Commun Transp Res

Research Article

Vision Language Model–Guided Semantic Distillation Network for Pedestrian Crossing Intention Prediction

Yuanzhe Li[1,✉], Cheng Wang[2], Zhuoqi Zeng[3], Steffen Müller[1]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640061

ABSTRACT: Pedestrian crossing intention prediction is a critical function of autonomous vehicles, which enables safe pedestrian–vehicle interactions in complex urban environments. However, accurately predicting pedestrian intention remains challenging, as it depends on multiple factors, including pedestrian behavior, vehicle dynamics, and surrounding context. To improve prediction performance, existing approaches increasingly incorporate more modalities, but this often results in substantial computational overhead. Achieving accurate intention prediction with a limited number of modalities remains an open challenge. In this paper, we propose a vision language model-guided semantic distillation network (VLM-GSDN) for pedestrian crossing intention prediction. During training, VLM-GSDN leverages pedestrian- and scene-level semantic descriptions generated by a vision language model to align local and global visual representations with their corresponding high-level semantics, while requiring only RGB images and image-space pedestrian motion for efficient inference. To incorporate motion cues, we introduce motion-guided modulation and predictive spatial modulation, which refine pedestrian representations using pedestrian motion and predicted future displacements. A selective temporal attention module is further incorporated to integrate temporal information and emphasize informative frames. Moreover, a multi-task learning strategy is adopted in which pedestrian action-state transition prediction serves as an auxiliary task to provide additional motion supervision. Extensive experiments on the JAAD and PIE datasets demonstrate that VLM-GSDN achieves competitive performance compared with state-of-the-art methods. Moreover, it exhibits promising cross-dataset generalization and maintains robust performance over longer prediction horizons.

KEYWORDS: pedestrian intention prediction; vision language model; semantic distillation; multi-task learning

Just Accepted

1 Introduction

Advances in artificial intelligence and sensing technologies have accelerated the deployment of autonomous vehicles (AVs), with the prospect of improving transportation safety and efficiency. However, frequent interactions between AVs and vulnerable road users, particularly pedestrians, pose significant safety challenges in urban environments. AVs should possess the capability to anticipate pedestrians' crossing intentions by integrating temporal behavioral cues with environmental context, analogous to experienced human drivers. Nevertheless, this task remains challenging, as pedestrian behavior is inherently stochastic and strongly influenced by contextual factors such as road geometry, dynamic traffic conditions, and interactions with surrounding road users (Li et al., 2026).

Extensive research has focused on predicting pedestrian intentions via onboard vision systems. Datasets such as JAAD (Rasouli et al., 2017) and PIE (Rasouli et al., 2019) provide real-world video recordings that support the analysis of pedestrian crossing behavior. Early studies primarily relied on unimodal inputs, such as visual appearance (Varytimidis et al., 2018), trajectory history (Saleh et al., 2018), or pose information (Wang and Papanikolopoulos, 2020), leveraging deep learning to extract intention-relevant cues. However, these approaches often struggle to generalize in complex traffic environments, where pedestrian behavior is heavily influenced by dynamic contextual factors and intricate interaction patterns. For example, approaches that rely solely on pedestrian pose information may overlook crucial contextual factors, such as traffic signs, that can substantially influence pedestrian crossing behavior.

Subsequent studies incorporated diverse visual inputs, including semantic maps (Yang et al., 2022), optical flows (Munir and Kucner, 2025), and depth maps (Li and Müller, 2025b), together with motion-related signals such as bounding boxes (Osman et al., 2024) and vehicle speed (Kotseruba et al., 2021), to better capture dynamic context and interaction patterns. This trend has motivated the development of sophisticated multimodal fusion approaches for integrating heterogeneous modalities. For example, PCPA (Kotseruba et al., 2021) employs separate recurrent neural network (RNN) for vehicle speed, skeleton data, and bounding box, while using a 3D convolutional neural network (CNN) to extract visual features. Temporal and modality attention are then applied to fuse these modality-specific features for pedestrian intention modeling. MTC (Li and Müller, 2025b) introduces cross-modality guided attention to enhance feature complementarity through bidirectional interactions among spatially aligned visual modalities at the local pedestrian and global scene levels. PTINet (Munir and Kucner, 2025) adopts a context-aware multi-task framework that jointly predicts

[1] Chair of Automotive Engineering, Technical University of Berlin, Berlin 13355, Germany. [2] School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK. [3] School of Engineering, Hainan Bielefeld University of Applied Sciences, Danzhou 571700, China.

✉ Corresponding author. E-mail: yuanzhe.li@campus.tu-berlin.de

Received: April 12, 2026; Revised: July 30, 2026; Accepted: October 23, 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/).

pedestrian intention and trajectory by integrating past trajectories, local contextual attributes, RGB images, and optical flow.

Despite the success of multimodal fusion approaches, several limitations remain. First, high-level semantic cues related to pedestrian behavior and scene context, such as hand gestures, gaze direction, traffic signals, and crosswalk markings, are often not explicitly modeled and must instead be learned implicitly from raw sensory inputs. Second, some visual modalities must be obtained using additional pretrained models. For example, Global PCPA (Yang et al., 2022) uses DeepLabV3 (Chen et al., 2017) to generate semantic maps, while MTC (Li and Müller, 2025b) employs Depth Anything V2 (Yang et al., 2024) to estimate depth maps, increasing computational cost and the risk of error propagation. Finally, integrating numerous heterogeneous modalities requires complex feature extraction and multimodal fusion pipelines, thereby increasing computational cost and hindering practical deployment.

Recent advances in vision language models (VLMs) have demonstrated great potential for pedestrian intention prediction due to their strong scene understanding and reasoning capabilities. One line of research comprises VLM-based approaches that directly employ VLMs for the prediction task. Huang et al. (2024) and Ham et al. (2024, 2026) evaluate open-source and proprietary VLMs in zero-shot settings. Ling et al. (2026) further use chain-of-thought (CoT) reasoning data generated by Gemini-2.5 to fine-tune a lightweight open-source VLM. Another line of research comprises language-enhanced approaches that incorporate VLM-derived semantics as complementary inputs or auxiliary guidance for intention prediction. TAMformer (Osman et al., 2024) employs BLIP-2 to generate frame-wise textual descriptions, which are integrated with local image, pose, and bounding-box features for early pedestrian action prediction. GLCCL (Lin and Hu, 2026) uses VLM-generated global scene descriptions as auxiliary guidance to enhance motion representations, while TCP (Xiao et al., 2025) aligns trajectory features with semantics derived from VLM-generated pedestrian behavior descriptions during training.

Nevertheless, both VLM-based and language-enhanced approaches still have several limitations. VLM-based approaches typically rely on large-scale VLMs, which may incur relatively high computational cost and inference latency and can also be susceptible to hallucinations. For language-enhanced approaches, using VLM-generated descriptions as input modalities, as in TAMformer, introduces additional overhead for text generation and processing. GLCCL and TCP instead use VLM-generated semantics as auxiliary supervision during training but primarily focus on either global scene context or local pedestrian behavior, without fully exploiting the complementarity between these two semantic levels.

To address these limitations, this paper proposes a vision language model-guided semantic distillation network (VLM-GSDN) for pedestrian crossing intention prediction. VLM-GSDN leverages a VLM to generate complementary pedestrian- and scene-level semantic descriptions, providing more comprehensive high-level semantic cues than existing methods that focus on a single semantic level. Importantly, these semantic descriptions are used only during training for language-guided distillation, while the VLM is not required at inference, thereby introducing no additional inference overhead. VLM-GSDN subsequently relies only on RGB images and image-space pedestrian motion for inference, without requiring additional derived visual modalities, thereby reducing computational cost. The main contributions are as follows:

(1) A novel pedestrian intention prediction framework is proposed, incorporating a language-guided distillation module that distills VLM-generated pedestrian- and scene-level semantics into local and global visual representations during training. At inference, the framework requires only RGB images and pedestrian motion, enabling efficient deployment.

(2) A motion-guided modulation module is introduced to enrich pedestrian visual features with observed motion cues, while a predictive spatial modulation module uses predicted future displacements to refine the temporally aggregated representation, thereby producing motion-aware pedestrian representations.

(3) A selective temporal attention module is introduced to selectively aggregate sequential visual features through a learnable summary token and asymmetric cross-attention, thereby highlighting salient frames and reducing computational overhead.

(4) A multi-task learning strategy is adopted, where pedestrian action-state transition prediction is employed as an auxiliary task to provide additional motion supervision.

(5) Extensive evaluations on the JAAD and PIE datasets demonstrate that VLM-GSDN achieves competitive performance compared with state-of-the-art methods while maintaining robust performance over longer prediction horizons. In addition, ablation studies verify the effectiveness of the proposed network architecture.

2 Related works

2.1. Pedestrian crossing intention prediction

Pedestrian crossing intention prediction has gained increasing attention in recent years, with many methods developed and evaluated on benchmark datasets such as JAAD (Rasouli et al., 2017) and PIE (Rasouli et al., 2019), which provide real-world video recordings. Early work mainly focused on static contextual cues. For instance, Varytimidis et al. (2018) predicts crossing intention using features from cropped pedestrian head and leg regions, while Wang and Papanikolopoulos (2020) infer intention from single-frame pose keypoints using a shallow neural network. However, relying solely on static cues ignores temporal continuity, making these methods less effective in handling intermittent motion patterns and short-term pauses in pedestrian behavior.

Subsequent work has shifted toward spatio-temporal modeling based on a single modality. For example, Lorenzo et al. (2020) combine a convolutional autoencoder with RNN to learn visual-temporal dependencies from image sequences. Pedestrian Graph (Cadena et al., 2019) and STMA-GCN (Ling et al., 2023) instead model pose-based temporal dynamics using graph convolutional networks (GCNs), with STMA-GCN further enhancing spatio-temporal representation learning through multiple attention mechanisms. By operating on low-dimensional pose keypoints rather than raw images, these methods reduce computational

Just Accepted

overhead and improve real-time efficiency. However, single-modality methods often struggle in complex scenarios, where pedestrian intention is influenced by vehicle dynamics, environmental context, and social interactions, which are not fully observable from any single modality.

Recent studies have shown that multimodal fusion can improve prediction performance by capturing richer contextual and behavioral cues, motivating the development of architectures that exploit complementary information across heterogeneous modalities. For instance, PCPA (Kotseruba et al., 2021) encodes non-visual modalities (bounding boxes, pose keypoints, and vehicle speed) with independent RNNs and extracts visual features using a 3D CNN. Multimodal fusion is then achieved through temporal attention and modality attention, enabling the model to quantify the importance of different frames and modalities. A subsequent extension (Yang et al., 2022) further incorporates semantic maps as an additional modality to provide structured scene context, including road geometry and traffic semantics. DPCIAN (Yang et al., 2023) uses a dual-channel architecture, extracting spatio-temporal features from skeletons via multi-scale GCN layers and modeling scene-traffic object interactions with a dedicated relation module. Gated-S2R-PCP (Bai et al., 2025) transfers multimodal knowledge from synthetic data through knowledge distillation, style transfer, and distribution alignment, followed by adaptive gated fusion for pedestrian intention prediction. PTINet (Munir and Kucner, 2025) introduces a context-aware multi-task framework that jointly predicts pedestrian intention and trajectory by integrating past trajectories, local contextual attributes, and global visual and motion features. Several studies have introduced attention-based spatial fusion mechanisms to exploit complementary information across visual modalities. ACIT (Li and Müller, 2025a) employs an attention-guided visual modality interaction module, using optical flow to highlight salient regions in RGB images and semantic maps. MTC (Li and Müller, 2025b) further introduces cross-modality guided attention to enable bidirectional interactions between depth and RGB features, as well as between depth and semantic features.

Recently, rapid advances in vision language models (VLMs) have attracted increasing attention to their application in pedestrian intention prediction. One line of research focuses on VLM-based approaches that directly leverage the multimodal reasoning capabilities of VLMs to understand traffic-scene context, analyze pedestrian behavior, and infer pedestrian intentions. LLaMAPed (Ham et al., 2024), GPT-4V-PBP (Huang et al., 2024), and OmniPredict (Ham et al., 2026) employ VideoLLaMA2, GPT-4V, and GPT-4o, respectively, in a zero-shot setting. They use task-specific prompts that explicitly define the input modality and task requirements, enabling contextual reasoning for intention inference. Azarmi et al. (2025a) employ GPT-4V in a zero-shot setting while optimizing hierarchical prompts through an automatic prompt engineering framework. Hussien et al. (2025) employ GPT-4 within a retrieval-augmented generation framework to generate interpretable explanations by incorporating retrieved contextual knowledge into the prompt, without task-specific fine-tuning. However, off-the-shelf VLMs used in zero-shot settings lack task-specific adaptation and generally achieve lower predictive accuracy than dedicated prediction models. To address this limitation, Munir et al. (2025) propose PedVLM, which fine-tunes a CLIP-based vision encoder and a T5 language model on multimodal inputs, including RGB images, optical flow, and textual prompts, to predict pedestrian intentions and generate interpretable explanations. Mishra et al. (2026) propose PedestrianQA, a video-based question-answering benchmark, and employ LoRA-based instruction fine-tuning to adapt an open-source VLM for intention prediction. Ling et al. (2026) propose VLMPed-CoT, which distills CoT reasoning data generated by Gemini-2.5 and employs a two-stage fine-tuning strategy to adapt the lightweight Qwen2.5-VL-3B model to this task. Overall, VLM-based approaches can provide interpretable explanations, but directly using large-scale VLMs for intention prediction incurs high computational cost and inference latency, while posing risks of hallucination.

Another line of research focuses on language-enhanced approaches that incorporate VLM-derived semantics into dedicated prediction models. Osman et al. (2024) employ BLIP-2 to generate frame-wise textual descriptions, which are encoded by CLIP and integrated with local image, pose, and bounding-box features in a multimodal Transformer for early pedestrian action prediction. Beyond their use as explicit input modalities, VLM-derived semantics can also serve as semantic anchors or auxiliary supervision. Uziel and Bialer (2025) propose ClipCross, which uses optimized CLIP text embeddings as semantic anchors to extract high-level semantics from pedestrian-focused visual representations for intention prediction. Lin and Hu (2026) propose GLCCL, which leverages VLM-generated global traffic-scene descriptions as auxiliary guidance within a global-local context co-learning framework to enhance motion representations derived from pedestrian bounding-box trajectories, pose, and ego-vehicle speed. TCP (Xiao et al., 2025) uses a VLM to generate fine-grained pedestrian behavior descriptions and aligns trajectory features with the resulting text embeddings during training.

Compared with multimodal fusion approaches relying on heterogeneous inputs, VLM-GSDN adopts a compact inference configuration requiring only RGB images and pedestrian motion. Unlike approaches that directly use large-scale VLMs for intention prediction, VLM-GSDN uses VLM-generated semantic descriptions only during training and transfers their high-level semantics into visual representations through language-guided distillation, thereby avoiding online VLM-based semantic description generation at inference. Furthermore, whereas GLCCL (Lin and Hu, 2026) and TCP (Xiao et al., 2025) primarily exploit scene- and pedestrian-level semantics, respectively, VLM-GSDN jointly distills both semantic levels into visual representations, providing more comprehensive semantic supervision. This design enables efficient intention prediction with low computational overhead.

2.2. Vision language models for traffic scene understanding

VLMs, such as Qwen (Bai et al., 2025), MiniCPM (Yu et al., 2026), and Phi-3 (Abdin et al., 2024), have demonstrated strong multimodal understanding capabilities and are increasingly adopted to traffic scene understanding (TSU) for high-level

Just Accepted

semantic description generation and contextual reasoning. For example, DriveVLM (Tian et al., 2024) leverages Qwen-VL to enhance driving scene understanding and planning by using multi-view images as input and a CoT reasoning mechanism to generate scene descriptions, situation analyses, and hierarchical planning outputs. DriveVLM-Dual further integrates the VLM-based reasoning module into conventional autonomous driving pipelines, aiming to address bottlenecks in spatial reasoning while improving computational efficiency. InternDrive (Zhang and Nie, 2024) applies visual instruction tuning to InternVL-1.5 and uses structured queries to enhance holistic understanding of driving scenes. TrafficPerceiver (Kuang et al., 2026) combines instruction-driven scene interpretation with object-level segmentation and employs reinforcement learning to improve robustness under adverse and visually degraded traffic conditions. ScVLM (Shi et al., 2025) proposes a hybrid framework that combines supervised and contrastive learning and uses event classification to support narrative generation for safety-critical driving events, thereby improving description accuracy and mitigating VLM hallucinations.

