-term dependencies still constrains the overall effectiveness of these methods.

Compared with RNN-based models, spatiotemporal attention GCNs (STA-GCN) and transformer-based models offer stronger spatiotemporal modeling by using attention rather than gating mechanisms to capture temporal dependencies, while GCNs are particularly effective for non-Euclidean structures such as skeletons. Zhou et al. (2023) integrated pose keypoints, image crops, vehicle speed, bounding boxes, and global scene images and used a transformer-based architecture to model cross-modal dependencies for prediction. Yang et al. (2024) and Chen et al. (2025b) adopt GCN-based architectures to extract features from skeletons, vehicle speed, and bounding boxes, achieving state-of-the-art accuracy in pedestrian intention prediction. However, their performance remains unstable: For example, although the accuracy on $JAAD_{all}$ exceeds 89%, the recall is still below 70%, indicating overly conservative predictions and limited generalization in real-world scenarios.

2.2 Pedestrian crossing intention prediction based on MLLMs

With the development of LLMs, an increasing number of studies have focused on using MLLMs to predict pedestrian crossing intentions (Ham et al., 2026; Huang et al., 2024; Munir et al., 2025). Huang et al. (2024) used GPT-4 V as the reasoning model and input scene images with pedestrians marked by red bounding boxes to predict pedestrian crossing intentions. Ham et al. (2026) employed GPT-4o as the reasoning model, using local context, vehicle speed, bounding boxes, and scene images as inputs. They design prompts to stimulate the reasoning capabilities of LLMs and achieve improved performance. However, without the guidance of domain knowledge, their performance still falls behind that of deep learning-based models. Moreover, these approaches rely on large-scale LLMs as inference models, which demand substantial computational resources and incur high inference latency, thereby limiting their practical deployment and real-world applicability.

2.3 Explicit and implicit CoT for reasoning task

CoT prompting, which guides LLMs to reason step by step, is effective and has been widely used in tasks such as mathematics and code generation (Wei et al., 2022). However, it works best in large-scale LLMs and degrades in smaller models, while large models are resource intensive and difficult to deploy. To enhance CoT reasoning in smaller models for specific tasks, many works distill CoT data from large LLMs and then fine-tune smaller models on these data (Fu et al., 2023; Ho et al., 2022; Shridhar et al., 2023; Thawakar et al., 2025; Xu et al., 2024). For example, Fu et al. (2023) used code-davinci-002 to generate CoT solutions and applied instruction tuning to train Flan-T5, yielding substantial performance gains. Similarly, Ho et al. (2022) elicit CoT data from GPT-3 (175B) via zero-shot prompting and fine-tune smaller models such as GPT-2 and T5, showing that these fine-tuned models not only outperform prompt-based baselines but can even surpass the teacher model in many cases.

Although CoT can improve performance, explicit CoT reasoning introduces extra computation and latency, making it less suitable for latency-sensitive tasks. Recent works therefore explore implicit reasoning to enhance model capability (Deng et al., 2023, 2024; Yu, 2024). Unlike explicit reasoning, which outputs step-by-step rationales, implicit reasoning internalizes the reasoning process, improving inference speed. Deng et al. (2024) gradually removed intermediate CoT steps and fine-tuned the model to learn implicit CoT, achieving improved reasoning ability with only a slight performance gap to explicit CoT. In related work, Deng et al. (2023) propose an emulator that simulates intermediate CoT states and trains a student model to answer directly from these implicit representations. Yu (2024) further shows that models trained with Deng et al. (2023) can perform intermediate reasoning and exhibit CoT-like behavior on target tasks, leading to performance improvement.

3 Methodology

3.1 Problem definition

Following the formulation in Kotseruba et al. (2021), we define pedestrian crossing intention prediction as a binary classification problem. Specifically, the goal is to predict whether a pedestrian i will intend to cross $(I \in \{0,1\})$ at a future time $t+n$, where $n \in \{30,\ 60\}$ corresponds to approximately 1–2 s ahead. To perform the prediction, we leverage the current scene image along with the historical sequences over consecutive frames, including pedestrian image crop sequence $P_i^t = \{p_i^{t-m+1}, p_i^{t-m+2}, \ldots, p_i^t\}$, bounding box sequence $B_i^t = \{b_i^{t-m+1}, b_i^{t-m+2}, \ldots, b_i^t\}$, and vehicle speed sequence $V_i^t = \{v_i^{t-m+1}, v_i^{t-m+2}, \ldots, v_i^t\}$.

Pedestrian crossing intention prediction in the wild remains a challenging task. To enhance the reasoning capability of LVLMs for this problem, we propose PedLVLM-CoT, a structured framework based on CoT prompting. As illustrated in Fig. 2, PedLVLM-CoT consists of four main stages: prompt design, data distillation, supervised fine-tuning, and joint training.

1) Prompt design. It uses multimodal data (scene images, pedestrian crops, vehicle speed, and bounding boxes) to construct a structured CoT prompt, which serves as input to an LVLM.

2) Data distillation. It uses the structured prompt to activate teacher LVLM (Gemini-2.5), enabling the generation of CoT-style reasoning data for pedestrian intention prediction.

3) Supervised fine-tuning. It employs a two-stage supervised fine-tuning strategy to fine-tune a smaller LVLM (Qwen2.5VL-3B[1]) for pedestrian intention prediction. In the first stage, the model is guided to perform explicit reasoning based on the CoT data; in the second stage, it learns to make final predictions in an implicit manner.

4) Joint training. It uses the joint dataset (JAAD and PIE) to fine-tune the lightweight LVLM for pedestrian crossing intention.

3.2 Prompt design

LLM-based models possess strong reasoning abilities due to self-supervised learning on massive corpora. However, directly using