Unlike many existing pedestrian intention prediction methods, which primarily infer scene context from raw sensory inputs, we leverage the scene understanding capabilities of VLMs to extract high-level semantics of the global traffic scene, including traffic signs, crosswalks, and traffic lights that encode right-of-way information. Such semantics can be difficult to capture effectively from raw sensory inputs alone.

2.3. Vision language models for pedestrian behavior understanding

VLMs have shown strong capability in capturing fine-grained pedestrian behaviors. Recent studies have increasingly focused on pedestrian behavior understanding, aiming to capture context-dependent behaviors for safe interactions between AVs and pedestrians. To support such studies, Kong et al. (2024) introduces the WTS dataset, a multi-view traffic video benchmark with fine-grained pedestrian behavior annotations, rich textual descriptions, and 3D gaze cues, along with a large language models (LLMs)-based caption metric to evaluate dense video-to-text event understanding. Bossen et al. (2025) evaluates the capability of advanced VLMs to describe and classify dynamic nonverbal pedestrian behaviors, such as traffic gestures, under a zero-shot setting. Gao et al. (2025) distill semantic knowledge from GPT-4V into vision networks to recognize diverse pedestrian attributes, behaviors, and surrounding scene context. Choi et al. (2026) introduces a LLaVA-based vision-language behavior detector that uses prompt-engineering schemes, including zero-shot CoT with self-consistency and few-shot prompting, to identify pedestrians distracted by texting while walking. Tran et al. (2025) uses GPT-4V as a diagnostic tool to analyze pedestrian gesture misrecognition in pedestrian-vehicle interactions, identifying recurring failure factors associated with visibility and contextual information.

Unlike many existing pedestrian intention prediction methods that implicitly learn pedestrian behavioral cues from raw sensory inputs, we leverage the pedestrian behavior understanding capabilities of VLMs to explicitly generate fine-grained behavioral semantics, such as motion states, spatial positions, and gaze directions, providing rich intention-relevant cues that can be difficult to model explicitly from raw sensory inputs alone.

3 Proposed method

3.1. Problem formulation

We formulate pedestrian crossing intention prediction as a short-horizon anticipation task. Given a 0.5 second front-view video clip comprising T_{obs}=16 consecutive frames sampled at 30 Hz and captured by the ego-vehicle, the task is to predict whether the target pedestrian will cross the street, with a prediction horizon of 1-2 s relative to the event time. The event time is defined as the crossing onset for positive samples and, for negative samples, the last observable frame.

At inference, VLM-GSDN leverages three input streams. (1) Local RGB image sequence $L=\left[L^{t-T_{\text{obs}}+1},L^{t-T_{\text{obs}}+2},...,L^{t}\right]$, which is cropped from the original frames according to the pedestrian's bounding box and captures the pedestrian's appearance as well as fine-grained behavioral cues, such as body posture, motion patterns, gestures, and head orientation. (2) Global RGB image sequence $G=\left[G^{t-T_{\text{obs}}+1},G^{t-T_{\text{obs}}+2},...,G^{t}\right]$, which consists of the raw frames captured by the ego-vehicle camera and provides scene-level visual context, including traffic signs, crosswalks, and surrounding traffic participants. (3) Pedestrian motion sequence $M=\left[m^{t-T_{\text{obs}}+1},m^{t-T_{\text{obs}}+2},...,m^{t}\right]$, where $m^{\tau}\in\mathbb{R}^{4}$ denotes the image-space pedestrian motion at time step τ, thereby characterizing the pedestrian's dynamic motion cues. Following Rasouli and Kotseruba (2023), m^{τ} is formulated as the frame-to-frame differences in the pedestrian bounding-box coordinates:

$$m^{\tau}=\left(x_{tl}^{\tau}-x_{tl}^{\tau-1},y_{tl}^{\tau}-y_{tl}^{\tau-1},x_{br}^{\tau}-x_{br}^{\tau-1},y_{br}^{\tau}-y_{br}^{\tau-1}\right) \quad (1)$$

where $(x_{tl}^{\tau},y_{tl}^{\tau})$ and $(x_{br}^{\tau},y_{br}^{\tau})$ denote the top-left and bottom-right corners of the pedestrian bounding box at time step τ, respectively.

Following common practice in related work, the pedestrian bounding-box annotations provided by the datasets are used to construct the local RGB and motion sequences. As this study focuses on pedestrian intention prediction rather than detection and tracking, this setting ensures a consistent evaluation protocol and enables fair comparison with existing methods.

During training, two semantic descriptions are additionally introduced as auxiliary inputs for semantic distillation: (1) Pedestrian-level semantic description D_L, capturing the pedestrian's location, motion, pose, head orientation, and related behavioral cues. (2) Scene-level semantic description D_G, capturing the surrounding traffic scene, including traffic lights, traffic signs, crosswalks, and other contextual elements.

The architecture of VLM-GSDN is illustrated in Fig. 1, and its core modules are described in Sections 3.2-3.10.

3.2. Semantic description generation

Visual features extracted by visual backbones often lack explicit high-level semantics, which may limit their ability to capture global traffic context and pedestrian behavior effectively. In contrast,

Just Accepted

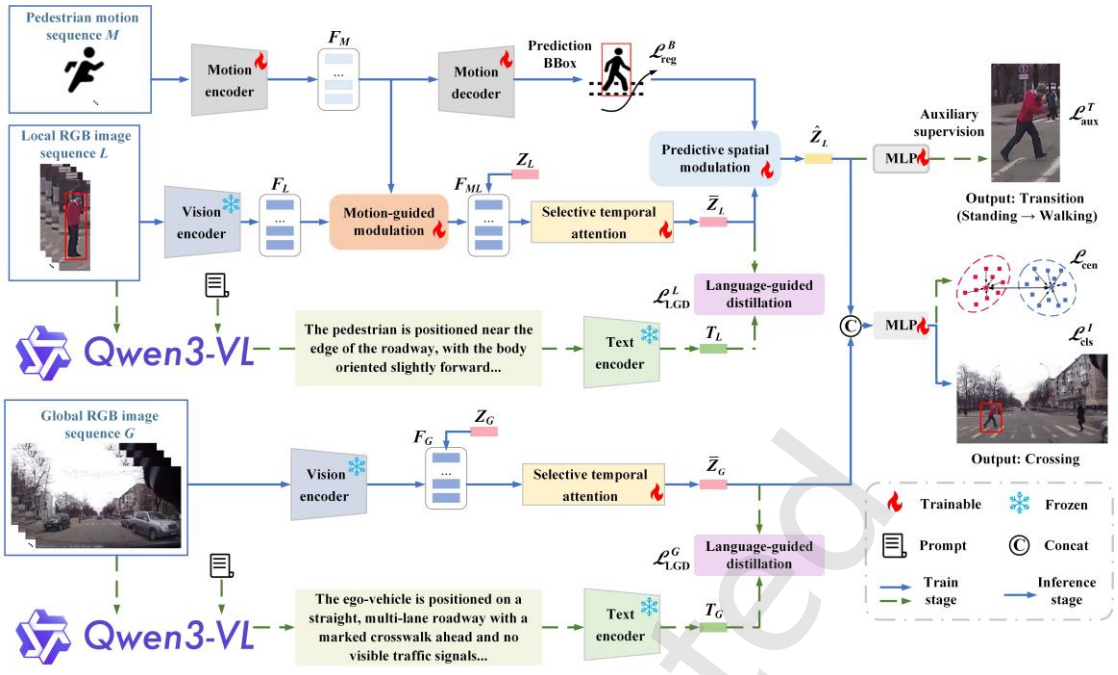


Fig. 1. Network architecture of the proposed VLM-GSDN.

VLMs have shown strong capabilities in understanding both holistic driving scenes (Tian et al., 2024) and fine-grained pedestrian behaviors (Bossen et al., 2025). Motivated by this, we use VLM-generated semantic descriptions during training as auxiliary semantic supervision to align visual representations with high-level semantics.

To obtain semantic descriptions, we propose a prompt-driven generation pipeline, as illustrated in Fig. 2. Instead of feeding images directly into the VLM with generic instructions, we adopt a hierarchical guidance strategy to guide the VLM toward generating informative and task-relevant descriptions. Specifically, we first leverage Qwen3-Max-Thinking (Yang et al., 2025), which exhibits strong reasoning capabilities, to identify critical factors that influence pedestrian crossing decisions from two perspectives: the pedestrian-level perspective (Fig. 2(a)) and the scene-level perspective (Fig. 2(b)). By reasoning from the viewpoint of a professional driver, Qwen3-Max-Thinking organizes these factors into two structured factor sets, corresponding to pedestrian-level and scene-level factors. For pedestrian-level, the factor set includes pedestrian position and road relation, motion cues, posture, orientation, and other behavior-related attributes. For scene-level, the factor set includes road layout and ego-vehicle position, critical traffic elements, and the spatial distribution of traffic participants. The definitions of these factors and their corresponding contents are illustrated in Fig. 2.

The identified key factors are further processed through a Qwen3-Max-Thinking-based prompt generation process, which reformulates them into structured, task-specific natural-language prompts rather than simply concatenating them as isolated keywords. Specifically, the generated prompt consists of several elements, including a predefined role, a task description, relevant constraints, and the selected key factors. Accordingly, this process yields two structured prompts that separately incorporate pedestrian-behavior and scene-context factors, guiding the VLM to generate informative and task-relevant semantic descriptions.

Finally, based on the pedestrian-level prompt P_L and the scene-level prompt P_G, we employ the open-source Qwen3-VL-8B (Bai et al., 2025), which demonstrates strong visual understanding capability, to produce semantic descriptions from two complementary perspectives. Specifically, conditioned on P_L, Qwen3-VL-8B takes five pedestrian images as input and generates pedestrian-level semantic description D_L, while conditioned on P_G, it takes the global scene image from the last frame of the observation window as input and generates the scene-level semantic description D_G. This process can be formulated as follows:

$$D_L = \mathrm{VLM}(L^{t-4\Delta}, \ldots, L^{t-\Delta}, L^t, P_L) \tag{2}$$

$$D_G = \mathrm{VLM}(G^t, P_G) \tag{3}$$

where Δ denotes the sampling interval between adjacent local RGB image frames and is set to 3. VLM refers to the pretrained Qwen3-VL-8B model.

Representative examples, together with the corresponding prompts and generated semantic descriptions, are shown in Fig. 3. The generated semantic descriptions D_L and D_G contain rich and complementary semantic information. Specifically, D_L focuses on fine-grained pedestrian-level cues, such as position relative to road elements, motion patterns, posture, and head orientation, as shown in Fig. 3(a), while D_G captures holistic scene context, including the ego-vehicle position, road layout, traffic control devices, and spatial distribution of surrounding participants, as shown in Fig. 3(b).

3.3. Vision encoder

RGB images provide informative cues about pedestrian behavior and surrounding context and can be directly captured by onboard cameras, unlike additional visual modalities derived from pretrained models, such as semantic maps (Yang et al., 2022) and depth maps (Li and Müller, 2025b), making RGB images more

Just Accepted

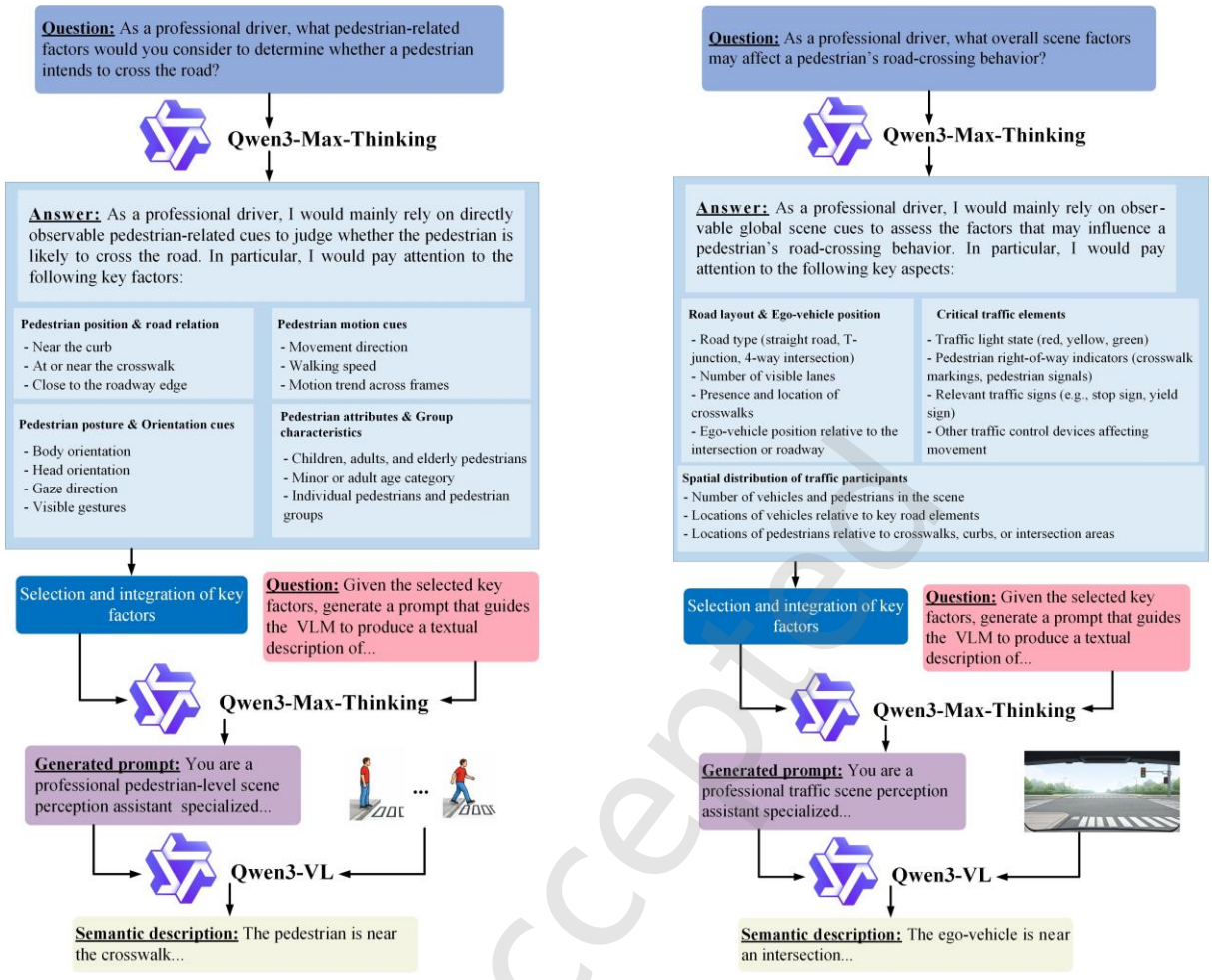


Fig. 2. Pipeline of prompt-driven semantic description generation: (a) pedestrian-level pipeline; (b) scene-level pipeline.

practical and computationally efficient. To preserve complementary information across local and global visual contexts, we employ a vision encoder to separately encode local and global RGB image sequences into compact visual features.

We adopt DINOv2-Reg (Darcet et al., 2023) as the visual backbone due to its strong representation capability and effectiveness across a wide range of computer vision tasks. To balance performance and computational cost, we use the DINOv2-Reg-Base model pretrained on the LVD-142M dataset. All visual inputs are resized to 224×224, consistent with the backbone's input resolution. For each frame, a compact visual feature is obtained by average pooling the patch tokens from the final layer of DINOv2-Reg, excluding the CLS token and the four register tokens. Formally, given a local or global RGB image sequence I, where $I \in \{L, G\}$, the vision encoder produces the corresponding visual feature sequence $F_I \in \mathbb{R}^{T_{\text{obs}} \times 768}$, where F_L and F_G denote the local and global visual feature sequence, respectively:

$$F_I = \text{Encoder}_{\text{image}}(I) \tag{4}$$

where $\text{Encoder}_{\text{image}}$ denotes the pre-trained DINOv2-Reg backbone.

3.4. Text encoder

The text encoder is designed to encode the semantic descriptions into compact text representations. Transformer-based text encoders have shown strong capability in modeling complex linguistic semantics. Accordingly, we adopt the text encoder of the pre-trained Long-CLIP model (Zhang et al., 2024), which preserves CLIP's vision-language alignment capability while extending the maximum input length to support longer text sequences. This is particularly beneficial for our task, as the semantic descriptions generated by the VLM often exceed CLIP's maximum input length (Radford et al., 2021).