[1] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct

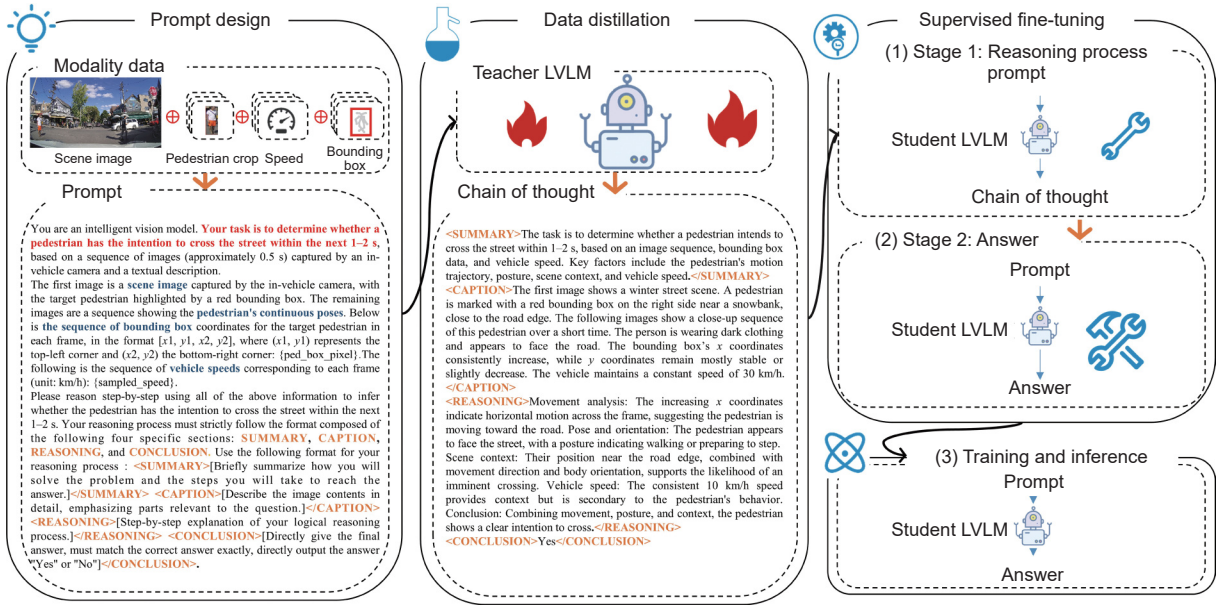


Fig. 2 Framework of PedLVLM-CoT for pedestrian crossing intention prediction consists of four main stages. First, it leverages multimodal inputs, including the scene image, pedestrian image crop sequence, vehicle speed sequence, and bounding box sequence, to construct a structured CoT prompt, which serves as input to an LVLM. The scene image and pedestrian crop sequence are provided to the LVLM as multi-image inputs, while the bounding box and vehicle speed sequences are encoded in the prompt as textual tokens. Second, the CoT prompt is used to activate Gemini-2.5, which generates CoT-style reasoning data tailored for the intention prediction task. Third, a two-stage supervised fine-tuning strategy is employed to train a smaller LVLM (e.g., Qwen2.5VL-3B) using the generated CoT data. Finally, the fine-tuned LVLM is deployed to perform pedestrian crossing intention prediction.

raw data as input and prompting the LLM with simple instructions to make decisions for downstream tasks is often ineffective. This is primarily because such approaches fail to effectively aggregate and organize the input multimodal information, thereby limiting the LLM's ability to fully engage its reasoning capabilities.

Inspired by Xu et al. (2024), we propose a structured CoT prompt to enhance the performance of LVLMs on pedestrian crossing intention prediction by encouraging step-by-step reasoning.

As shown in Fig. 3, the CoT prompt decomposes the answer generation process into four structured reasoning stages: planning, perception, reasoning, and decision.

Planning. It briefly explains the plan and steps involved in solving the pedestrian crossing intention prediction task, serving as a high-level summary of the question. This component outlines the primary aspects of the problem to be addressed and is enclosed within the special tag pair <Summary> and </Summary>.

Perception. It provides a concise overview of the visual elements and the temporal and spatial features of sequential data, including scene images, pedestrian image crop sequences, bounding box sequences, and vehicle speed sequences, all of which are relevant to the pedestrian crossing intention prediction task. It is used to perceive task-related features from multisource data and is enclosed within the special tag pair <Caption> and </Caption>.

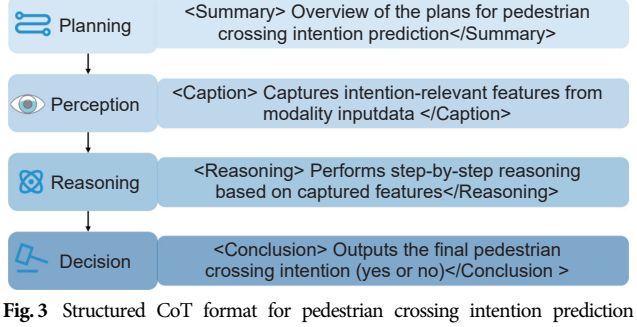


Fig. 3 Structured CoT format for pedestrian crossing intention prediction consists of four stages: planning, perception, reasoning, and decision-making.

Reasoning. It provides a step-by-step structured reasoning process based on the modality features. This component is designed to guide the LVLM in performing logical reasoning and is enclosed within the special tag pair <Reasoning> and </Reasoning>.

Decision. It provides the final prediction (Yes or No) based on the preceding reasoning and is enclosed within the special tag pair <Conclusion> and </Conclusion>.

As discussed above, we use the scene image, pedestrian image crop sequence, bounding box sequence, and vehicle speed sequence as inputs. Following the format of "task instruction, input description, and CoT reasoning with output", we construct the final prompt for the LVLM to perform pedestrian crossing intention prediction, as illustrated in Fig. 2a.

Compared to simple instruction prompts, the proposed CoT prompt enables the LVLM to reason more like a human, improving performance by guiding it through step-by-step reasoning: making a plan, observing the current situation and pedestrian behavior, reasoning based on these observations, and drawing a final conclusion.

3.3 Data distillation

CoT prompting can elicit language models to solve complex reasoning tasks step by step. However, prompt-based CoT methods typically rely on very large models (≥ 100 B), which are prohibitively expensive to deploy at scale. Moreover, manually annotating reasoning data is both time-consuming and resource-intensive. Inspired by previous work (Shridhar et al., 2023; Thawakar et al., 2025), we adopt the zero-shot-CoT prompting method on teacher models to automatically generate reasoning samples.

First, we use the teacher LVLM (e.g., Gemini-2.5) as the teacher model to generate CoT reasoning steps along with the result r.

Each raw sample s consists of multimodal input data X and its corresponding answer a. Following the method described in Section 3.2, we construct a prompt q to guide Gemini-2.5 in generating a reasoning explanation and final decision r in the format: "< Summary>...</Summary>, <Caption>...</Caption>, <Reasoning>...</Reasoning>, <Conclusion> (Yes or No) </Conclusion>".

Next, we filter the generated sample reasoning results r and reformat them into prompt-completion pairs. For filtering, we perform both format validation and answer consistency checks. First, we use regular expressions to strictly match whether the teacher model's output conforms to the format: "< Summary>... </Summary>, <Caption>...</Caption>, <Reasoning>... </Reasoning>, <Conclusion> (Yes or No) </Conclusion>". Then, we compare the conclusion in r with the ground truth answer a. An example of the filtered results is shown in Fig. 2b. Note that this filtering step leads to the loss of some training samples. For the remaining samples, we construct prompt-completion pairs using the prompt q and the reasoning output r, forming pairs in the format (X, q, r).

3.4 Supervised fine-tuning

Although the CoT prompt guides the LVLM to reason step by step, the performance of small LVLMs on the

downstream task remains inconsistent. On the one hand, small LVLMs lack sufficient CoT reasoning capabilities, which limits their effectiveness in pedestrian crossing intention prediction. On the other hand, this task has specific reasoning logic and domain rules. For example, the model should prioritize cues such as the pedestrian's pose and movement over irrelevant features such as clothing color. Additionally, vehicle speed is a critical factor because higher speeds generally suggest that it is not a safe moment for the pedestrian to cross. To address these challenges, we propose a two-stage fine-tuning method to better adapt the LVLM to this task.

In the first stage, we use (X, q, r) to train the small LVLM (e.g., Qwen2.5-VL-3B), as illustrated in Fig. 4a. This stage aims to teach the model to decompose complex tasks into logical reasoning steps. The training objective is to maximize the likelihood of generating the reasoning output r, conditioned on the input (X, q). The objective function is defined as

$$\mathcal{L}_{\mathrm{SFT}_{\mathrm{stage1}}} = -\mathbb{E}_{(\mathrm{X},p,y)\sim D} \sum_{t=1}^{T} \log \pi_\theta \left(y_t | \mathrm{X}, p, y_{<t}\right) \tag{1}$$

where D denotes the training dataset, consisting of samples in the form of (X, q, r). π_θ represents the model's token distribution. y_t denotes the token at time step t in the target sequence, and $y_{<t}$ represents the sequence of all preceding tokens. The resulting model, denoted as π_{CoT}, serves as the initialization for the next stage, providing a robust foundation for reasoning ability. In stage2, we use (X, q, a) (Fig. 4b) to refine π_{CoT}. In this stage, it aims to teach the model to directly output the result through implicit reasoning (Deng et al., 2024). The training objective is to maximize the

likelihood of generating the final answer a, conditioned on the input (X, q). The objective function is defined as

$$\mathcal{L}_{\mathrm{SFT}_{\mathrm{stage2}}} = -\mathbb{E}_{(\mathrm{X},p,a)\sim D} \sum_{t=1}^{T} \log \pi_{\mathrm{CoT}} \left(y_t | \mathrm{X}, p, \mathrm{y}_{<t}\right) \tag{2}$$

where π_{CoT} represents the initial model from stage1.