Given the pedestrian-level or scene-level semantic descriptions D_i, where $i \in \{L, G\}$, they are first tokenized and then processed by the Transformer layers of the Long-CLIP text encoder to obtain the corresponding text representation. Formally, the text representation $T_i \in \mathbb{R}^{512}$ is computed as:

$$T_i = \text{Encoder}_{\text{text}}(D_i) \tag{5}$$

where $\text{Encoder}_{\text{text}}$ denotes the pre-trained Long-CLIP text encoder.

3.5. Motion-guided modulation

The motion-guided modulation (MGM) module leverages frame-level pedestrian motion to adaptively inject motion information into the local visual feature sequence F_L, producing motion-aware representations. A motion encoder is designed to first embed the pedestrian motion sequence and then perform temporal modeling to capture informative motion patterns, as illustrated in Fig. 4.

Just Accepted

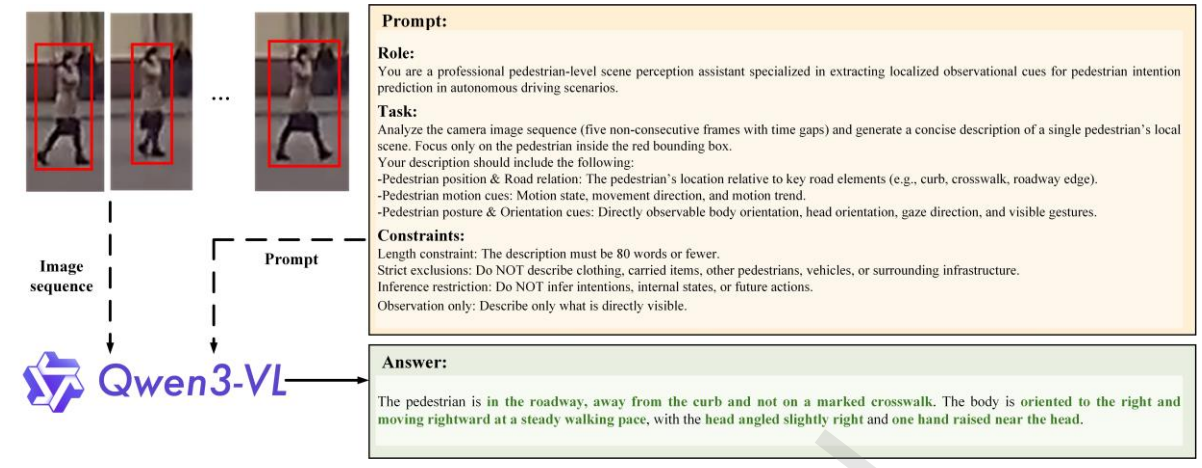


Fig. 3. Examples of prompt-driven semantic descriptions for pedestrian- and scene-level understanding: (a) pedestrian-level semantic description; (b) scene-level semantic description.

Specifically, the pedestrian motion sequence $M \in \mathbb{R}^{T_{\text{obs}} \times 4}$ is first projected into a high-dimensional representation $E_M \in \mathbb{R}^{T_{\text{obs}} \times 128}$ through an embedding layer, and positional encoding (PE) is then added to incorporate temporal information, resulting in $\bar{M} = E_M + \text{PE}$.

The motion encoder is composed of L Transformer encoder layers, each comprising a multi-head self-attention (MHSA) sublayer followed by a feed-forward network (FFN) sublayer. For l-th layer, the input motion feature $\bar{M}^{l-1}$ is first processed by the MHSA sublayer, followed by layer normalization and residual connection, which can be expressed as:

$$\hat{M}^{l} = \text{Norm}\left(\text{MHSA}(\bar{M}^{l-1})\right) + \bar{M}^{l-1} \tag{6}$$

Then $\hat{M}^{l}$ is fed into the FFN sublayer, followed by layer normalization and residual connection, formulated as:

$$\bar{M}^{l} = \text{Norm}\left(\text{FFN}(\hat{M}^{l})\right) + \hat{M}^{l} \tag{7}$$

The final output of the motion encoder, denoted as $F_M = \left[f_M^{t-T_{\text{obs}}+1}, \ldots, f_M^{t-1}, f_M^{t} \right]$, is a temporally enriched motion feature sequence, where $f_M^{\tau} \in \mathbb{R}^{128}$ represents the motion feature at time step τ. Subsequently, MGM modulates F_L via a feature-wise linear modulation (FiLM) mechanism, as illustrated in Fig. 5. Specifically, for each time step τ, the motion feature f_M^{τ} is fed into a FiLM generator to produce the modulation parameters $\gamma^{\tau} \in \mathbb{R}^{768}$ and $\beta^{\tau} \in \mathbb{R}^{768}$, formulated as:

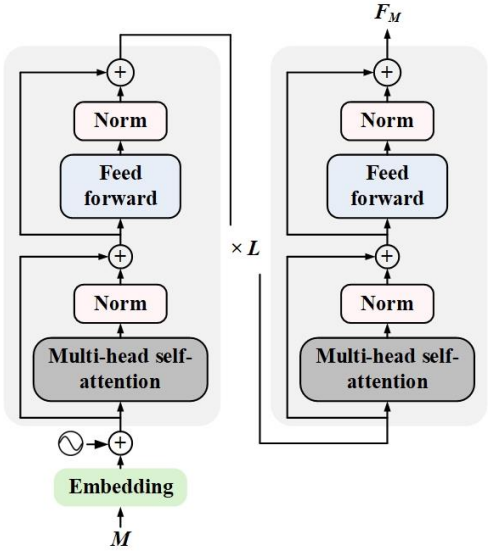


Fig. 4. Architecture of motion encoder.

Just Accepted

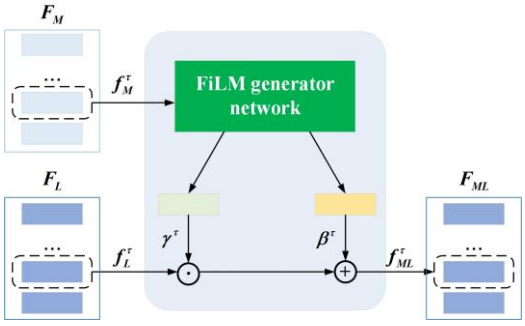


Fig. 5. Architecture of feature-wise linear modulation.

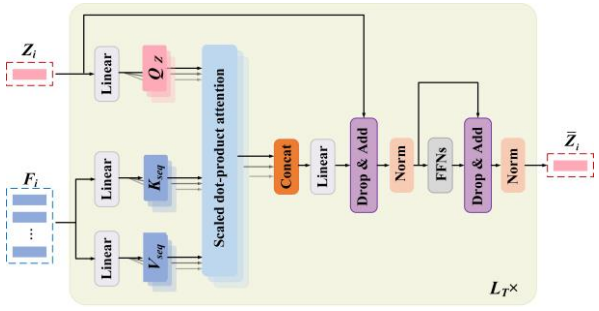


Fig. 6. Architecture of selective temporal attention.

$$\gamma^\tau = W_\gamma f_M^\tau + b_\gamma \tag{8}$$

$$\beta^\tau = W_\beta f_M^\tau + b_\beta \tag{9}$$

where W_γ and W_β are learnable projection matrices, γ^τ and β^τ are modulation parameters, and b_γ and b_β are the corresponding bias terms. Given the local visual feature $f_L^\tau \in \mathbb{R}^{768}$, the modulated local visual feature $f_{ML}^\tau \in \mathbb{R}^{768}$ is computed as:

$$f_{ML}^\tau = \gamma^\tau \odot f_L^\tau + \beta^\tau \tag{10}$$

where $\odot$ denotes element-wise multiplication.

3.6. *Selective temporal attention*

The selective temporal attention (STA) module is designed to integrate visual feature sequences while emphasizing salient frames. Specifically, summary tokens Z_L and Z_G are inserted into visual feature sequence F_{ML} and F_G, respectively. As illustrated in Fig. 6, STA adopts an asymmetric cross-attention architecture in which the interaction is constrained to a unidirectional information flow from feature sequence to summary token, allowing the summary token to selectively capture informative temporal cues while preserving a clear measure of each frame's contribution. Unlike conventional self-attention, where the query, key, and value are all derived from the input sequence, STA reduces computational cost by using only the summary token as the query, while projecting the key and value from the visual feature sequence. STA consists of multiple layers, each comprising a multi-head cross-attention (MHCA) followed by an FFN. The output of the h-th attention head is computed as:

$$H^h = \mathrm{Softmax}\left(\frac{Q_Z^h \cdot K_{seq}^h}{\sqrt{d}}\right) \cdot V_{seq}^h \tag{11}$$

where Q_Z^h, K_{seq}^h, and V_{seq}^h denote the query, key, and value projections of the h-th attention head, respectively, and d is the feature dimension used for scaling.

The summary token Z_i, where $i \in \{L, G\}$, assigns higher attention weights to informative frames while suppressing less relevant or redundant ones. The outputs of all attention heads are concatenated and projected through a multilayer perceptron (MLP).

$$\mathrm{MHCA}(Q_Z, K_{seq}, V_{seq}) = \mathrm{MLP}\left[H^1, H^2, ..., H^n\right] \tag{12}$$

The resulting features are further refined by the FFN with residual connection and layer normalization. Notably, only the summary token Z_i is iteratively updated, while the visual feature sequence remains unchanged throughout the attention process. By stacking multiple STA layers, the updated summary token $\bar{Z}_i$ progressively encodes salient temporal information and serves as a compact temporal representation.

3.7. *Predictive spatial modulation*

The predictive spatial modulation (PSM) module predicts the pedestrian's future bounding-box displacement and uses the resulting spatial cue to modulate the aggregated temporal representation $\bar{Z}_L$, enriching it with future motion information. The prediction horizon is set to 1 second. Specifically, a motion decoder composed of linear layers takes the last-frame motion feature F_M^t as input and predicts the future bounding-box displacement $B_{\mathrm{pre}} \in \mathbb{R}^4$ relative to the last observed frame. The decoder is optimized using mean squared error (MSE) loss.

$$B_{\mathrm{pre}} = \mathrm{Decoder}(F_M^t) \tag{13}$$

$$\mathrm{L}_{\mathrm{reg}}^B = \frac{1}{N}\sum_{n=1}^{N}\left\| B_{\mathrm{pre}}^{(n)} - B_{\mathrm{gt}}^{(n)} \right\|_2^2 \tag{14}$$

where $\mathrm{L}_{\mathrm{reg}}^B$ denotes the bounding-box displacement regression loss, $B_{\mathrm{pre}}^{(n)}$ and $B_{\mathrm{gt}}^{(n)}$ are the predicted and ground-truth future bounding-box displacement for the n-th sample, respectively, and N denotes the batch size. Here, $\|\cdot\|_2^2$ represents the squared ℓ_2 norm.

Subsequently, B_{pre} is incorporated into a FiLM-based modulation mechanism to modulate the aggregated temporal representation $\bar{Z}_L$ from the STA module. Specifically, the modulation parameters $\gamma^s \in \mathbb{R}^{256}$ and $\beta^s \in \mathbb{R}^{256}$ are computed as:

$$\gamma^s = W_\gamma^s B_{\mathrm{pre}} + b_\gamma^s \tag{15}$$

$$\beta^s = W_\beta^s B_{\mathrm{pre}} + b_\beta^s \tag{16}$$

where W_γ^s and W_β^s are learnable projection matrices, and b_γ^s and b_β^s are the corresponding bias terms.

The modulated summary token $\hat{Z}_L \in \mathbb{R}^{256}$ is obtained by:

$$\hat{Z}_L = \gamma^s \odot \bar{Z}_L + \beta^s \tag{17}$$

3.8. *Language-guided distillation*

The language-guided distillation (LGD) module transfers high-

Just Accepted

level semantic knowledge from pedestrian- and scene-level descriptions to visual representations during training. Specifically, the text encoder maps the corresponding descriptions into compact text representations, with which $\bar{Z}_L$ and $\bar{Z}_G$ are aligned to incorporate pedestrian- and scene-level semantics, respectively. For each $i \in \{L, G\}$, the aggregated temporal representation $\bar{Z}_i$ is projected through a linear layer to obtain a 512-dimensional embedding $\bar{Z}_i' \in \mathbb{R}^{512}$, matching the dimensionality of the corresponding text representation T_i. Cosine similarity is then employed to measure the semantic consistency between $\bar{Z}_i'$ and T_i in the shared embedding space:

$$\cos\left(\bar{Z}_i', T_i\right) = \frac{\bar{Z}_i' \cdot T_i}{\left\|\bar{Z}_i'\right\| \left\|T_i\right\|} \tag{18}$$

Accordingly, the language-guided distillation loss is defined as:

$$\mathrm{L}_{\mathrm{LGD}}^{i} = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \cos\left(\bar{Z}_i'^{(n)}, T_i^{(n)}\right)\right) \tag{19}$$

where $\bar{Z}_i'^{(n)}$ is the projected visual representation of the n-th sample, $T_i^{(n)}$ is the corresponding text representation. Minimizing $\mathrm{L}_{\mathrm{LGD}}^{i}$ encourages the visual representations to capture the high-level semantic information encoded in the corresponding text representations, thereby improving their semantic expressiveness.

The LGD module is applied only during training, while the text encoder and semantic descriptions are not required at inference, allowing VLM-GSDN to benefit from language-guided semantic supervision without introducing additional inference overhead.

3.9. Auxiliary supervision

The auxiliary supervision (AS) module introduces pedestrian action-state transition prediction as an auxiliary task, providing additional motion supervision to improve the primary intention prediction task. The transition labels are constructed from the pedestrian action annotations in JAAD and PIE. Let $a^t \in$ {standing, walking} denote the action state at the final frame of the observation window, and $a^{t+H_{AS}}$ the action state 1 second later, where H_{AS}=30 frames. The transition label is defined as follows:

$$y_{aux} = \begin{cases} 0, a^{t+H_{AS}} = a^t, \\ 1, a^t = \text{standing}, a^{t+H_{AS}} = \text{walking}, \\ 2, a^t = \text{walking}, a^{t+H_{AS}} = \text{standing}. \end{cases} \tag{20}$$

The no-transition category includes standing-to-standing and walking-to-walking cases, whereas the other two categories represent standing-to-walking and walking-to-standing transitions, respectively. Specifically, the modulated summary token $\hat{Z}_L$ generated by the PSM module is first processed by an MLP and then fed into a three-class classification head, where the auxiliary task is optimized using cross-entropy loss:

$$\mathrm{L}_{\mathrm{aux}}^{T} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=0}^{2} y_{\mathrm{aux},c}^{(n)} \log(\hat{y}_{\mathrm{aux},c}^{(n)}) \tag{21}$$

where $y_{\mathrm{aux},c}^{(n)}$ and $\hat{y}_{\mathrm{aux},c}^{(n)}$ denote the ground-truth class indicator and predicted probability for class c of the n-th sample, respectively.

3.10. Center-constrained classification

The summary tokens $\bar{Z}_G$ from STA module and $\hat{Z}_L$ from PSM module are first concatenated and then fed into an MLP with hidden dimensions of 256 and 128, followed by a binary classification head to produce the final intention prediction. To mitigate class imbalance and emphasize hard samples, the classification head is optimized using focal loss, defined as

$$\mathrm{L}_{\mathrm{cls}}^{I} = -\frac{1}{N}\sum_{n=1}^{N}\alpha_{y^{(n)}}\left(1 - p_{y^{(n)}}^{(n)}\right)^{\gamma}\log\left(p_{y^{(n)}}^{(n)}\right) \tag{22}$$

where $\alpha_{y^{(n)}}$ is the class weighting factor, γ is the focusing parameter, and $p_{y^{(n)}}^{(n)}$ is the predicted probability of the ground-truth class for the n-th sample.

Although focal loss improves classification by emphasizing hard samples, it does not explicitly enforce intra-class compactness. Consequently, representations belonging to the same intention class may remain dispersed in the feature space. To address this limitation, we introduce a class-weighted center loss on the hidden representation produced by the first MLP layer. This loss encourages samples to cluster around their corresponding class centers, thereby enhancing feature discriminability. Formally, the class-weighted center loss is defined as:

$$\mathrm{L}_{\mathrm{cen}} = \frac{1}{N}\sum_{n=1}^{N}\omega_{y^{(n)}}\left\|Z_{LG}^{(n)} - c_{y^{(n)}}\right\|_2^2 \tag{23}$$

where $Z_{LG}^{(n)}$ denotes the hidden representation of the i-th sample produced by the first MLP layer, $c_{y^{(n)}}$ is the learnable center corresponding to class $y^{(n)}$ and $\omega_{y^{(n)}}$ is the class-specific weighting factor.