3.5 Unified fine-tuning

In real-world scenarios, pedestrian crossing intention prediction involves a wide range of diverse and complex environments, which increases the difficulty of accurate modeling. The widely used public datasets, PIE and JAAD, capture different types of scenes and driving contexts. However, most existing approaches train and evaluate models separately on these datasets, limiting the model's generalization capability across domains.

To address this issue and improve real-world robustness, we perform joint fine-tuning on both the JAAD and PIE datasets. During the inference stage, the fine-tuned model is evaluated on multiple subsets, including $\mathrm{JAAD}_{\mathrm{all}}$, $\mathrm{JAAD}_{\mathrm{beh}}$, and PIE, to assess its cross-domain performance.

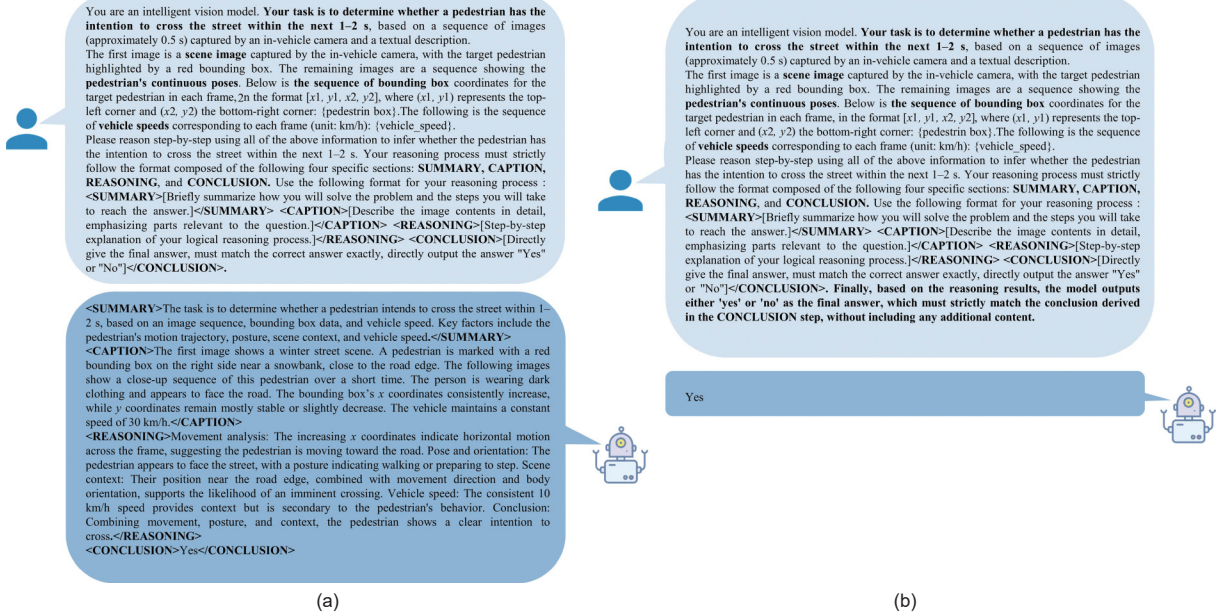


Fig. 4 Prompt-completion pairs used for fine-tuning are shown as follows: (a) represents the training data for stage1, which includes the prompt and the CoT reasoning process with answer; (b) represents the training data for stage2, which includes the prompt and the direct answer.

清华大学出版社 Tsinghua University Press | SciOpen

4 Experiments and evaluations

4.1 Dataset

4.1.1 JAAD

The JAAD dataset (Rasouli et al., 2017) is a dedicated resource for autonomous driving research, comprising 346 video clips, each ranging from 5 to 10 s in duration. The clips were collected in North America (60 clips) and Europe (286 clips) and mainly depict scenes in environments such as parking lots, garages, and city streets. The behavioral subset of JAAD (denoted as $JAAD_{beh}$) includes annotations for 495 crossing pedestrians and 191 pedestrians showing an intention to cross. The full dataset ($JAAD_{all}$) further incorporates 2100 additional pedestrians who are visible but located far from the road and exhibit no intention of crossing. For consistency and fair comparison, we adopt the same data split as proposed in (Kotseruba et al., 2021), using 177 clips for training, 29 for validation, and 117 for testing.

4.1.2 PIE

The PIE dataset (Rasouli et al., 2019) serves as another real-world benchmark for pedestrian crossing intention prediction. It was recorded in downtown Toronto, Canada, during the daytime under both sunny and overcast weather conditions. The scenes are primarily captured in cities streets. The dataset includes 512 crossing instances and 1322 noncrossing instances, encompassing all pedestrians near the road, regardless of whether they exhibit hesitation. To ensure fair comparison, we adopt the same data split protocol as suggested in (Kotseruba et al., 2021), using set01, set02, and set04 for training, set05 and set06 for validation, and set03 for testing.

4.2 Benchmark models and evaluation metrics

We evaluate our approach against several state-of-the-art methods on two widely used benchmark datasets: JAAD and PIE. The comparison covers CNN-based methods, including ATGC (Rasouli et al., 2017), C3D (Tran et al., 2016), I3D (Carreira and Zisserman, 2017), TwoStream (Simonyan and Zisserman, 2014), and Fussi-Net (Piccoli et al., 2020); RNN-based methods, including ConvLSTM (Shi et al., 2015), SingleRNN (Kotseruba et al., 2020), Stacked RNN (Yue-Hei Ng et al., 2015), MultiRNN (Bhattacharyya et al., 2018), SFRNN (Rasouli et al., 2020), PCPA (Kotseruba et al., 2021), Global PCPA (Yang et al., 2022), and TrouSPI-Net (Gesnouin et al., 2021); GCN-based methods, including pedestrian graph (Cadena et al., 2019), pedestrian graph+ (Cadena et al., 2022), ST CrossingPose (Zhang et al., 2022), STMA-GCN PedCross (Ling et al., 2023b), PedAST-GCN (Ling et al., 2024b), Faster-PCPN (Yang et al., 2024), and MB-STGCN (Chen et al., 2025b); and transformer-based methods, represented by PIT (Zhou et al., 2023); and MLLM-based methods, including GPT-4 V (Huang et al., 2024), OmniPredict (Ham et al., 2026), and LLaMAPed (Ham et al., 2025).

To ensure a fair and comprehensive comparison, we evaluate model performance using five metrics: accuracy, F1 score, precision, and recall, following the evaluation protocol proposed in (Kotseruba et al., 2021). The definitions of these metrics are as

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{6}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

4.3 Experimental setup

We utilize Gemini-2.5 as the teacher LVLM for data distillation and employ Qwen2.5-VL-3B as the backbone models for fine-tuning. For data preparation, since the vehicle speed in the JAAD dataset is estimated, we scale the values by a factor of 10 to align with the scale used in the PIE dataset. Pedestrian image crops are extracted from the scene images using the provided bounding boxes. We adopt a two-stage fine-tuning strategy. In stage1, Gemini-distilled CoT labels are filtered by consistency with the ground truth, resulting in 6200 training samples (3932 JAAD and 2268 PIE). In stage2, we further fine-tune the stage1 model on a larger nondistilled dataset with 13,383 samples (8612 JAAD and 4770 PIE). During joint training, samples from JAAD and PIE are simply merged into a single dataset, without dataset-specific loss weights or any sampling, so each instance contributes equally.

For model training, we employ the low-rank adaptation (LoRA) technique to fine-tune the Qwen2.5-VL model. The hyperparameters are experimented and set as follows: rank $r = 16$, $\alpha = 32$, dropout = 0.1, a learning rate of 1×10^{-4}, and epoch = 1, applied consistently across both stage1 and stage2. All experiments are conducted on a server equipped with 4 × NVIDIA A6000 GPUs.

4.4 Results

Table 1 shows the performance comparison on the JAAD and PIE datasets. The LVLMPed-CoT model achieves comparable or superior results compared to both traditional deep learning-based models and LVLM-based models. Specifically, the LVLMPed-CoT model achieves a significant improvement of 11% in accuracy on $JAAD_{beh}$ compared to state-of-the-art models while maintaining comparable performance on the $JAAD_{all}$ and PIE datasets. The improvement can be attributed to the use of multimodal data as input, the incorporation of CoT reasoning, the adoption of a two-stage fine-tuning strategy, and joint training for a unified model. Traditional deep learning-based models, particularly those using GCN and Transformer as backbone networks, achieve strong results due to their powerful spatiotemporal feature extraction capabilities (Chen et al., 2025b; Ling et al., 2024b; Yang et al., 2024; Zhou et al., 2023). However, although some models report higher accuracy, their performance is not robust. For instance, certain models achieve high accuracy at the cost of very low recall (Chen et al., 2025b; Yang et al., 2024), indicating overly conservative predictions. This lack of balance hinders their generalization and limits their applicability in real-world scenarios. LVLM-based models typically employ zero-shot prompting to engage general-purpose LLMs for pedestrian crossing intention prediction (Ham et al., 2026; Huang et al., 2024). These approaches use multimodal inputs and carefully crafted instructions to guide the LLMs. However, their performance is often limited due to a lack of alignment with task-specific semantics. They struggle to capture fine-grained features essential for accurate judgment, such as pedestrian posture, orientation, and motion cues, and fail to extract consistent decision-making "rules" from the captions alone. Moreover, without task-specific fine-tuning, these models

Table 1 Comparison of model performance on the PIE and JAAD datasets. $JAAD_{beh}$ is a subset of the JAAD dataset containing only pedestrian instances with behavioral annotations. In contrast, $JAAD_{all}$ includes all detected pedestrians, regardless of interaction. Acc denotes accuracy, F1 represents the F1 score, P is precision, and R is recall

Model name	Year	Model Variant	Use frame	Input data	PIE				$JAAD_{beh}$				$JAAD_{all}$			
					Acc	F1	P	R	Acc	F1	P	R	Acc	F1	P	R
ATGC	2017	VGG16	1	G, L, W	71	41	49	36	49	71	63	82	82	55	49	63
	2017	ResNet50	1	G, L, W	70	38	47	32	46	54	58	51	81	52	47	56
ConvLSTM	2015	VGG19 +LSTM	16	I	58	39	32	49	53	64	64	64	63	32	24	48
	2015	ResNet50 +LSTM	16	I	54	26	23	29	59	69	68	70	63	33	25	49
TwoStream	2014	VGG16	16	I, OF	64	32	33	31	56	66	66	66	60	43	29	83
SingleRNN	2020	GRU	16	I, B, S, Int	83	67	70	64	58	67	67	68	65	34	26	49
SingleRNN	2020	LSTM	16	I, B, S, Int	81	64	67	61	51	61	63	59	78	54	44	70
MultiRNN	2018	GRU	16	—	83	71	69	73	61	74	64	86	79	58	45	79
StakedRNN	2015	GRU	16	I, OF	82	67	67	68	60	66	73	61	79	58	46	79
HerarchicalRNN	2015	GRU	16	K	82	67	68	66	53	63	64	61	80	59	47	79
SFRNN	2020	GRU	16	I, G, K, B, S	82	69	67	70	51	63	61	64	84	65	54	84
C3D	2015	3DConv	16	I	77	52	63	44	61	75	63	91	84	65	57	75
I3D	2017	3DConv	16	I	80	62	67	58	62	73	68	79	81	63	66	61
	2017	Opticflow +3DConv	16	I, OF	81	72	60	90	62	75	65	88	84	63	55	73
FUSSI-Net	2020	DenseNet	16	B, K	—	—	—	—	59	69	66	73	60	40	27	73
PCPA	2021	3DConv	16	I, B, K, S	86	78	69	89	50	59	61	58	70	51	36	87
Global PCPA	2021	VGG+GRU	16	I, B, K, S, SGM	—	—	—	—	62	73	65	85	83	63	51	82
TrouSPI-Net	2021	GRU	16	B, K, S, ED	88	80	73	89	64	76	66	91	85	56	57	55
Pedestrian Graph	2019	GCN	16	K	76	48	62	39	62	70	71	68	80	55	46	68
Pedestrian Graph +	2022	Conv+GCN	32	I, K, S, SGM	89	81	83	79	70	76	77	75	86	65	58	75
ST CrossingPose	2022	ST-GCN	16	K	—	—	—	—	63	74	66	83	—	—	—	—
PIT	2023	Transformer	16	I, K, S, B, G	91	82	85	79	70	81	71	93	87	66	54	85
STMA-GCN PedCross	2023	STA-GCN	16	K	—	—	—	—	69	80	68	97	—	—	—	—
PedAST-GCN	2024	STA-GCN	16	K, S, B	91	83	88	79	69	79	68	93	89	68	67	69
Faster-PCPN	2024	GCN	16	K, S, B	94	89	89	88	—	—	—	—	89	65	73	58
MB-STGCN	2025	GCN	16	K, S, B	94	89	91	87	71	76	81	71	91	72	75	69
GPT4V	2023	MLLM	10	G	—	—	—	—	57	65	82	54	—	—	—	—
OmniPredict	2024	MLLM	16	G, I, B, S	—	—	—	—	67	65	66	65	—	—	—	—
LLaMAPed	2025	MLLM	16	G, I, B, S	—	—	—	—	58	59	67	52	—	—	—	—
LVLMPed-CoT	2025	MLLM	8 (skip 1 from 16)	G, I, B, S	90	82	83	81	82	84	90	80	90	74	69	80

Note: G represents the scene image; L signifies if the pedestrian is looking; W indicates if the pedestrian is walking; I is the pedestrian cropped image; OF refers to optical flow; B denotes the bounding box; S is the vehicle speed; Int indicates the pedestrian's intention to cross; K denotes pose keypoints; SGM refers to segmentation maps; and ED denotes relative distances between pairs of skeletal joints. MLLM represents the multimodal large language model.

are unaware of the expected label formats or reasoning patterns, leading to suboptimal and hard-to-evaluate predictions.

4.5 Ablation study

In this section, we will conduct the ablation study on the CoT prompt and the two-stage fine-tuning strategy.

4.5.1 Impact of the CoT prompt

To evaluate the impact of the CoT prompt, we test both the Qwen2.5-VL-3B and Gemini-2.5 models with and without the CoT prompt in a zero-shot scenario. The CoT prompt follows the template described in Section 3.2, while the template used without CoT is provided in the Appendix. As shown in Table 2, Qwen2.5-VL-3B without the CoT prompt achieves a high recall rate but low accuracy. The model tends to predict all samples as positive, indicating that it lacks the ability to perceive and understand fine-grained visual patterns and relies primarily on pretrained common sense to make direct judgments. As a result, the model makes decisions in a single step, which can easily lead to a simple common sense bias. In contrast, when using the CoT prompt, the performance of Qwen2.5-VL-3B improved significantly. Although the recall rate decreases, the model begins to reason step by step based on the prompts, allowing it to focus more on fine-grained visual information before making its final predictions. Compared with Qwen2.5-VL-3B, Gemini-2.5 demonstrates superior performance both with and without the CoT prompt. This can be attributed to Gemini-2.5's built-in CoT reasoning mechanism and its larger parameter size. Furthermore, Gemini-2.5 consistently achieves better results on all datasets when using the CoT prompt compared to when CoT is not used.