Accordingly, the overall loss function for pedestrian intention prediction is defined as:

$$\mathrm{L}_{\mathrm{all}} = \mathrm{L}_{\mathrm{cls}}^{I} + \alpha \cdot \left(\mathrm{L}_{\mathrm{LGD}}^{L} + \mathrm{L}_{\mathrm{LGD}}^{G}\right) + \beta \cdot \mathrm{L}_{\mathrm{reg}}^{B} + \gamma \cdot \mathrm{L}_{\mathrm{aux}}^{T} + \lambda \cdot \mathrm{L}_{\mathrm{cen}} \tag{24}$$

where α, β, γ, and λ are hyper-parameters that balance the contributions of the language-guided distillation, bounding-box displacement regression, transition prediction, and center losses defined in Eqs. (19), (14), (21), and (23), respectively.

4 Experiment

4.1. Datasets and implementation details

We evaluate VLM-GSDN on two public benchmarks, JAAD (Rasouli et al., 2017) and PIE (Rasouli et al., 2019). JAAD comprises 346 high-resolution videos captured in Eastern Europe and North America. We use two JAAD subsets, namely $\mathrm{JAAD}_{\mathrm{beh}}$ and $\mathrm{JAAD}_{\mathrm{all}}$. $\mathrm{JAAD}_{\mathrm{all}}$ subset is augmented with an additional 2,100 samples of pedestrians that are not crossing. Following Benchmark (Kotseruba et al., 2021), we split the JAAD dataset into 177 videos for training, 29 for validation, and 117 for testing. PIE dataset consists of high-resolution video captured in Toronto, with frame-level annotations for 1,842 pedestrians. We adopt the standard split

Just Accepted

Table 1. Comparison results with state-of-the-art methods on the JAAD and PIE datasets.

Models	Inference input	$JAAD_{beh}$					$JAAD_{all}$					PIE				
		Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
Global PCPA (Yang et al., 2022)	I, B, Po, V, S	0.62	0.54	0.74	0.65	0.85	0.83	0.82	0.63	0.51	0.81	0.89	0.86	0.80	0.79	0.81
PIT (Zhou et al., 2023)	I, Po, V	0.70	0.65	0.81	0.71	0.93	0.87	0.87	0.66	0.54	0.85	0.91	0.90	0.82	0.85	0.79
DPCIAN (Yang et al., 2023)	Po, I	0.71	0.66	0.78	0.73	0.85	0.89	0.77	0.59	0.61	0.58	0.91	0.88	0.83	0.83	0.84
STFF-MANet (Zhang et al., 2024)	O, Po, B, V, I, S	0.66	0.58	0.77	0.67	0.89	0.89	0.80	0.67	0.68	0.67	0.89	0.88	0.82	0.79	0.85
PedAST-GCN (Ling et al., 2024)	Po	0.69	0.66	0.79	0.68	0.93	0.89	0.83	0.68	0.67	0.69	0.91	0.94	0.83	0.88	0.79
PVIM (Xu et al., 2024)	I, V, Po, B	-	-	-	-	-	-	-	-	-	-	0.91	0.90	0.84	0.82	0.87
GANet (Zeng et al., 2024)	B, PM, S, I, V	0.74	0.54	0.84	0.78	0.91	0.84	0.83	0.67	0.57	0.81	0.74	0.54	0.84	0.78	0.91
RAIDN (Yang et al., 2024b)	Po, I	0.70	0.68	0.75	0.75	0.77	0.89	0.80	0.66	0.65	0.72	0.92	0.89	0.85	0.82	0.89
MTMGN (Yang et al., 2024a)	S, Po, D	0.70	0.70	0.83	0.79	0.87	0.89	0.89	0.73	0.66	0.89	0.90	0.87	0.92	0.95	0.90
TCL (Liang et al., 2024)	I, B, Po	0.74	0.61	0.85	0.75	0.98	0.87	0.92	0.71	0.68	0.74	0.87	0.88	0.92	0.89	0.96
ACIT (Li and Müller, 2025a)	I, O, S, V, B	0.70	0.62	0.79	0.69	0.93	0.89	0.83	0.69	0.65	0.74	-	-	-	-	-
Gated-S2R-PCP (Bai et al., 2025)	B, I, DM, S	0.66	0.68	0.77	0.71	0.85	0.88	0.84	0.79	0.77	0.82	0.92	0.90	0.82	0.83	0.80
LSOP-Net (Liu et al., 2025)	V, B, Po, I, S	0.65	0.54	0.78	0.65	0.98	0.85	0.75	0.58	0.56	0.61	0.89	0.87	0.81	0.80	0.82
GTransPDM (Xie et al., 2025)	B, V, Po	-	-	-	-	-	0.87	0.78	0.64	0.64	-	0.90	0.87	0.82	0.86	-
MTC (Li and Müller, 2025b)	V, B, I, S, DM	0.71	0.65	0.80	0.72	0.90	0.90	0.82	0.70	0.70	0.70	0.89	0.85	0.80	0.85	0.75
PIP-Net (Azarmi et al., 2025a)	B, Po, V, I, O, DM	-	-	-	-	-	-	-	-	-	-	0.91	0.90	0.84	0.85	0.84
LIM (Chen et al., 2025)	Po	0.72	0.75	0.77	0.85	0.67	0.89	0.68	0.91	0.69	0.67	0.91	0.84	0.96	0.84	0.85
PTINet (Munir and Kucner, 2025)	B, I, O, LC	-	-	-	-	-	0.96	-	0.92	-	-	0.98	-	0.96	-	-
SAM (Elgazwy et al., 2025)	V, B, Po, I, S	0.67	0.60	0.77	0.68	0.90	0.83	0.80	0.62	0.47	0.82	-	-	-	-	-
SafeCrossNet (Du et al., 2026)	Po, B, D, I, S	-	-	-	-	-	0.90	0.85	0.73	0.70	0.79	0.91	0.88	0.82	0.76	0.79
GATEPose (AlShami et al., 2026)	Po	-	-	-	-	-	0.85	0.93	0.87	0.86	0.89	0.87	0.90	0.86	0.83	0.88
GPT-4V-PBP (Huang et al., 2024)	I, P	0.57	0.61	0.65	0.82	0.54	-	-	-	-	-	-	-	-	-	-
LLaMAPed (Ham et al., 2024)	I, B, V, P	0.58	0.59	0.59	0.67	0.52	-	-	-	-	-	-	-	-	-	-
GPT-4V (Azarmi et al., 2025b)	I, V, PC, P	0.74	0.67	0.66	0.68	0.65	-	-	-	-	-	0.81	0.74	0.73	0.75	0.72
ClipCross (Uziel and Bialer, 2025)	I, B	-	-	-	-	-	0.89	0.93	-	-	-	0.84	0.85	-	-	-
PedVLM (Munir et al., 2025)	I, B, O, V, LC, P	-	-	-	-	-	-	0.81	0.94	0.89	0.99	-	0.52	0.74	0.65	0.85
TCP (Xiao et al., 2025)	B	-	-	-	-	-	0.89	0.91	0.68	0.70	0.66	0.91	0.94	0.84	0.84	0.83
PedFeatRulesKG (Hussien et al., 2025)	PM, D, LC, PC	-	-	-	-	-	0.83	-	0.87	0.86	0.88	-	-	-	-	-
VLMPed-CoT (Ling et al., 2026)	I, B, V, P	0.82	-	0.84	0.90	0.80	0.90	-	0.74	0.69	0.80	0.90	-	0.82	0.83	0.81
OmniPredict (Ham et al., 2026)	I, B, V, P	0.67	0.65	0.65	0.66	0.65	-	-	-	-	-	-	-	-	-	-
GLCCL (Lin and Hu, 2026)	B, Po, V	-	-	-	-	-	0.91	0.89	0.75	0.72	0.78	0.93	0.93	0.88	0.91	0.85
PedestrianQA (Mishra et al., 2026)	I, B, V, LC, PC, P	0.55	-	0.71	-	-	-	-	-	-	-	0.67	-	0.80	-	-
VLM-GSDN	I, PM	0.76	0.85	0.83	0.73	0.97	0.91	0.93	0.74	0.75	0.73	0.92	0.95	0.85	0.88	0.83

Notes: The abbreviations used for input data are image (I); 2D bounding boxes (B); ego-vehicle velocity (V); pedestrian motion (PM); pose (Po); semantic maps (S); optical flow (O); distance (D); depth maps (DM); local contextual features (LC); pedestrian physical cues (PC); prompt (P). All comparative results are taken directly from the corresponding original publications. A dash (–) indicates that the corresponding result was not reported.

(Yang et al., 2022): sets 01, 02, and 04 for training, sets 05 and 06 for validation, and set 03 for testing. From the annotated sequences, we extract clips with a sliding window, using 80% overlap on JAAD dataset and 60% on PIE dataset.

The batch size is set to 16, and the model is trained for 100 epochs on an NVIDIA GeForce RTX 3080 Ti GPU. The motion encoder is implemented using a 2-layer Transformer with 4 attention heads and a hidden dimension of 128. The STA module consists of 2 layers with a feature dimension of 256, where the MHCA uses 8 attention heads. The hyper-parameters α, β, γ, and λ are set to 0.5, 1.5, 0.3, and 0.1, respectively.

4.2. Comparative results

To evaluate the performance of VLM-GSDN, we compare it with state-of-the-art (SOTA) methods. We also include representative VLM-based approaches that directly employ VLMs for intention prediction, as well as language-enhanced approaches that incorporate VLM-derived semantics as additional inputs or auxiliary guidance. For evaluation, we use five standard metrics for intention prediction: accuracy (Acc), precision (P), recall (R), F1 score, and area under the ROC curve (AUC). The comparative results are summarized in Table 1.

On the $JAAD_{beh}$ dataset, VLM-GSDN achieves the second-best accuracy of 76%, lower than VLMPed-CoT (Ling et al., 2026). However, VLM-GSDN achieves a recall that is 17 percentage points (pp) higher than that of VLMPed-CoT, indicating a stronger ability to identify crossing cases. VLM-GSDN outperforms the third-best methods, TCL (Liang et al., 2024), GANet (Zeng et al., 2024) and GPT-4V (Azarmi et al., 2025b), by 2 pp in accuracy while using the fewest input modalities. Specifically, TCL and GPT-4V use three and four input modalities, respectively, whereas GANet uses five and additionally relies on a pretrained model to generate semantic maps. VLM-GSDN achieves the best AUC, surpassing LIM (Chen et al., 2025), by 10 pp. In terms of recall, VLM-GSDN achieves 97%, only 1 pp behind LSOP-Net (Liu et al., 2025) and TCL.

On the $JAAD_{all}$ dataset, VLM-GSDN achieves an accuracy of 91%, trailing only PTINet (Munir and Kucner, 2025). Nevertheless, VLM-GSDN requires fewer input modalities and avoids reliance on an auxiliary optical-flow estimator and manually annotated high-level local contextual features. It outperforms MTC (Li and Müller, 2025b) and SafeCrossNet (Du et al., 2026) by 1 pp in accuracy while using three and four fewer input modalities, respectively. Unlike these methods, VLM-GSDN also does not require additional pretrained models for depth or semantic map estimation. Moreover, it achieves the highest AUC of 93%.

On the PIE dataset, VLM-GSDN attains 92% accuracy, trailing PTINet (Munir and Kucner, 2025) and falling 1 pp below GLCCL

Just Accepted

Table 2. Comparison of future bounding-box prediction performance.

Methods	$JAAD_{beh}$		$JAAD_{all}$		PIE	
	Disp. MAE↓(px)	CFDE ↓(px)	Disp. MAE↓(px)	CFDE ↓(px)	Disp. MAE↓(px)	CFDE ↓(px)
PIE-Traj (Rasouli et al., 2019)	31.58	50.81	21.95	35.63	13.32	21.36
BiTraP (Yao et al., 2021)	27.82	42.28	27.65	45.99	15.74	24.19
PedCMT (Chen et al., 2024)	114.35	177.81	58.45	92.47	63.39	76.04
PTINet (Munir and Kucner, 2025)	34.83	54.71	22.28	35.65	15.88	25.45
TCP (Xiao et al., 2025)	52.91	71.25	34.67	50.94	39.94	60.23
VLM-GSDN	31.24	48.54	21.52	34.56	12.24	19.75

Notes: PIE-Traj and BiTraP are dedicated pedestrian bounding-box prediction methods, whereas PedCMT, PTINet, and TCP incorporate future bounding-box prediction as an auxiliary task for intention prediction. All compared methods are reproduced using their publicly released implementations. For methods that predict a sequence of future bounding boxes, the reported metrics are computed using only the predicted bounding box at the 1-s horizon. For PedCMT and TCP, the prediction horizon is adjusted to 1 s for a fair comparison.

Table 3. Contribution and interaction analysis of core modules. All metrics are reported in %.

LGD	STA	MGM	PSM	AS	CCC	$JAAD_{beh}$					$JAAD_{all}$					PIE				
						Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
√	√	√	√	√	×	73.05	75.75	81.10	72.24	92.44	89.91	90.62	70.62	71.96	69.33	90.28	94.32	82.75	82.64	82.87
√	√	√	√	×	√	74.48	86.94	82.76	71.69	97.88	90.08	92.45	70.96	72.66	69.33	91.38	94.10	84.18	87.06	81.47
√	√	√	×	√	√	75.23	83.84	82.64	73.59	94.22	90.33	91.31	73.00	71.31	74.77	89.68	93.50	81.78	81.25	82.31
√	√	×	√	√	√	73.79	81.88	81.81	72.29	94.22	90.02	90.74	69.12	75.28	63.89	84.80	88.51	70.20	78.33	63.59
√	√	×	×	√	√	72.36	81.32	81.63	69.87	98.13	89.77	91.43	69.82	72.06	67.71	83.81	92.36	74.46	66.94	83.89
√	×	√	√	√	√	74.22	77.81	82.42	71.87	96.60	90.02	92.20	71.23	71.79	70.69	90.09	94.21	82.52	81.99	83.05
×	√	√	√	√	√	72.46	81.27	81.25	70.79	95.33	89.23	91.30	66.48	72.92	61.09	86.71	93.72	78.31	72.45	85.20
√	√	√	√	√	√	75.97	85.35	83.43	73.37	96.69	91.06	92.61	74.01	75.24	72.81	91.95	94.57	85.23	88.16	82.50

Table 4. Performance comparison of different text encoders. All metrics are reported in %.

Text encoder	$JAAD_{beh}$					$JAAD_{all}$					PIE				
	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
BERT	73.26	81.29	81.38	72.11	93.37	88.35	91.24	68.54	64.94	72.56	88.99	94.00	80.61	79.95	81.28
CLIP	74.22	77.56	81.76	73.35	92.35	89.99	91.96	72.15	70.23	74.17	91.19	94.30	84.53	83.61	85.47
MPNet	71.61	82.28	81.08	69.54	97.20	89.74	91.53	71.65	69.29	74.17	90.23	94.23	82.87	81.78	83.99
Long-CLIP	75.97	85.35	83.43	73.37	96.69	91.06	92.61	74.01	75.24	72.81	91.95	94.57	85.23	88.16	82.50

(Lin and Hu, 2026). However, unlike GLCCL, VLM-GSDN does not require pedestrian pose estimates. It also matches the accuracy of Gated-S2R-PCP (Bai et al., 2025) and RAIDN (Yang et al., 2024b). Notably, VLM-GSDN is trained solely on PIE, without leveraging additional large-scale synthetic data as in Gated-S2R-PCP or relying primarily on pose cues as in RAIDN. VLM-GSDN achieves the best AUC, outperforming PedAST-GCN (Ling et al., 2024) and TCP (Xiao et al., 2025) by 1 pp.