4.5.2 Impact of the two-stage fine-tuning strategy

To evaluate the impact of the two-stage fine-tuning strategy, we compare model performance across four settings: zero-shot, few-shot, one-stage fine-tuning, and two-stage fine-tuning. In the few-shot scenario, we randomly select one positive and one negative

Table 2 Ablation study of chain of thought

MLLMs	Method	PIE				$JAAD_{beh}$				$JAAD_{all}$			
		Acc	F1	P	R	Acc	F1	P	R	Acc	F1	P	R
Qwen2.5-VL-3B	Without CoT	28.41	43.9	28.16	99.53	62.57	76.98	62.57	100	17.69	29.82	17.52	100
	CoT	41.54	38.95	27.57	66.32	53.52	65.91	60.91	71.81	30.06	26.18	16.05	71.01
Gemini-2.5	Without CoT	42.98	35.81	26.21	56.52	66.56	76.43	68.36	86.66	43.46	34.65	21.71	85.73
	CoT	48.13	32.26	23.96	49.33	67.41	75.91	70.61	82.07	46.72	37.91	24.88	79.6

sample as demonstrations. For one-stage fine-tuning, we fine-tune the LLM directly on the answer. In the two-stage fine-tuning approach, we first fine-tune the model on the process and then further fine-tune it on the answer. As shown in Table 3, compared with the zero-shot setting, the few-shot approach achieves a notable improvement in performance. However, the overall results for both zero-shot and few-shot scenarios remain relatively limited. This can be attributed to the high variability of the few-shot examples, which may hinder the model's ability to generalize to real reasoning tasks. In contrast, one-stage fine-tuning results in significant performance improvements over both zero-shot and few-shot settings, as it effectively transforms general knowledge into specialized capabilities and systematically enhances the model's reasoning abilities. Compared with one-stage fine-tuning, two-stage fine-tuning further improves performance, especially on the PIE dataset, with an accuracy increase of approximately 1.8%. This improvement is mainly due to enhanced perception and reasoning abilities gained through the additional fine-tuning stage.

4.6 Model inference speed and accuracy

Inference speed plays an important role in pedestrian crossing intention prediction tasks. To investigate the impact of input data, specifically, image quality and sequence length (i.e., different sampling intervals from a 16-frame sequence), we conducted comparative experiments with different input configurations. As shown in Table 4, the scene image quality significantly affects both the accuracy and inference speed. Compared to using scene images with a resolution of 1920×1080 pixels, the performance decreases significantly when the resolution is reduced to 960×540 pixels. This is mainly because scene images are crucial for pedestrian intention prediction, as they contain many visual cues relevant to determining crossing intentions. For example, they provide information about the distance between pedestrians and the road, the orientation of pedestrians relative to the road, and the surrounding traffic conditions. As the image quality decreases, the model cannot clearly capture these details, especially when pedestrians are far from the vehicle. At the same time, higher image quality also demands significantly more computational resources and increases inference time. This is primarily because vision encoders (such as ViT) extract features by dividing images into patches with a fixed stride (e.g., 14) (Bai et al., 2025). Higher-resolution images produce more patches, thereby increasing the computational load and prolonging the inference time. However, increasing the resolution of the Cropped Pedestrian images results in a longer inference time, while the prediction performance remains unchanged. This is mainly because the pixel size of most pedestrians is already concentrated within 150×300.

In addition, sequence length, which is determined by applying different sampling intervals to a 16-frame sequence, plays an important role in model performance. When using a 4-frame sequence sampled with a stride of 4 (i.e., skip 3 from 16) as input, the model's performance drops significantly compared to using a 4-frame sequence sampled with a stride of 2 (i.e., skip 1 from 16), although the inference time is slightly reduced. The main reason is that using too few frames leads to the loss of key pedestrian moments, making it difficult for the model to effectively capture pedestrian motion information.

4.7 Interpretability analysis with and without fine-tuning

As discussed in Section 4.5.1, compared with the model without

Table 3 Ablation study of the two-stage fine-tuning strategy

Strategy	PIE				$JAAD_{beh}$				$JAAD_{all}$			
	Acc	F1	P	R	Acc	F1	P	R	Acc	F1	P	R
Zero-shot	41.54	38.95	27.57	66.32	53.52	65.91	60.91	71.81	30.06	26.18	16.05	71.01
Few-shot	46.35	39.08	28.73	61.08	60.56	70.93	65.79	76.94	44.57	23.45	15.46	48.51
One stage fine-tuning	88.18	77.84	82.41	73.74	82.19	85.14	89.05	81.56	90.34	74.81	68.78	81.99
Two stages fine-tuning	89.96	81.96	82.94	81.01	81.45	84.29	89.66	79.52	90.18	73.88	69.05	79.44

Note: All results were obtained using Qwen2.5-VL-3B.

Table 4 Evaluation of model inference speed and accuracy

Input	PIE				$JAAD_{beh}$				$JAAD_{all}$				Inference time (ms)
	Acc	F1	P	R	Acc	F1	P	R	Acc	F1	P	R	
1 Scene (960×540) + 8 Cropped pedestrians (150×300)	85.95	72.65	80.36	66.29	75.81	78.43	88.73	70.26	88.80	68.74	67.13	70.43	410
1 Scene (1920×1080) + 4 Cropped pedestrians (150×300)	84.59	68.32	81.07	59.03	77.25	80.62	86.32	75.62	88.01	68.83	63.10	75.70	890
1 Scene (1920×1080) + 8 Cropped pedestrians (150×300)	89.96	81.96	82.94	81.01	81.45	84.29	89.66	79.52	90.18	73.88	69.05	79.44	970
1 Scene (1920×1080) + 8 Cropped pedestrians (300×600)	89.12	80.15	82.40	78.03	82.46	85.44	88.89	82.24	89.32	72.92	65.49	82.24	1300

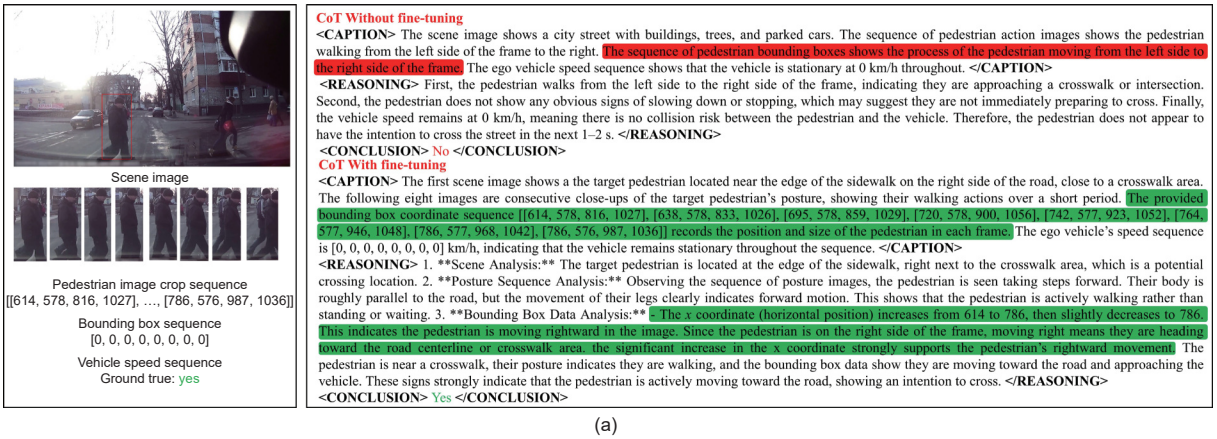
Note: All results were obtained using Qwen2.5-VL-3B. "1 Scene (960×540) + 8 Cropped pedestrians (150×300)" denotes using one scene image with a resolution of 960×540 pixels, together with a sequence of 8 cropped pedestrian images (each 150×300 pixels), sampled with a stride of 2 from an original sequence of 16 images, as input. All experiments were conducted on an NVIDIA A6000 GPU with 48 GB of memory.

fine-tuning (zero-shot), the two-stage fine-tuned model achieves substantial accuracy gains, improving accuracy by 48.42% on PIE, 27.93% on $JAAD_{beh}$, and 60.12% on $JAAD_{all}$, reaching 89.96%, 81.45%, and 90.18%, respectively. To further examine how fine-tuning affects the model's reasoning ability, we compare its outputs generated through the CoT reasoning process both before and after stage1 fine-tuning. Representative examples are shown in Fig. 5 (Due to space limitations, we retained only the caption, reasoning process, and conclusion stages of the CoT output). After CoT fine-tuning, the model demonstrates fine-grained information captioning, more rigorous logical reasoning, and stronger information integration and comprehensive decision-making abilities.