To further assess whether PSM module captures meaningful future-oriented spatial information, we quantitatively evaluate its future bounding-box displacement prediction capability against five existing methods under a unified evaluation setting. All metrics are evaluated at the 1-s prediction horizon. We report the mean absolute error of bounding-box displacement (Disp. MAE) and center final displacement error (CFDE), both measured in pixels (px), where lower values indicate more accurate future spatial prediction. As shown in Table 2, VLM-GSDN achieves the best performance on $JAAD_{all}$ and PIE for both metrics. Specifically, compared with PIE-Traj (Rasouli et al., 2019), VLM-GSDN reduces Disp. MAE and CFDE by 0.43 and 1.07 px on $JAAD_{all}$, and by 1.08 and 1.61 px on PIE, respectively. On $JAAD_{beh}$, VLM-GSDN ranks second, trailing BiTraP (Yao et al., 2021) by 3.42 px in Disp. MAE and 6.26 px in CFDE. These results demonstrate the effectiveness of PSM in capturing meaningful future spatial information.

The strong comparative performance of VLM-GSDN while relying on a compact set of input modalities can be attributed to three main factors. First, VLM-GSDN leverages the semantic understanding capabilities of VLMs to generate high-level semantic descriptions that capture rich scene- and pedestrian-level information, which can be difficult to extract directly from raw input modalities. The LGD module distills this high-level semantic knowledge into visual representations during training, thereby enhancing their semantic richness while eliminating the need for online semantic description generation at inference. Second, VLM-GSDN enhances motion modeling through the MGM and PSM modules. By incorporating observed pedestrian motion dynamics and predicted future spatial information into visual representations, VLM-GSDN learns more discriminative motion-aware features. Third, pedestrian action-state transition prediction provides additional motion supervision, while the class-weighted center loss improves feature discriminability.

4.3. Ablation studies

(1) Contribution of core modules: We evaluate the contribution of each module by individually removing language-guided distillation (LGD), selective temporal attention (STA), motion-guided modulation (MGM), predictive spatial modulation (PSM), auxiliary supervision (AS), and center-constrained classification

Just Accepted

Table 5. Performance comparison of different vision encoders. All metrics are reported in %.

Vision encoder	$JAAD_{beh}$					$JAAD_{all}$					PIE				
	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
VGG16	69.96	71.66	80.55	67.71	99.41	88.81	85.49	66.34	70.00	63.04	88.34	92.24	77.28	85.54	70.48
EfficientNet	72.36	81.89	81.43	70.24	96.86	88.81	90.83	67.39	68.73	66.10	88.89	92.94	79.05	84.21	74.49
ViT	73.31	79.62	82.07	70.79	97.62	89.41	88.95	68.49	71.36	65.85	88.86	93.46	79.81	81.47	78.21
DeiT	73.74	79.87	81.88	72.05	94.82	90.17	91.67	71.59	72.33	70.86	89.54	94.11	80.94	83.12	78.86
Swin-V2	74.11	81.43	81.94	72.70	93.88	90.17	87.86	67.74	79.43	59.05	90.20	93.47	81.49	87.00	76.63
DINOv2-Reg	75.97	85.35	83.43	73.37	96.69	91.06	92.61	74.01	75.24	72.81	91.95	94.57	85.23	88.16	82.50

Table 6. Performance comparison of different loss functions for center-constrained classification. All metrics are reported in %.

Loss	$JAAD_{beh}$					$JAAD_{all}$					PIE				
	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
$L_{cls}^{I} + L_{tri}$	73.21	81.62	81.86	71.02	96.60	89.11	90.19	68.85	68.88	68.82	90.46	94.27	83.92	79.83	88.45
$L_{cls}^{I} + L_{con}$	74.85	84.92	83.11	71.67	98.90	89.85	86.85	71.10	70.83	71.37	91.46	92.64	84.71	85.35	84.08
$L_{cls}^{I} + L_{var}$	74.85	81.48	82.40	73.31	94.05	89.60	89.90	70.71	69.66	71.79	90.85	88.76	82.52	89.27	76.72
$L_{cls}^{I} + L_{cen}$	75.97	85.35	83.43	73.37	96.69	91.06	92.61	74.01	75.24	72.81	91.95	94.57	85.23	88.16	82.50

(CCC). To further examine the interaction between MGM and PSM, we additionally remove both modules simultaneously. The results are reported in Table 3. When CCC is removed, accuracy consistently decreases across all three datasets, demonstrating its contribution to discriminative feature learning. Removing AS also degrades accuracy, indicating that the auxiliary pedestrian action-state transition prediction task provides effective supervision for learning more informative motion-aware representations. The joint use of MGM and PSM achieves better overall performance than using either module alone. More importantly, simultaneously removing both modules causes a more pronounced performance degradation, reducing accuracy by 3.61, 1.29, and 8.14 pp on $JAAD_{beh}$, $JAAD_{all}$, and PIE, respectively. This indicates that MGM and PSM provide complementary motion cues, with MGM capturing observed motion dynamics from frame-level pedestrian motion and PSM introducing future-oriented spatial cues from predicted bounding-box displacement. Removing STA also degrades performance, as treating all frames equally prevents the model from prioritizing the most informative temporal observations. Finally, when LGD is removed, substantial performance degradation is observed, with accuracy decreases of 3.51, 1.83, and 5.24 pp on $JAAD_{beh}$, $JAAD_{all}$, and PIE, respectively. This indicates that LGD provides effective semantic supervision by aligning visual representations with high-level semantic information.

(2) Comparison of text encoders: We further compare different text encoders to evaluate their effectiveness in encoding the generated semantic descriptions. The results are shown in Table 4. Long-CLIP (Zhang et al., 2024) achieves the best accuracy, while CLIP (Radford et al., 2021) consistently ranks second. CLIP's inferior performance is likely due to its limited input length, which may truncate textual descriptions and result in incomplete semantic information. BERT (Devlin et al., 2019) and MPNet (Song et al., 2020) perform worse because they are designed for unimodal language modeling and lack explicit alignment with visual representations. In contrast, CLIP and Long-CLIP are trained with cross-modal objectives to align textual and visual representations, making them better suited for cross-modal semantic alignment.

(3) Comparison of vision encoders: We also compare different vision encoders to assess their effectiveness. The results are shown in Table 5. Compared with CNN-based encoders such as VGG16 (Simonyan and Zisserman, 2014) and EfficientNet (Tan and Le, 2019), Transformer-based encoders, including DINOv2-Reg (Darcet et al., 2023), ViT (Dosovitskiy et al., 2020), DeiT (Touvron et al., 2021), and Swin-V2 (Liu et al., 2022), achieve better performance, highlighting their stronger capability to extract expressive visual representations. Moreover, DINOv2-Reg achieves the best overall performance among the evaluated vision encoders across all three datasets, supporting its selection as the vision encoder in VLM-GSDN.

(4) Ablation on center-constrained classification: To further evaluate CCC, we replace class-weighted center loss L_{cen} in CCC with triplet loss L_{tri}, contrastive loss L_{con}, or variance loss L_{var}, while retaining the focal loss L_{cls}^{I}. The results are shown in Table 6. All alternative loss formulations underperform class-weighted center loss. This may be because triplet and contrastive losses rely on explicit triplet or pair construction, which can make optimization more sensitive to class imbalance. Moreover, variance loss does not explicitly constrain samples toward class centers, limiting its ability to enforce intra-class compactness. In contrast, the class-weighted center loss explicitly promotes intra-class compactness without requiring triplet or pair construction, while accounting for class imbalance, thereby producing more stable and discriminative representations.

(5) Effect of auxiliary supervision under long-horizon prediction: To evaluate the contribution of AS to long-horizon prediction, we

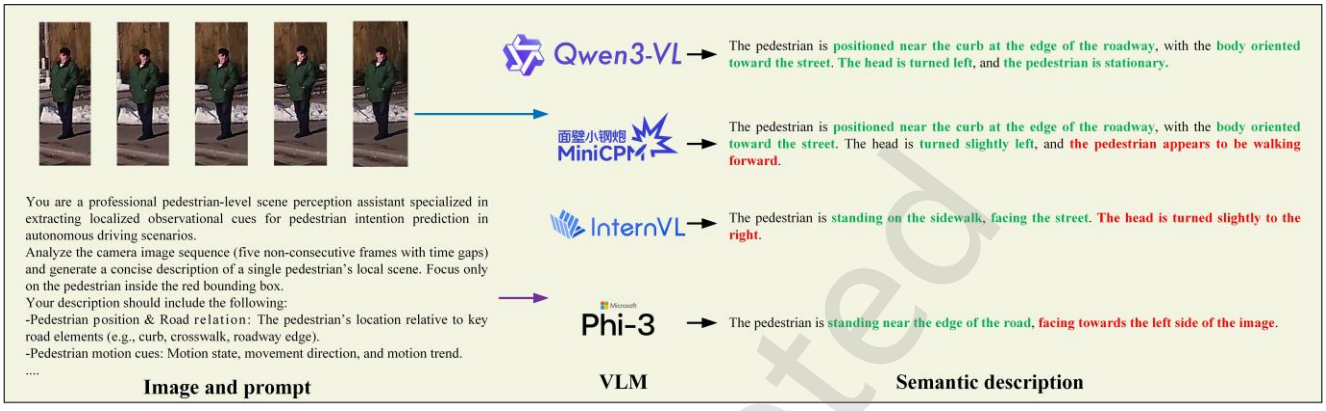
Table 7. Effect of auxiliary supervision (AS) under long-horizon prediction.

Dataset	TTE (s)	Acc (%)			F1 (%)		
		w/o AS	w/ AS	Δ (pp)	w/o AS	w/ AS	Δ (pp)
$JAAD_{beh}$	2-3	68.82	69.97	+1.15	79.41	80.87	+1.46
	3-4	67.95	68.60	+0.65	77.38	77.47	+0.09
$JAAD_{all}$	2-3	85.97	86.18	+0.21	62.49	65.35	+2.86
	3-4	85.99	86.09	+0.10	59.76	62.69	+2.93
PIE	2-3	83.19	83.65	+0.46	65.38	71.10	+5.72
	3-4	78.44	80.94	+2.50	64.47	66.51	+2.04

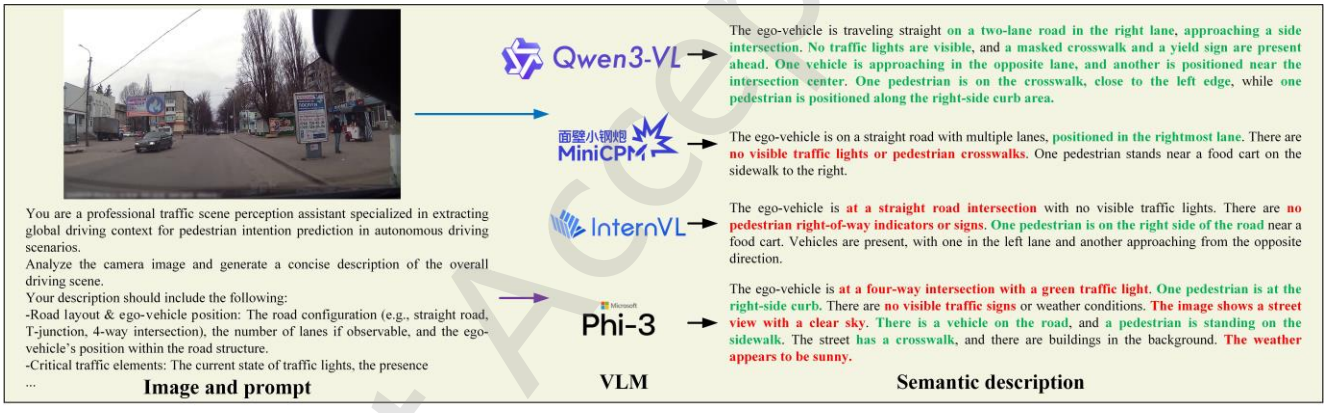
Just Accepted

Table 8. Sensitivity analysis of VLM-GSDN to different semantic description generators. All metrics are reported in %.

Sem. generator	$JAAD_{beh}$			$JAAD_{all}$			PIE		
	Acc	AUC	F1	Acc	AUC	F1	Acc	AUC	F1
None	72.47 ± 0.21	77.86 ± 2.26	81.16 ± 0.17	89.00 ± 0.57	91.67 ± 1.91	68.21 ± 3.22	86.87 ± 0.22	92.50 ± 0.70	76.53 ± 1.54
Phi-3-Vision	75.56 ± 0.16	83.53 ± 0.43	83.28 ± 0.12	90.50 ± 0.11	92.80 ± 1.34	71.89 ± 1.58	90.95 ± 0.43	93.76 ± 0.73	84.14 ± 0.91
MiniCPM-V4.6	75.04 ± 0.56	81.68 ± 0.48	82.77 ± 0.39	90.34 ± 0.34	92.49 ± 1.27	72.82 ± 0.92	90.47 ± 0.35	94.48 ± 0.96	83.65 ± 0.73
InternVL3	75.71 ± 0.35	82.74 ± 0.41	83.03 ± 0.17	90.70 ± 0.28	91.85 ± 0.98	72.95 ± 1.04	91.19 ± 0.34	93.59 ± 0.84	84.28 ± 0.79
Qwen3-VL	75.94 ± 0.43	85.14 ± 0.59	83.41 ± 0.24	90.96 ± 0.23	92.93 ± 0.36	73.87 ± 0.45	91.91 ± 0.20	93.80 ± 0.59	84.76 ± 0.63



(a)



(b)

Fig. 7. Comparison of semantic descriptions generated by different VLMs: (a) pedestrian-level semantic descriptions; (b) scene-level semantic descriptions.

compare VLM-GSDN with and without AS under extended time-to-event (TTE) settings: 2-3 s and 3-4 s. As shown in Table 7, introducing AS consistently improves both accuracy and F1 score across all datasets and TTE settings. The improvements are particularly evident in F1 score on $JAAD_{all}$, with gains of 2.86 and 2.93 pp, and on PIE, with gains of 5.72 and 2.04 pp under the 2-3 s and 3-4 s settings, respectively. These results suggest that action-state transition supervision provides complementary motion cues for intention prediction when the prediction horizon is extended.

4.4. Analysis of language-guided distillation

4.4.1. Sensitivity to semantic description generators

To assess whether the performance gains of VLM-GSDN depend on the choice of semantic description generator, we compare Qwen3-VL (Bai et al., 2025) with Phi-3-Vision (Abdin et al., 2024), MiniCPM-V 4.6 (Yu et al., 2026), and InternVL3 (Zhu et al., 2025). The network architecture and training settings are kept unchanged. We also include a no-semantic baseline in which neither semantic descriptions nor LGD are used. For each setting, we report the mean and standard deviation obtained from five independent runs with different random seeds. The results are shown in Table 8. Overall, using semantic descriptions generated by different VLMs leads to notable performance improvements over the no-semantic baseline. Among the evaluated generators, Qwen3-VL achieves the highest mean accuracy on all three datasets, reaching 75.94 ± 0.43%, 90.96 ± 0.23%, and 91.91 ± 0.20% on $JAAD_{beh}$, $JAAD_{all}$, and PIE, respectively. Compared with InternVL3, the corresponding mean accuracy improvements are 0.23, 0.26, and 0.72 pp. These results indicate that language-guided distillation remains effective across different semantic description generators, while the observed performance variations demonstrate that the choice of generator can influence prediction performance.

To provide qualitative insights into these performance differences, Fig. 7 presents challenging pedestrian- and scene-level cases together with semantic descriptions generated by different VLMs. Correct and incorrect content in the generated descriptions

Just Accepted

Table 9. Quantitative reliability evaluation of VLM-generated semantic descriptions.

Dataset	Semantic level	# Claims	Correct claims ↑ (%)	Incorrect claims ↓ (%)	Hallucinated claims ↓ (%)
JAAD	Pedestrian-level	402	79.85	18.16	1.99
JAAD	Scene-level	545	78.72	17.06	4.22
PIE	Pedestrian-level	412	74.27	24.03	1.70
PIE	Scene-level	679	76.88	17.38	5.74



Fig. 8. Representative incorrect and hallucinated claims in VLM-generated semantic descriptions: (a) incorrect pedestrian-level claim; (b) hallucinated pedestrian-level claim; (c) incorrect scene-level claim; (d) hallucinated scene-level claim.

is highlighted in green and red, respectively. In the pedestrian-level case shown in Fig. 7(a), all VLMs correctly identify the pedestrian's position, indicating comparable capability in coarse spatial perception. However, clear differences emerge in the recognition of fine-grained pedestrian states. Qwen3-VL and MiniCPM-V 4.6 correctly capture the pedestrian's head orientation, whereas only Qwen3-VL accurately identifies the stationary motion state in this case. The scene-level case in Fig. 7(b) presents a challenging traffic scenario in which the ego-vehicle is relatively far from the intersection and the crosswalk markings are partially degraded. Qwen3-VL and MiniCPM-V 4.6 correctly identify the vehicle's position, while Qwen3-VL and Phi-3-Vision recognize the presence of the crosswalk. Qwen3-VL additionally identifies the small yield sign and the distant pedestrian at the intersection, whereas Phi-3-Vision introduces irrelevant weather-related information. These examples illustrate qualitative differences among the evaluated VLMs in capturing fine-grained pedestrian states and driving-scene context, providing further insight into the observed variations in prediction performance.

4.4.2. Reliability and robustness of semantic descriptions

To assess the reliability of VLM-generated semantic descriptions, we randomly select 150 samples for each dataset and semantic level. The descriptions are decomposed into independently verifiable claims, which are assessed against the corresponding visual evidence and available dataset annotations. Evaluable claims are categorized as correct, incorrect, or hallucinated, corresponding to accurate descriptions, incorrect descriptions of object attributes, states, or relations, and assertions of nonexistent objects or events, respectively. As shown in Table 9, correct claims account for 79.85% and 74.27% of the evaluated claims in pedestrian-level descriptions on JAAD and PIE, respectively, with corresponding proportions of 78.72% and 76.88% for scene-level descriptions. The highest proportion of incorrect claims is observed in pedestrian-level descriptions on PIE (24.03%), whereas the highest hallucination rate occurs in scene-level descriptions on PIE (5.74%). Overall, these results indicate that the generated descriptions are generally reliable across pedestrian- and scene-level semantics, while still exhibiting residual errors and hallucinations.

Fig. 8 provides representative examples of incorrect and hallucinated claims. Correct, incorrect, and hallucinated content is highlighted in green, red, and orange, respectively. In Fig. 8(a), the descriptions correctly capture the pedestrians' motion states but misidentify Pedestrian A's gaze target as the ground rather than the phone and describe Pedestrian B as facing straight ahead rather than looking to the right, toward the ego-vehicle. In Fig. 8(b), the description of Pedestrian C misinterprets the observed hand-raising gesture as signaling to another person, whereas that of Pedestrian D infers an unverified mental state from the observed walking behavior. These examples illustrate two representative pedestrian-level semantic errors: inaccurate descriptions of observable attributes and unsupported behavioral interpretations. In Fig. 8(c), the description fails to identify the crosswalk markings,

Just Accepted

Table 10. Robustness analysis under different semantic description shuffling ratios.

Shuffling ratio (%)	$JAAD_{beh}$		$JAAD_{all}$		PIE	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
w/o sem. dist	72.46	81.25	89.23	66.48	86.71	78.31
10	75.27	83.23	90.72	72.41	91.37	84.09
20	74.48	82.53	90.40	73.08	90.32	83.19
40	72.35	81.44	89.72	70.97	88.58	80.96
0	75.97	83.43	91.06	74.01	91.95	85.23

Table 11. Complementarity analysis of pedestrian- and scene-level semantic distillation.

Pedestrian-level	Scene-level	$JAAD_{beh}$		$JAAD_{all}$		PIE	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
×	×	72.46	81.25	89.23	66.48	86.71	78.31
√	×	74.75	82.49	90.52	71.24	90.39	82.42
×	√	73.26	82.03	90.67	73.05	90.78	83.32
√	√	75.97	83.43	91.06	74.01	91.95	85.23

Table 12. Performance and efficiency comparison of different semantic utilization strategies.

Methods	Semantic desc.		$JAAD_{beh}$	$JAAD_{all}$	PIE	Latency↓	Params.↓	Peak mem.↓	GFLOPs↓
	Train	Infer.	Acc/F1↑ (%)	Acc/F1↑ (%)	Acc/F1↑ (%)	(ms)	(M)	(GB)	
w/o sem.	×	×	72.46/81.25	89.23/66.48	86.71/78.31	90.24	90.59	0.71	1506.58
Online sem. (Qwen3-VL-2B)	√	√	76.50/82.95	92.02/78.29	92.82/86.97	2741.05	~2153.26	4.74	25,910.27
Online sem. (Qwen3-VL-8B)	√	√	76.82/83.77	92.50/77.74	93.49/88.29	5327.45	~8153.26	18.26	105,230.91
Ours (LGD)	√	×	75.97/83.43	91.06/74.01	91.95/85.23	90.24	90.59	0.71	1506.58

Note: All efficiency metrics cover the full inference pipeline, including local and global DINOv2-Reg feature extraction and intention prediction. The online variants additionally include semantic generation and text processing, whereas LGD requires neither at inference.

which are less discernible under rainy conditions, and misclassifies an overhead priority-road sign as a yield sign. In Fig. 8(d), the description introduces a nonexistent yield sign on the right side of the intersection. Together, these examples show that scene-level descriptions may either mischaracterize existing traffic elements or introduce unsupported contextual information.

To evaluate robustness to unreliable semantic distillation, we randomly select 10%, 20%, and 40% of the training samples, shuffle their associated semantic descriptions, and retrain the model under each setting. For comparison, we include the baseline without shuffling (0%), which uses the original VLM-generated descriptions, and a variant without semantic distillation. As shown in Table 10, prediction performance generally decreases as the shuffling ratio increases. At a shuffling ratio of 10%, accuracy decreases by only 0.70, 0.34, and 0.58 pp on $JAAD_{beh}$, $JAAD_{all}$, and PIE, respectively, relative to the baseline. Even when this ratio increases to 20%, the model outperforms the variant without semantic distillation in both accuracy and F1, indicating that the benefits of semantic distillation are retained despite partial shuffling of the descriptions. At a shuffling ratio of 40%, the model still outperforms the variant without semantic distillation on $JAAD_{all}$ and PIE, with comparable performance on $JAAD_{beh}$. Overall, semantic distillation exhibits a certain degree of robustness to unreliable semantic supervision, remaining beneficial at low to moderate shuffling ratios, while its effectiveness generally decreases as the proportion of shuffled descriptions increases.

4.4.3. Effectiveness of language-guided distillation

To investigate the complementary contributions of pedestrian- and scene-level semantic distillation, we compare the individual pedestrian-level and scene-level distillation settings with a no-semantic baseline and the complete model. The results are reported in Table 11. Both individual distillation settings improve performance over the no-semantic baseline, indicating that pedestrian- and scene-level semantics each provide task-relevant information for intention prediction. Pedestrian-level distillation performs better on $JAAD_{beh}$, whereas scene-level distillation provides larger gains on $JAAD_{all}$ and PIE. More importantly, jointly applying pedestrian- and scene-level semantic distillation consistently yields the best performance across all three datasets. Compared with the best individual setting, the complete model further improves Acc/F1 by 1.22/0.94 pp on $JAAD_{beh}$, 0.39/0.96 pp on $JAAD_{all}$, and 1.17/1.91 pp on PIE, respectively. These results demonstrate that pedestrian- and scene-level semantic distillation capture complementary information, and their joint use provides more comprehensive semantic guidance for intention prediction.

To evaluate the effectiveness of language-guided distillation (LGD) and assess the performance-efficiency trade-off of incorporating semantic descriptions at inference, we compare the proposed LGD strategy with a no-semantic baseline and two online semantic input variants. Specifically, the Qwen3-VL-8B variant uses the same semantic description generator as LGD, enabling a direct comparison between online semantic input and LGD-based semantic distillation, whereas the Qwen3-VL-2B variant assesses the feasibility of lightweight online semantic generation. In both online variants, the generated semantic descriptions are incorporated as an additional input modality during training and inference. The results are reported in Table 12.

Just Accepted

Table 13. Performance comparison of different semantic description generation strategies. All metrics are reported in %.

Description generation strategy	$JAAD_{beh}$					$JAAD_{all}$					PIE				
	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R	Acc	AUC	F1	P	R
Template (Base)	72.09	82.13	81.51	69.61	98.30	88.73	92.46	70.73	64.76	77.91	89.36	95.02	82.29	77.42	87.80
Template (Comp.)	75.49	80.93	83.01	73.31	95.67	89.50	92.21	69.67	70.36	68.99	90.59	94.58	82.65	85.93	79.61
Generative (Qwen3-VL)	75.97	85.35	83.43	73.37	96.69	91.06	92.61	74.01	75.24	72.81	91.95	94.57	85.23	88.16	82.50

Table 14. Effect of LGD and CCC on feature compactness and relative separability.

Module config.		$JAAD_{beh}$				$JAAD_{all}$				PIE			
LGD	CCC	Acc↑	D_{intra}↓	D_{inter}↑	R_{sep}↑	Acc↑	D_{intra}↓	D_{inter}↑	R_{sep}↑	Acc↑	D_{intra}↓	D_{inter}↑	R_{sep}↑
×	×	0.69	9.34	4.85	0.52	0.87	12.74	9.32	0.73	0.86	8.99	6.29	0.70
√	×	0.73	9.06	4.90	0.54	0.90	7.08	6.16	0.87	0.90	8.25	6.10	0.74
×	√	0.72	3.22	3.91	1.22	0.89	1.88	4.27	2.28	0.87	3.25	6.34	1.95
√	√	0.76	2.41	4.11	1.70	0.91	1.44	4.44	3.07	0.92	2.39	6.79	2.84

All efficiency measurements were obtained on a single NVIDIA GeForce RTX 3090 Ti GPU under identical inference settings.

LGD consistently outperforms the no-semantic baseline without additional inference overhead, as both settings share the same inference architecture and differ only in the use of semantic distillation during training. Incorporating online-generated semantic descriptions as an additional input modality further improves several prediction metrics. The Qwen3-VL-8B variant achieves the highest accuracy on all three datasets and the best F1 scores on $JAAD_{beh}$ and PIE, while the Qwen3-VL-2B variant obtains the best F1 score on $JAAD_{all}$. However, these improvements come with considerable computational overhead. Compared with LGD, the Qwen3-VL-2B and Qwen3-VL-8B variants increase end-to-end latency from 90.24 ms to 2741.05 ms and 5327.45 ms, respectively, while also exhibiting substantially higher end-to-end parameter counts, peak GPU memory usage, and GFLOPs. Although LGD incurs a modest performance decrease compared with online semantic input, it provides a substantially better performance-efficiency trade-off, achieving 30.38× and 59.04× faster inference than the Qwen3-VL-2B and Qwen3-VL-8B variants, respectively.

4.4.4. Comparison of generative and template-based descriptions

To assess the effect of semantic description generation strategies on language-guided distillation, we compare free-form, image-conditioned descriptions generated by Qwen3-VL with two template-based variants constructed from the structured pedestrian-behavior and scene annotations in JAAD and PIE, as illustrated in Fig. 9. Template (Base) maps each annotation label to a fixed expression and inserts the resulting expressions into a predefined sentence template, whereas Template (Comp.) uses the same template but randomly selects from multiple candidate expressions for each label to increase linguistic diversity. Unlike both template-based variants, Qwen3-VL does not rely on predefined labels or fixed sentence templates.

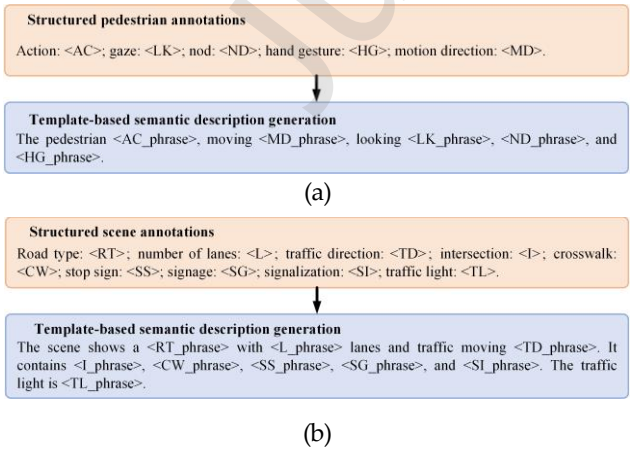


Fig. 9. Template-based semantic description generation from structured pedestrian and scene annotations: (a) pedestrian-level semantic description generation; (b) scene-level semantic description generation.

As shown in Table 13, Template (Comp.) consistently outperforms Template (Base), demonstrating the benefit of linguistic diversity in template-based semantic descriptions. Nevertheless, the generative descriptions achieve the best overall performance, suggesting that their benefits extend beyond linguistic diversity. Free-form, image-conditioned generation captures richer contextual information and finer-grained pedestrian semantics than structured annotations and manually designed templates, thereby providing more informative supervision for language-guided distillation.

4.5. Feature compactness and relative separability analysis

Language-guided distillation (LGD) transfers high-level semantic knowledge into visual representations but does not explicitly constrain their class-wise distribution. To assess the additional contribution of center-constrained classification (CCC) beyond LGD, we analyze whether its class-weighted center loss further improves feature compactness and relative separability under different LGD and CCC configurations. Specifically, we report the intra-class distance D_{intra}, inter-class distance D_{inter}, and relative separation ratio $R_{sep}=D_{inter}/D_{intra}$, where lower D_{intra} indicates greater intra-class compactness, while higher R_{sep} indicates better relative class separability. The results are shown in Table 14.

Compared with the baseline without LGD and CCC, introducing LGD alone consistently improves prediction accuracy and slightly increases R_{sep} across all three datasets, suggesting that distilling high-level semantics into visual representations modestly improves feature discriminability. In contrast, introducing CCC alone substantially reduces D_{intra} and increases R_{sep} across all three datasets, confirming that the class-weighted center loss promotes intra-class compactness and relative class separability. More importantly, combining LGD and CCC achieves the best accuracy,

Just Accepted

Table 15. Cross-dataset generalization performance between JAAD and PIE dataset.

Train → Test	Model	Acc	AUC	F1	P	R
$JAAD_{all}$ → PIE	Global PCPA	0.59	0.52	0.32	0.31	0.34
	LSOP-Net	0.65	0.49	0.18	0.26	0.13
	MTC	0.63	0.53	0.32	0.33	0.32
	VLM-GSDN	0.67	0.66	0.37	0.41	0.33
PIE → $JAAD_{all}$	Global PCPA	0.21	0.28	0.15	0.09	0.39
	LSOP-Net	0.24	0.35	0.19	0.12	0.52
	MTC	0.30	0.42	0.23	0.14	0.60
	VLM-GSDN	0.57	0.54	0.27	0.19	0.46

Table 16. Cross-dataset generalization performance on the WatchPed dataset.

Train → Test	Acc	AUC	F1	P	R
$JAAD_{beh}$ → WatchPed	0.45	0.49	0.55	0.46	0.67
$JAAD_{all}$ → WatchPed	0.38	0.32	0.21	0.28	0.16
PIE → WatchPed	0.72	0.79	0.65	0.86	0.52

the lowest D_{intra}, and the highest R_{sep}. By contrast, D_{inter} does not exhibit a consistent trend across datasets, with the joint use of LGD and CCC achieving the highest value only on PIE. This can be attributed to the fact that CCC primarily promotes intra-class compactness rather than explicitly maximizing inter-class distances. Compared with LGD alone, adding CCC increases R_{sep} from 0.54 to 1.70 on $JAAD_{beh}$, from 0.87 to 3.07 on $JAAD_{all}$, and from 0.74 to 2.84 on PIE, while reducing D_{intra} from 9.06 to 2.41, 7.08 to 1.44, and 8.25 to 2.39, respectively. These results demonstrate that CCC complements the high-level semantic supervision provided by LGD by explicitly constraining class-wise feature distributions, thereby improving intra-class compactness and relative class separability.

4.6. Cross-dataset generalization analysis

To evaluate the cross-dataset generalization of VLM-GSDN, we conduct cross-dataset evaluation between $JAAD_{all}$ and PIE. Models trained on the source dataset are directly evaluated on the target dataset without target-domain fine-tuning. VLM-GSDN is compared with three representative baselines: Global-PCPA (Yang et al., 2022), LSOP-Net (Liu et al., 2025), and MTC (Li and Müller, 2025b). The results are reported in Table 15. Although all methods exhibit performance degradation due to domain shift, VLM-GSDN consistently achieves the best performance. When trained on $JAAD_{all}$ and evaluated on PIE, it achieves 67% accuracy, outperforming LSOP-Net by 2 pp. In the reverse setting, it achieves 57% accuracy and surpasses MTC by 27 pp. These results demonstrate the promising cross-dataset generalization of VLM-GSDN and suggest that the high-level semantic knowledge introduced through language-guided distillation contributes to more transferable feature representations.