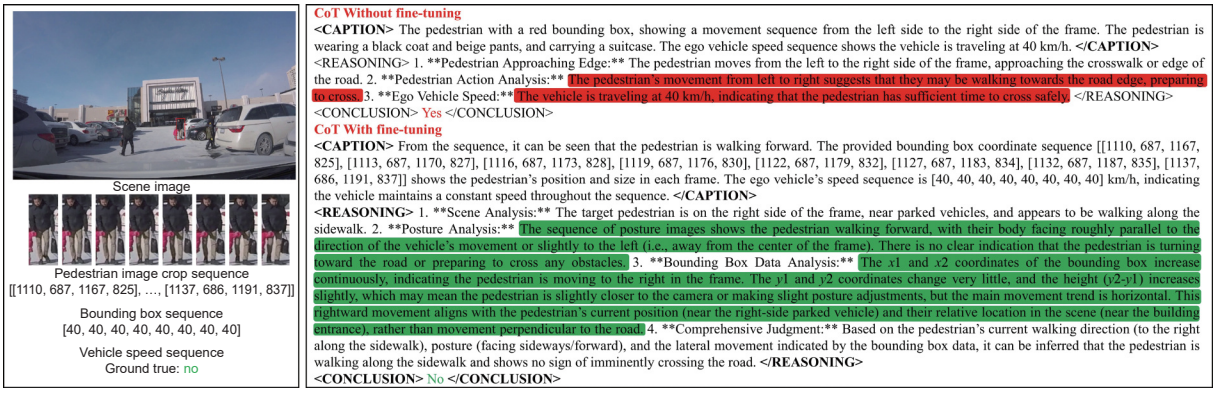
As shown in Fig. 5a, without fine-tuning, the LLM mainly focuses on superficial information, such as the approximate movement of pedestrians (highlighted in red). In contrast, with fine-tuning, the LLM is able to capture more detailed information in the caption stage and perform deeper analysis based on this information.

and ultimately make more informed decisions in the reasoning stage, as indicated by the green labeled text.

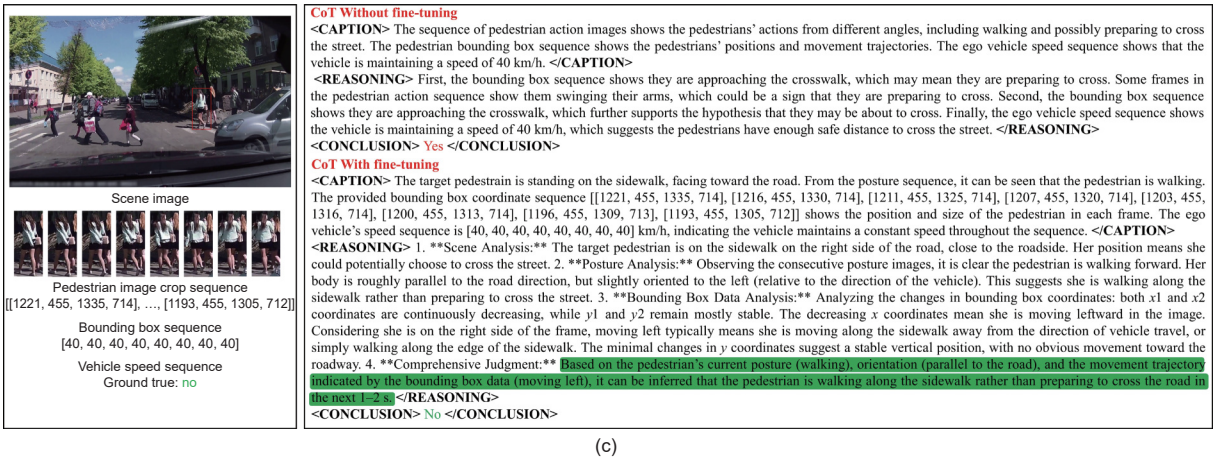
As shown in Fig. 5b, without fine-tuning, the LLM frequently exhibits reasoning biases. For example, it incorrectly states, "The pedestrian's movement from left to right suggests that they may be walking toward the road edge, preparing to cross." (highlighted in red), when in reality, the pedestrian is moving toward a parking area rather than the road, making it unlikely they intend to cross. This mistake can be attributed to the LLM's scene cognition bias. Additionally, the LLM claims, "The vehicle is traveling at 40 km/h, indicating that the pedestrian has sufficient time to cross safely."



(a)



(b)



(c)

Fig. 5 Comparison of the CoT reasoning process without fine-tuning and after stage1 fine-tuning.

清华大学出版社 Tsinghua University Press | SciOpen

However, 40 km/h is not a safe speed for pedestrians to cross, as it implies that vehicles in the vicinity may be moving relatively fast. This error reflects a failure in the LLM's common sense reasoning.

As shown in Fig. 5c, the fine-tuned LLM demonstrates stronger information integration and more comprehensive decision-making abilities. In real-world prediction scenarios, multiple sources of information may point to different outcomes. In this example, the pedestrian is walking along the roadside with their head slightly turned toward the road, which could suggest an intention to cross. However, the pedestrian's overall posture and movement indicate that they are mainly moving parallel to the flow of traffic, suggesting they are more likely to be walking along the road rather than intending to cross. After fine-tuning, the LLM is able to synthesize these conflicting cues and make an accurate prediction, whereas the non-fine-tuned LLM captures only partial information.

4.8 Analysis of failure cases

To further investigate the decision boundary of our model, we conduct an error analysis on failure cases. The main failure factors include occlusion by obstacles, remote distance, and contextually confusing scenarios. As shown in Figs. 6a and 6b, when pedestrians are heavily occluded or located far away from the ego vehicle, their pose and trajectory cues become difficult to perceive reliably, which leads to incorrect predictions. In addition, as illustrated in Figs. 6c and 6d, the model may also fail in highly confusing contexts. In Fig. 6c, the pedestrian appears near the center area in front of the vehicle, and the pedestrian's appearance blends into the surroundings due to clothing and background similarity, making it hard to distinguish the body posture; although the pedestrian's motion is parallel to the vehicle's driving direction, the model still produces an incorrect prediction. In Fig. 6d, at an intersection without a crosswalk, the vehicle is in the early stage of a right turn, and there is no salient interaction between the vehicle and the pedestrian's walking direction; meanwhile, the complex traffic flow further increases contextual ambiguity, which can mislead the model's intention inference and result in prediction failure.

(a) GT: cross, PR: no cross (b) GT: cross, PR: no cross

(c) GT: no cross, PR: cross (d) GT: cross, PR: no cross

Fig. 6 Typical examples of detection failures. "GT" is the ground truth classification value, and "PR" is the predicted value. The occlusion of obstacles, remote distance, and contextually confusing scenarios are the main reasons for failure detection.