To further assess cross-dataset generalization, we directly evaluate models trained on $JAAD_{beh}$, $JAAD_{all}$, and PIE on the unseen WatchPed dataset (Abbasi et al., 2023) without fine-tuning. WatchPed contains 255 pedestrian-vehicle interaction videos and covers diverse long-tail scenarios, including nighttime scenes and distant pedestrians, providing a challenging testbed for evaluating transferability under complex real-world conditions. The results

Table 17. Performance comparison under different prediction horizons.

TTE (s)	Models	$JAAD_{beh}$		$JAAD_{all}$		PIE	
		Acc	F1	Acc	F1	Acc	F1
0.5-1	Global PCPA	0.63	0.74	0.85	0.67	0.92	0.87
	LSOP-Net	0.71	0.80	0.86	0.61	0.92	0.85
	MTC	0.74	0.80	0.91	0.78	0.90	0.81
	VLM-GSDN	0.80	0.86	0.92	0.76	0.94	0.89
1-2	Global PCPA	0.62	0.74	0.83	0.63	0.89	0.80
	LSOP-Net	0.65	0.78	0.85	0.58	0.89	0.81
	MTC	0.71	0.80	0.90	0.70	0.89	0.80
	VLM-GSDN	0.76	0.83	0.91	0.74	0.92	0.85
2-3	Global PCPA	0.53	0.65	0.79	0.57	0.78	0.65
	LSOP-Net	0.66	0.77	0.83	0.54	0.83	0.68
	MTC	0.68	0.79	0.83	0.62	0.80	0.60
	VLM-GSDN	0.70	0.81	0.86	0.65	0.84	0.71

are presented in Table 16. Models trained on $JAAD_{beh}$ and $JAAD_{all}$ achieve relatively limited transfer performance, with accuracies of 45% and 38%, respectively. In particular, the $JAAD_{all}$-trained model obtains only 16% recall, possibly because the predominance of negative samples in $JAAD_{all}$ biases the model toward the negative class, leading to more false negatives on the more balanced WatchPed dataset. By contrast, the PIE-trained model achieves the best transfer performance, attaining 72% accuracy and the highest AUC, F1 score, and precision. This suggests that training on PIE yields more transferable representations, potentially due to its greater diversity in traffic scenes and pedestrian behaviors. Overall, these results suggest that cross-dataset generalization is influenced by the source-domain distribution, while the PIE-trained VLM-GSDN shows promising transfer performance under challenging real-world conditions.

4.7. Prediction horizon analysis

To evaluate VLM-GSDN across different prediction horizons, we conduct experiments under three time-to-event (TTE) settings: 0.5-1 s, 1-2 s, and 2-3 s. The 1-2 s setting follows the benchmark protocol (Kotseruba et al., 2021), whereas the other two represent short- and extended-horizon prediction, respectively. We compare VLM-GSDN with three representative baselines: Global-PCPA (Yang et al., 2022), LSOP-Net (Liu et al., 2025), and MTC (Li and Müller, 2025b). The results are reported in Table 17. Overall, performance generally degrades as TTE increases, as pedestrian intentions become more uncertain and discriminative cues less explicit at longer horizons. Under the short-horizon setting, VLM-GSDN achieves the highest accuracies of 80%, 92%, and 94% on $JAAD_{beh}$, $JAAD_{all}$, and PIE, outperforming the second-best methods by 6, 1, and 2 pp, respectively. Under the extended-horizon setting, VLM-GSDN achieves the highest accuracy and F1 score across all three datasets, indicating robust performance at longer prediction horizons.

4.8. Computational efficiency evaluation

To evaluate computational efficiency, we compare VLM-GSDN with representative baselines on $JAAD_{all}$ in terms of accuracy, GFLOPs, peak GPU memory, inference latency, model size, and parameter count. All measurements are obtained on a single NVIDIA GeForce RTX 3090 Ti GPU under identical inference settings. We report computational costs at both the prediction-

Not Accepted

Table 18. Comparison of computational costs with representative baselines and temporal attention variants on the $JAAD_{all}$ dataset.

Models	Acc	GFLOPs↓		Peak mem.↓ (MB)		Latency↓ (ms)		Size↓ (MB)		Params↓ (M)	
		Pred.	Full	Pred.	Full	Pred.	Full	Pred.	Full	Pred.	Full
PCPA (Kotseruba et al., 2021)	0.77	77.20	7834.92	289.32	864.03	31.18	160.42	118.80	189.76	31.17	51.19
Global PCPA (Yang et al., 2022)	0.83	154.30	3350.76	32.39	2246.35	35.70	690.91	374.20	602.46	60.99	122.32
Razali et al. (2021)	-	233.70	238.85	430.44	439.27	17.98	26.31	94.10	101.24	25.31	23.52
MTMGN (Yang et al., 2024a)	0.89	2.22	-	156.04	-	13.19	-	8.02	-	2.01	-
ACIT (Li and Müller, 2025a)	0.89	2.36	3061.77	308.11	1074.01	43.93	408.18	62.50	506.40	5.15	121.51
Gated-S2R-PCP (Bai et al., 2025)	0.88	313.98	-	775.21	-	22.89	-	687.33	-	176.70	-
MTC (Li and Müller, 2025b)	0.90	4.78	7093.64	82.21	1505.58	36.23	760.07	99.70	913.50	8.25	216.85
TCP (Xiao et al., 2025)	0.89	0.04	-	157.13	-	3.29	-	147.39	-	36.85	-
OmniPredict (Ham et al., 2026)	-	-	-	-	-	-	1160.00	-	-	-	-
VLMPed-CoT (Ling et al., 2026)	0.90	-	62,472.85	-	8478.48	-	1146.62	-	~7520	-	~3750
VLM-GSDN w/ FTSA	0.90	0.13	1506.64	35.63	789.55	10.59	96.37	16.06	362.37	4.01	90.59
VLM-GSDN	0.91	0.07	1506.58	18.63	710.05	4.44	90.24	16.06	362.37	4.01	90.59

Note: Pred. denotes the prediction stage alone, whereas Full denotes the complete inference pipeline, including, where applicable, upstream models for derived-modality generation, data preprocessing and visual feature extraction. For VLM-GSDN, Full includes DINOv2-Reg-based feature extraction for both local and global images, followed by intention prediction. A dash (–) indicates that the corresponding metric is not applicable, was not reported in the original publication, or could not be reproduced using the publicly available implementation.

stage and full-pipeline levels. The results are reported in Table 18.

At the prediction-stage level, VLM-GSDN achieves the highest accuracy, the lowest peak GPU memory, and the second-lowest GFLOPs, inference latency, model size, and parameter count. Its lower computational cost than representative multimodal methods, such as Global PCPA (Yang et al., 2022), ACIT (Li and Müller, 2025a), and MTC (Li and Müller, 2025b), can largely be attributed to its reduced reliance on heterogeneous inputs and complex fusion operations. Although slightly slower than TCP (4.44 vs. 3.29 ms), VLM-GSDN achieves higher accuracy (0.91 vs. 0.89) and substantially lower peak GPU memory usage (18.63 vs. 157.13 MB).

At the full-pipeline level, VLM-GSDN ranks second in GFLOPs, peak GPU memory, and inference latency, and third in model size and parameter count. Razali et al. (2021) achieve the lowest latency because their method processes a single RGB frame without temporal modeling. By avoiding additional pretrained models for derived modalities such as optical flow, semantic maps, or depth maps, VLM-GSDN achieves lower latency than Global PCPA, ACIT, and MTC. Compared with representative VLM-based methods, VLM-GSDN is 12.85× and 12.71× faster than OmniPredict (Ham et al., 2026) and VLMPed-CoT (Ling et al., 2026), respectively. It also requires only approximately 2.4% of the parameters and 8.4% of the peak GPU memory used by VLMPed-CoT.

To quantify the computational benefits of selective temporal attention (STA) module, we replace it with full temporal self-attention (FTSA) while keeping all other components unchanged. Compared with FTSA, STA reduces the prediction-stage GFLOPs, peak GPU memory, and latency by approximately 46.15%, 47.71%, and 58.07%, respectively, while improving accuracy from 0.90 to 0.91. FTSA computes pairwise interactions among all T temporal elements, producing a $T\times T$ attention map with $O(T^2d)$ computational and $O(T^2)$ memory complexity. In contrast, STA uses a learnable summary token as a single query, reducing the attention map to $1\times T$ and the corresponding complexities to $O(Td)$ and $O(T)$. At the full-pipeline level, the reductions are smaller because DINOv2-Reg feature extraction dominates the overall computational cost.

4.9. Robustness to bounding-box perturbations

For fair comparison with baseline methods, VLM-GSDN is trained and evaluated using ground-truth pedestrian bounding boxes. In real-world applications, however, detector-generated bounding boxes may contain localization errors, scale variations, and temporal jitter. To assess the effect of such errors, we add Gaussian noise with standard deviations of 0, 0.05, 0.1, 0.2, and 0.4 to the annotated bounding boxes during inference, where 0 denotes no perturbation. The results are reported in Table 19.

Prediction performance generally decreases as the noise standard deviation increases, indicating that larger bounding-box deviations can impair motion-cue modeling. On $JAAD_{beh}$, VLM-GSDN remains stable, with accuracy decreasing only from 75.97% to 74.85% as the noise standard deviation increases from 0 to 0.4. On $JAAD_{all}$ and PIE, performance is relatively stable under mild perturbations but declines more noticeably at standard deviations of 0.2 and above. These results indicate that VLM-GSDN remains relatively stable under mild bounding-box perturbations but becomes more sensitive as localization errors increase, highlighting the importance of reliable bounding-box localization in practical deployment.

4.10. Qualitative results

We present qualitative results in several representative scenarios, as shown in Fig. 10. The ground-truth bounding boxes of the target pedestrians are marked in red, while the predicted bounding boxes are marked in green. GT denotes ground truth, and C and NC denote Crossing and Not Crossing, respectively. The right panel

Table 19. Robustness of VLM-GSDN to varying levels of Gaussian bounding-box perturbations.

Noise std	$JAAD_{beh}$		$JAAD_{all}$		PIE	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
0	75.97	83.43	91.06	74.01	91.95	85.23
0.05	75.23	83.02	89.65	71.82	90.38	82.31
0.1	75.01	82.87	88.12	69.18	87.42	77.61
0.2	74.96	82.83	86.48	65.21	86.08	75.11
0.4	74.85	82.76	85.83	63.19	85.12	72.69

Not Accepted

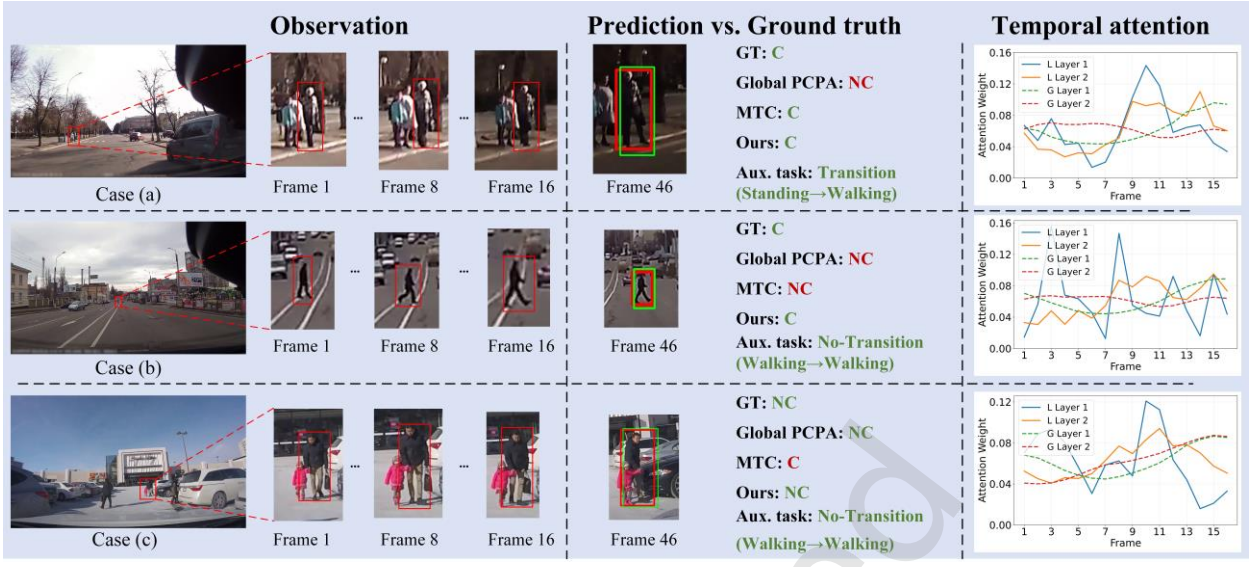


Fig. 10. Qualitative analysis of the intention prediction results.

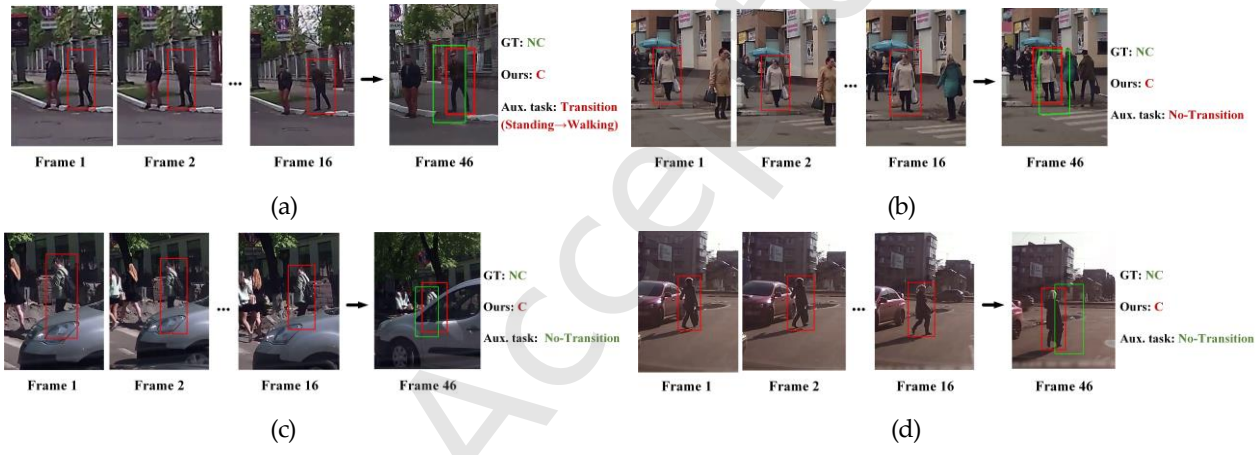


Fig. 11. Representative failure cases of VLM-GSDN under challenging conditions: (a) stationary pedestrian with ambiguous posture; (b) misleading crossing cues; (c) partial occlusion; (d) misleading motion tendency.

shows the temporal attention weights of the summary token over different frames in the STA module. We compare VLM-GSDN with Global PCPA (Yang et al., 2022) and MTC (Li and Müller, 2025b).

In case (a), a pedestrian is standing still beside the crosswalk and looking toward the ego-vehicle. By leveraging language-guided distillation, VLM-GSDN can capture such high-level semantic cues and correctly predict both the crossing intention and the action-state transition from standing to walking, whereas Global PCPA fails. In case (b), without yielding-related traffic signs, a distant pedestrian continues moving. VLM-GSDN can implicitly capture the high-level semantic cues underlying this continuous motion and thus correctly predict the crossing intention, whereas MTC and Global PCPA fail. In case (c), in a parking lot, a pedestrian is walking toward the ego-vehicle. From short-term observations, VLM-GSDN accurately predicts the future bounding-box displacement and correctly predicts a no-crossing intention. Moreover, the temporal attention shown on the right indicates that the model selectively focuses on key frames of the local RGB image sequence, whereas the attention over the global RGB image sequence is more dispersed and evenly distributed.