5 Conclusions

Pedestrian crossing intention prediction is critical for intelligent vehicles. This study proposes a lightweight LVLM that incorporates a CoT mechanism to enhance reasoning. A lightweight LVLM (e.g., Qwen-VL 3B) serves as the core reasoning model, guided by CoT prompts to perform step-by-step inference. A larger teacher model (e.g., Gemini 2.5) is employed for data distillation, and a two-stage fine-tuning strategy is adopted to enhance the perception and reasoning capabilities of the lightweight LVLM for accurate pedestrian crossing intention prediction.

The model was trained on a joint dataset comprising JAAD and PIE and evaluated on three public datasets: $JAAD_{all}$, $JAAD_{beh}$, and PIE. The results demonstrate that the proposed LVLMPed-CoT achieves strong performance in terms of accuracy, generalization, and robustness for pedestrian crossing intention prediction. Specifically, LVLMPed-CoT achieved a significant accuracy improvement of approximately 11% (from 71% to 82%) on $JAAD_{beh}$ compared to state-of-the-art methods. The ablation analysis verifies the importance of the CoT prompt design and the two-stage fine-tuning strategy. Specifically, the CoT prompt guides the lightweight LVLM to reason step by step, while the two-stage fine-tuning strategy enhances its perception and reasoning capabilities for the target task. Most prediction failures arise under challenging visual conditions, including long-range observations, severe occlusions, and contextually confusing scenarios.

Further analysis investigates the impact of input data sequence length and image quality on both accuracy and inference time. The results show that scene image quality significantly affects both accuracy and inference time, while sequence length has a notable impact on accuracy. In addition, interpretability analysis comparing models with and without fine-tuning reveals the critical role of fine-tuning in enhancing the CoT reasoning capability of lightweight LVLM.

Future work will focus on deploying the proposed model in real-world vehicle platforms and enhancing its performance and inference efficiency through techniques such as quantization and pruning.

Appendix

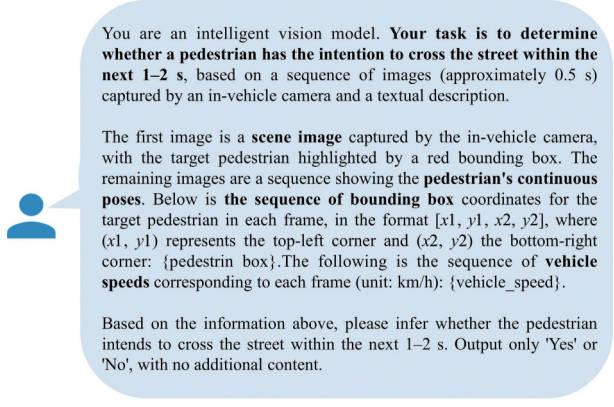


Fig. A1 Prompt template without CoT.

Author contributions

Yancheng Ling: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhenlin Qin**: Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Leizhen Wang**: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Zhendong Liu**: Writing – review & editing, Formal analysis, Conceptualization. **Yang Liu**: Writing – review & editing, Formal analysis, Conceptualization. **Zhenliang Ma**: Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Replication and data sharing

The source codes and replication package are available on ETS data at https://doi.org/10.26599/ETSD.2025.9190070.

Acknowledgements

This work was in part financially supported by Digital Futures.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Bai, J., Fang, J., Lv, Y., Lv, C., Xue, J., Li, Z., 2025. Gating syn-to-real knowledge for pedestrian crossing prediction in safe driving. *IEEE Trans Intell Transp Syst*, **26**, 7509–7522.

Bhattacharyya, A., Fritz, M., Schiele, B., 2018. Long-term on-board prediction of people in traffic scenes under uncertainty. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4194–4202.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al., 2020. Language Models are Few-shot Learners. https://arxiv.org/abs/2005.14165

Cadena, P. R. G., Qian, Y., Wang, C., Yang, M., 2022. Pedestrian graph: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Trans Intell Transp Syst*, **23**, 21050–21061.

Cadena, P. R. G., Yang, M., Qian, Y., Wang, C., 2019. Pedestrian graph: Pedestrian crossing prediction based on 2D pose estimation and graph convolutional networks. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2000–2005.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4724–4733.

Chen, L., Dong, T., Li, X., Xu, X., 2025a. Logistics engineering management in the platform supply chain: An overview from logistics service strategy selection perspective. *Engineering*, **47**, 236–249.

Chen, X., Xu, W., Zhang, S., Cai, Y., 2025b. Pedestrian crossing intention prediction via progressive multimodal token fusion for autonomous driving. *IEEE Trans Intell Transp Syst*, **26**, 12959–12973.

Deng, Y., Choi, Y., Shieber, S., 2024. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step. https://arxiv.org/abs/2405.14838

Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., Shieber, S., 2023. Implicit Chain of Thought Reasoning via Knowledge Distillation. https://arxiv.org/abs/2311.01460

Fu, Y., Peng, H., Ou, L., Sabharwal, A., Khot, T., 2023. Specializing smaller language models toward multistep reasoning. In: International Conference on Machine Learning, 10421–10430.

Gao, Z., Jia, B., Xie, D., Wang, W., Wu, J., 2025. A discussion on the complexity and transit mechanisms of urban traffic systems. *Engineering*, **44**, 24–29.

Gesnouin, J., Pechberti, S., Stanciulcscu, B., Moutarde, F., 2021. TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), 01–07.

Ham, J.-S., Huang, J., Jiang, P., Moon, J., Kwon, Y., Saripalli, S., Kim, C., 2026. Multimodal understanding with GPT-4o to enhance generalizable pedestrian behavior prediction. *Comput Electr Eng*, **129**, 110741.

Ham, J. S., Kim, S., Huang, J., Jiang, P., Moon, J., Saripalli, S., et al., 2025. LLaMAPed: Multi-modal pedestrian crossing intention prediction. In: Computer Vision – ECCV 2024 Workshops, 150–167.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

Ho, N., Schmid, L., Yun, S. Y., 2022. Large Language Models are Reasoning Teachers. https://arxiv.org/abs/2212.10071

Huang, J., Jiang, P., Gautam, A., Saripalli, S., 2024. GPT-4V takes the wheel: Promises and challenges for pedestrian behavior prediction. *Proc AAAI Symp Ser*, **3**, 134–142.

Kotseruba, I., Rasouli, A., Tsotsos, J. K., 2020. Do they want to cross? Understanding pedestrian intention for behavior prediction. In: 2020 IEEE Intelligent Vehicles Symposium (IV), 1688–1693.

Kotseruba, I., Rasouli, A., Tsotsos, J. K., 2021. Benchmark for evaluating pedestrian action prediction. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 1257–1267.

Ling, Y., Ma, Z., 2024a. Pedestrian crossing intention prediction in the wild: A survey. *CHAIN*, **1**, 263–279.

Ling, Y., Ma, Z., Xie, B., Zhang, Q., Weng, X., 2023a. SA-BiGCN: Bi-stream graph convolution networks with spatial attentions for the eye contact detection in the wild. *IEEE Trans Intell Transp Syst*, **25**, 2089–2100.

Ling, Y., Ma, Z., Zhang, Q., Xie, B., Weng, X., 2024b. PedAST-GCN: Fast pedestrian crossing intention prediction using spatial–temporal attention graph convolution networks. *IEEE Trans Intell Transp Syst*, **25**, 13277–13290.

Ling, Y., Zhang, Q., Weng, X., Ma, Z., 2023b. STMA-GCN_PedCross: Skeleton based spatial-temporal graph convolution networks with multiple attentions for fast pedestrian crossing intention prediction. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), 500–506.

Liu, X., Zhou, Y., Gou, C., 2023. Learning from interaction-enhanced scene graph for pedestrian collision risk assessment. *IEEE Trans Intell Veh*, **8**, 4237–4248.

Munir, F., Azam, S., Mihaylova, T., Kyrki, V., Kucner, T. P., 2025. Pedestrian vision language model for intentions prediction. *IEEE Open J Intell Transp Syst*, **6**, 393–406.

Nazari, F., Noruzoliaee, M., Mohammadian, A. K., 2025. Autonomous vehicle adoption behavior and safety concern: A study of public perception. *Multimodal Transp*, **5**, 100252.

Piccoli, F., Balakrishnan, R., Perez, M. J., Sachdeo, M., Nunez, C., Tang, M., et al., 2020. Fussi-net: Fusion of spatiotemporal skeletons for intention prediction network. In: 2020 54th Asilomar Conference on Signals, Systems, and Computers, 68–72.

Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J., 2019. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 6261–6270.

Rasouli, A., Kotseruba, I., Tsotsos, J. K., 2017. Are theygoing to cross A benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 206–213.

Rasouli, A., Kotseruba, I., Tsotsos, J. K., 2020. Pedestrian Action Anticipation Using Contextual Feature Fusion in Stacked RNNs. https://arxiv.org/abs/2005.06582

Sakib, N., Paul, T., Ahmed, M. T., Al, Momin, K., Barua, S., 2024. Investigating factors influencing pedestrian crosswalk usage behavior in Dhaka city using supervised machine learning techniques. *Multimodal Transp*, **3**, 100108.

Sharma, N., Dhiman, C., Indu, S., 2025a. Cross-modal pedestrian behavior prediction: A dual-task approach with progressive denoising attention and CVAE. *IEEE Trans Intell Transp Syst*, **26**, 17110–17120.

Sharma, N., Dhiman, C., Indu, S., 2025b. Predicting pedestrian intentions with multimodal IntentFormer: A co-learning approach. *Pattern Recognit*, **161**, 111205.

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., Woo, W. C., 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. https://arxiv.org/abs/1506.04214

Shridhar, K., Stolfo, A., Sachan, M., 2023. Distilling reasoning capabilities into smaller language models. In: Findings of the Association for Computational Linguistics: ACL, 7059–7073.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for Large-scale Image Recognition. https://arxiv.org/abs/1409.1556

Soleimani, M., Saria, A. A., 2025. Towards Autonomy: A Comprehensive Technical and Ethical Review of Automated Vehicle Safety. *Multimodal Transp*, **5**, 100272.

Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., et al., 2023. Gemini: A family of Highly Capable Multimodal Models. https://arxiv.org/abs/2312.11805

Thawakar, O., Dissanayake, D., More, K., Thawkar, R., Heakl, A., Ahsan, N., et al., 2025. LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs. https://arxiv.org/abs/2501.06186

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2016. Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), 4489–4497.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., 2022. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, 24824–24837.

Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., et al., 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-step. https://arxiv.org/abs/2411.10440

Xu, Z., Zheng, N., 2024. Integrating connected autonomous shuttle buses as an alternative for public transport–A simulation-based study. *Multimodal Transp*, **3**, 100133.

Yao, X., Ren, R., Liao, Y., Liu, Y., 2025. Unveiling the Mechanisms of Explicit CoT Training: How Chain-of-Thought Enhances Reasoning Generalization. https://arxiv.org/abs/2502.04667

Yang, B., Zhu, J., Hu, C., Yu, Z., Hu, H., Ni, R., 2024. Faster pedestrian crossing intention prediction based on efficient fusion of diverse intention influencing factors. *IEEE Trans Transp Electrif*, **10**, 9071–9087.

Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., Özgüner, Ü., 2022. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Trans Intell Veh*, **7**, 221–230.

Yang, Y., Zhan, J., Xu, M., Liu, Y., Qu, X., 2026. Toward climate-neutral urban mobility: understanding shared e-scooter carbon emission patterns through multi-city evidence in Europe. *Transp Res Part A Policy Pract*, **203**, 104736.

Yu, Y., 2024. Do LLMs Really Think Step-by-step in Implicit Reasoning? https://arxiv.org/abs/2411.15862

Yue-Hei, Ng J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G., 2015. Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 4694–4702.

Zhang, S., Zhang, J., Yang, L., Chen, F., Li, S., Gao, Z., 2024. Physics guided deep learning-based model for short-term origin–destination demand prediction in urban rail transit systems under pandemic. *Engineering*, **41**, 276–296.

Zhang, X., Angeloudis, P., Demiris, Y., 2022. ST CrossingPose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Trans Intell Transp Syst*, **23**, 20773–20782.

Zhou, Y., Liu, X., Guo, Z., Cai, M., Gou, C., 2024. HKTSG: A hierarchical knowledge-guided traffic scene graph representation learning framework for intelligent vehicles. https://doi.org/10.1109/TIV.2024.3384989

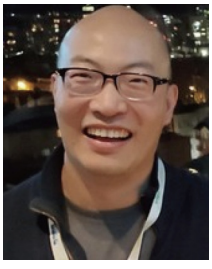
Zhou, Y., Tan, G., Zhong, R., Li, Y., Gou, C., 2023. PIT: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE Trans Intell Transp Syst*, **24**, 14213–14225.

Zhou, Y., Tang, J., Xiao, X., Lin, Y., Liu, L., Guo, Z., et al., 2025. Where, What, Why: Towards Explainable Driver Attention Prediction. https://arxiv.org/abs/2506.23088

Yancheng Ling received the B.S. degrees in traffic engineering and internet of things engineering from East China Jiaotong University in 2017, followed by the M.S. and Ph.D. degrees in traffic information engineering and control from South China University of Technology in 2019 and 2024, respectively. He was a visiting Ph.D. student at the Department of Civil and Architectural Engineering at KTH Royal Institute of Technology from 2022 to 2023. Currently, he is a Postdoctoral Researcher at KTH Royal Institute of Technology. His research interests primarily focus on intelligent transportation systems (ITSs), computer vision, natural language processing (NLP), and the application of large language models (LLMs) in transportation.

Zhenlin Qin received the B.S. and M.S. degrees from South China University of Technology in 2018 and 2021. He is pursuing the Ph.D. degree in transport science at KTH Royal Institute of Technology. His research interests include LLMs and individual mobility modeling.

Leizhen Wang received the M.S. degree in transportation engineering from Southeast University, China, in 2021. He is currently pursuing a Ph.D. degree in the Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Monash University, Australia. His research interests include large language models, reinforcement learning, and intelligent transport systems.



Zhendong Liu is a Researcher in rail vehicle technology at KTH Royal Institute of Technology. He has participated in more than 20 railway-related research projects and published 60 publications (conference and journal articles) about research on rail transport systems. His research interests cover traction and braking, electric power supply, driving optimization, pantograph-catenary dynamics, train thermal comfort, climate adaptation, circular economy, maglev trains and hyperloop systems.

Yang Liu is an Associate Research Fellow at the School of Vehicle and Mobility, Tsinghua University. He is a Recipient of the National High-Level Young Talents Program and a Marie Curie Fellow of the European Union. He has published over 40 papers as first/corresponding author, with more than 2600 google scholar citations and 5 ESI highly cited papers. He serves as an editorial board member of *Transportation Research Part E: Logistics and Transportation Review*, Associate Editor of *IEEE Transactions on Intelligent Vehicles* and Co-Chair of the WTC Technical Committee on Intelligent Driving and Mobility Services.

Zhenliang Ma is an Associate Professor in road traffic engineering at KTH Royal Institute of Technology. His research is mainly involved in statistics, machine learning, computer science-based modeling, simulation, optimization and control within the framework of selected mobility-related complex systems, which are intelligent transport systems and multimodal mobility systems.