To provide a more systematic analysis of model limitations, Fig. 11 presents representative failure cases under several challenging conditions. As shown in Fig. 11(a), a pedestrian remains stationary at the roadside while leaning forward and looking toward the ego-vehicle. This ambiguous posture resembles preparatory behavior for crossing, leading VLM-GSDN to incorrectly predict an action-state transition from standing to walking. The predicted future bounding box is also shifted toward the roadway relative to the ground truth, further contributing to the incorrect crossing prediction. As shown in Fig. 11(b), the pedestrian moves continuously toward the crosswalk during the observation window, but eventually stops before entering it. The approaching motion and the presence of a nearby crosswalk provide strong crossing-related cues, leading VLM-GSDN to incorrectly predict a crossing intention. As shown in Fig. 11(c), the target pedestrian is partially occluded by a nearby vehicle, reducing the visibility of body posture and motion cues and resulting in a false crossing prediction. As shown in Fig. 11(d), the pedestrian exhibits continuous lateral motion during the observation window, providing strong crossing-related motion cues. However, in the subsequent future frames, the pedestrian stops and yields to the

Just Accepted

ego-vehicle, making the observed motion tendency misleading and resulting in a false crossing prediction. Overall, these failure cases illustrate several challenging scenarios involving ambiguous motion and posture cues, partial occlusion, and potentially misleading contextual information, under which pedestrian crossing intention is inherently difficult to predict.

5 Conclusions

This paper presents a vision language model-guided semantic distillation network (VLM-GSDN) for pedestrian crossing intention prediction. VLM-GSDN leverages a VLM to capture scene-level context and fine-grained pedestrian semantics and distills the resulting high-level semantic knowledge into visual representations through language-guided distillation. Because VLM is used only during training, VLM-GSDN retains a compact inference architecture and incurs no additional VLM-related overhead at deployment. In addition, pedestrian motion and predicted bounding-box displacement are used to modulate visual features, yielding motion-aware pedestrian representations. Furthermore, pedestrian action-state transition prediction is introduced as an auxiliary task to enhance the modeling of pedestrian behavior dynamics, while the class-weighted center loss further improves feature discriminability. The comparative results demonstrate that VLM-GSDN achieves competitive performance against existing state-of-the-art methods, with accuracies of 76%, 91%, and 92% on the $JAAD_{beh}$, $JAAD_{all}$, and PIE datasets, respectively, while requiring fewer input modalities than methods with comparable performance. Ablation studies further confirm the effectiveness of its individual modules. Moreover, experimental results show that VLM-GSDN maintains stable performance under extended prediction horizons. It also achieves a full-pipeline inference latency of 90.24 ms. Compared with online semantic generation using Qwen3-VL-8B, VLM-GSDN achieves 59.04× faster inference with minimal performance loss, demonstrating a favorable performance-efficiency trade-off for practical deployment.

Future work will focus on practical deployment and model lightweighting. We will also investigate more robust modeling of ambiguous motion, partial occlusion, and misleading contextual cues to improve prediction reliability under challenging conditions. We will further investigate domain-specific fine-tuning of the open-source VLM to improve the factual accuracy and consistency of generated semantic descriptions, thereby reducing hallucinations and providing more reliable supervision for semantic distillation. In addition, we will explore the scalability of the proposed framework for fine-grained intention prediction by extending the current binary setting to multiple intention categories, such as immediate crossing, hesitant crossing, and no crossing.

CRediT authorship contribution statement

Yuanzhe Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft. **Cheng Wang:** Investigation, Writing – review & editing. **Zhuoqi Zeng:** Resources. **Steffen Müller:** Supervision, Project administration.

Replication and data sharing

The source codes and replication package are available on ETS data at https://doi.org/10.26599/ETSD.2026.9190104.

Acknowledgements

The authors gratefully acknowledge Navid Mohammad Imran (University of Memphis) for providing access to the WatchPed dataset and Keyuan Luo (Hainan Bielefeld University of Applied Sciences) for his contribution to dataset annotation. This work was supported in part by the China Scholarship Council (Grant No. 202206030065).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language polishing in parts of the manuscript. After utilizing this tool, the authors thoroughly reviewed and edited the content as necessary and took full responsibility for the final content of the published article.

References

Abbasi, J.A., Imran, N.M., Das, L.C., Won, M., 2023. WatchPed: pedestrian crossing intention prediction using embedded sensors of smartwatch. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 9574–9581.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., et al., 2024. Phi-3 technical report: a highly capable language model locally on your phone. https://doi.org/10.48550/arXiv.2404.14219

AlShami, A.K., Boult, T.E., Kalita, J., 2026. GATEPose: A graph attention transformer enhanced with pose and orientation angles for pedestrian crossing intention prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 1810–1819.

Azarmi, M., Rezaei, M., Wang, H., 2025a. PIP-Net: pedestrian intention prediction in the wild. *IEEE Trans Intell Transp Syst*, **26** (7), 9824–9837.

Azarmi, M., Rezaei, M., Wang, H., 2025b. Pedestrian intention prediction via vision-language foundation models. In: 2025 IEEE Intelligent Vehicles Symposium (IV), 1899–1904.

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** (6), 7509–7522.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al., 2025. Qwen3-VL: technical report. https://doi.org/10.48550/arXiv.2511.21631

Bossen, T.E.W., Møgelmose, A., Greer, R., 2025. Can vision-language models understand and interpret dynamic gestures from pedestrians? Pilot datasets and exploration

Just Accepted

towards instructive nonverbal commands for cooperative autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4818–4827.

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: Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), 2000–2005.

Chen, H., Sun, J., Liu, Y., Peng, Z., Hu, D., 2025. Causal confusion in pedestrian crossing intention prediction for autonomous vehicles: the role of ego-vehicle speed. *IEEE Trans Intell Transp Syst*, **26** (8), 12224–12237.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. https://doi.org/10.48550/arXiv.1706.05587

Chen, X., Zhang, S., Li, J., Yang, J., 2024. Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving. *IEEE Trans Intell Transp Syst*, **25** (9), 12538–12549.

Choi, S., Wu, J., An, Q., Yamashita, A., 2026. Texting-while-walking detection in real-world environments using vision-language models with prompt engineering. In: Proceedings of the IEEE/SICE International Symposium on System Integration (SII), 612–617.

Darcet, T., Oquab, M., Mairal, J., Bojanowski, P., 2023. Vision transformers need registers. https://doi.org/10.48550/arXiv.2309.16588

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al., 2020. An image is worth 16x16 words: transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929

Du, Q., Xu, L., Wu, Q., Ning, H., Wang, X., Lin, L., et al., 2026. SafeCrossNet: multi-modal fusion with social-aware for pedestrian crossing intention prediction. *Inf Fusion*, **126**, 103609.

Elgazwy, A., Elgazzar, K., Khamis, A., 2025. Predicting pedestrian crossing intentions in adverse weather with self-attention models. *IEEE Trans Intell Transp Syst*, **26** (3), 3250–3261.

Gao, H., Zhang, L., Zhao, Y., Yang, Z., Cao, J., 2025. Application of vision-language models to pedestrian behavior prediction and scene understanding in autonomous driving. In: Proceedings of the International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC), 1–8.

Ham, J.S., Kim, S., Huang, J., Jiang, P., Moon, J., Saripalli, S., et al., 2024. LLaMAPed: multi-modal pedestrian crossing intention prediction. In: European Conference on Computer Vision, 150–167.

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

Huang, J., Jiang, P., Gautam, A., Saripalli, S., 2024. GPT-4V takes the wheel: promises and challenges for pedestrian behavior prediction. In: Proceedings of the AAAI Symposium Series, 3 (1), 134–142.

Hussien, M.M., Melo, A.N., Ballardini, A.L., Salinas Maldonado, C., Izquierdo, R., Sotelo, M.Á., 2025. RAG-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models. *Expert Syst Appl*, **265**, 125914.

Kong, Q., Kawana, Y., Saini, R., Kumar, A., Pan, J., Gu, T., et al., 2024. WTS: a pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In: European Conference on Computer Vision, 1–18.

Kotseruba, I., Rasouli, A., Tsotsos, J.K., 2021. Benchmark for evaluating pedestrian action prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1258–1268.

Kuang, S., Gao, Y., Cong, S., Liu, Y., Wei, Y., 2026. TrafficPerceiver: a multimodal large language model with reinforcement learning for unified challenging traffic scene perception. *Commun Transp Res*, **6**, 9640008.

Li, Y., Müller, S., 2025a. ACIT: attention-guided cross-modal interaction transformer for pedestrian crossing intention prediction. In: Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC), 2018–2024.

Li, Y., Müller, S., 2025b. MTC: multimodal transformer with cross-modality guided attention for pedestrian crossing intention prediction. *IEEE Sens J*, **25** (24), 44929–44939.

Li, Y., Zhong, H., Müller, S., 2026. Multi-context fusion transformer for pedestrian crossing intention prediction in urban environments. *IET Intell. Transp. Syst.*, **20** (1), e70354.

Liang, H., Qiao, H., Huang, W., Wang, Q., Shang, M., Chen, L., 2024. Temporal-contextual event learning for pedestrian crossing intent prediction. In: International Conference on Neural Information Processing, 240–254.

Lin, C., Hu, X., 2026. Pedestrian crossing intention prediction based on cross-modal motion query and global–local context co-learning for autonomous driving. *IEEE Trans Intell Transp Syst*, **27** (5), 5804–5819.

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

Ling, Y., Ma, Z., Zhang, Q., Xie, B., Weng, X., 2024. PedAST-GCN: fast pedestrian crossing intention prediction using spatial–temporal attention graph convolution networks. *IEEE Trans Intell Transp Syst*, **25** (10), 13277–13290.

Ling, Y., Qin, Z., Wang, L., Liu, Z., Liu, Y., Ma, Z., 2026. VLMPed-CoT: A large vision-language model with a chain-of-thought mechanism for pedestrian crossing intention prediction. *Commun Transp Res*, **6** (1), 9640009.

Just Accepted

Liu, H., Liu, C., Chang, F., Lu, Y., Liu, M., 2025. Long–short observation-driven prediction network for pedestrian crossing intention prediction with momentary observation. *Neurocomputing*, **614**, 128824.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., et al., 2022. Swin Transformer V2: scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12009–12019.

Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D.F., Sotelo, M.A., 2020. RNN-based pedestrian crossing prediction using activity and pose-related features. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV), 1801–1806.

Mishra, N., Gangisetty, S., Jawahar, C.V., 2026. PEDESTRIANQA: A benchmark for vision-language models on pedestrian intention and trajectory prediction. https://doi.org/10.48550/arXiv.2605.24562

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.

Munir, F., Kucner, T.P., 2025. Context-aware multi-task learning for pedestrian intent and trajectory prediction. *Transp Res Part C Emerg Technol*, **178**, 105203.

Ni, R., Yang, B., Wei, Z., Hu, H., Yang, C., 2023. Pedestrians crossing intention anticipation based on dual-channel action recognition and hierarchical environmental context. *IET Intell Transp Syst*, **17** (2), 255–269.

Osman, N., Camporese, G., Ballan, L., 2024. Multi-modal transformer with language modality distillation for early pedestrian action anticipation. *Comput Vis Image Underst*, **249**, 104144.

Piccoli, F., Balakrishnan, R., Perez, M.J., Sachdeo, M., Nuñez, C., Tang, M., 2020. FUSSI-Net: fusion of spatio-temporal skeletons for intention prediction network. In: Proceedings of the 54th Asilomar Conference on Signals, Systems, and Computers, 68–72.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, 8748–8763.

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

Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K., 2019. PIE: a large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 6262–6271.

Rasouli, A., Kotseruba, I., 2023. PedFormer: pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 9844–9851.

Razali, H., Mordan, T., Alahi, A., 2021. Pedestrian intention prediction: a convolutional bottom-up multi-task approach. *Transp Res C Emerg Technol*, **130**, 103259.

Saleh, K., Hossny, M., Nahavandi, S., 2018. Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks. *IEEE Trans Intell Veh*, **3** (4), 414–424.

Shi, L., Jiang, B., Zeng, T., Guo, F., 2025. ScVLM: enhancing vision-language model for safety-critical event understanding. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops, 1061–1071.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. https://doi.org/10.48550/arXiv.1409.1556

Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y., 2020. MPNet: masked and permuted pre-training for language understanding. *Adv Neural Inf Process Syst*, **33**, 16857–16867.

Tan, M., Le, Q., 2019. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, 6105–6114.

Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., et al., 2024. DriveVLM: the convergence of autonomous driving and large vision-language models. https://doi.org/10.48550/arXiv.2402.12289

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, 10347–10357.

Tran, T.T.M., Yu, X., Parker, C., Berrío Perez, J.S., Worrall, S., Tomitsch, M., 2025. Understanding pedestrian gesture misrecognition: insights from vision-language model reasoning. In: Adjunct Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 138–144.

Uziel, R., Bialer, O., 2025. Optimizing vision-language model for road crossing intention estimation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 1702–1712.

Varytimidis, D., Alonso-Fernandez, F., Duran, B., Englund, C., 2018. Action and intention recognition of pedestrians in urban traffic. In: Proceedings of the International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 676–682.

Wang, Z., Papanikolopoulos, N., 2020. Estimating pedestrian crossing states based on single 2D body pose. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2205–2210.

Xiao, Y., Liu, W., Huang, W., Ma, J., Liu, R.W., Zhong, X., 2025. TCP: text-guided cascade network for pedestrian crossing intention prediction. *IEEE Trans Intell Transp Syst*, **27** (1), 831–841.

Xie, C., Lin, C., Zheng, X., Gong, B., López, A.M., 2025. GTransPDM: a graph-embedded transformer with positional decoupling for pedestrian crossing intention prediction. *IEEE Signal Process Lett*, **32**, 2109–2113.

Xu, L., You, S., He, G., Li, Y., 2024. Pedestrian-vehicle information modulation for pedestrian crossing intention prediction. *IEEE Trans Intell Veh*, **10** (3), 1919–1930.

Just Accepted

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al., 2025. Qwen3 technical report. https://doi.org/10.48550/arXiv.2505.09388

Yang, B., Wei, Z., Hu, H., Wang, R., Yang, C., Ni, R., 2023. DPCIAN: A novel dual-channel pedestrian crossing intention anticipation network. *IEEE Trans Intell Transp Syst*, **25** (6), 6023–6034.

Yang, B., Fan, Z., Hu, H., Hu, C., Ni, R., 2024a. Explainable pedestrian crossing intention prediction based on multi-task mutual guidance network. *IEEE Trans Intell Veh*, **10** (2), 1342–1352.

Yang, B., Wei, Z., Hu, C., Cai, Y., Wang, H., Hu, H., 2024b. Real-time pedestrian crossing anticipation based on an action–interaction dual-branch network. *IEEE Trans Intell Transp Syst*, **25** (12), 21021–21034.

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** (2), 221–230.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., et al., 2024. Depth Anything V2. *Adv Neural Inf Process Syst*, **37**, 21875–21911.

Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X., 2021. BiTraP: bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robot Autom Lett*, **6** (2), 1463–1470.

Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., et al., 2026. MiniCPM-V 4.5: cooking efficient MLLMs via architecture, data, and training recipe. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11704–11715.

Zambare, P., Thanikella, V.N., Liu, Y., 2025. Seeing beyond frames: zero-shot pedestrian intention prediction with raw temporal video and multimodal cues. In: Proceedings of the International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), 1–5.

Zeng, L., Xu, G., Han, Q., Ding, X., Ye, L., Hu, H., 2024. GANet: a pedestrian crossing intention prediction method based on group modelling and individual abnormal action detection. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), 1–8.

Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J., 2024. Long-CLIP: unlocking the long-text capability of CLIP. In: European Conference on Computer Vision, 310–325.

Zhang, X., Wang, X., Zhang, W., Wang, Y., Liu, X., Wei, D., 2024. Multi-attention network for pedestrian intention prediction based on spatio-temporal feature fusion. *Proc Inst Mech Eng Part D: J Automob Eng*, **238** (13), 4202–4215.

Zhang, Y., Nie, Y., 2024. Interndrive: a multimodal large language model for autonomous driving scenario understanding. In: Proceedings of the International Conference on Artificial Intelligence, Automation and High Performance Computing, 294–305.

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** (12), 14213–14225.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., et al., 2025. InternVL3: exploring advanced training and test-time recipes for open-source multimodal models. https://doi.org/10.48550/arXiv.2504.10479



Yuanzhe Li received the B.Eng. degree from Harbin Institute of Technology, China, in 2019, and the M.Sc. degree from Beijing Institute of Technology, China, in 2022. He is currently pursuing Ph.D. degree with the Chair of Automotive Engineering, Technical University of Berlin, Germany. His research interests include pedestrian intention prediction, and vehicle dynamics control.



Cheng Wang received the B.S. degree from Wuhan University of Technology, China, in 2014, and the M.S. degree from the School of Automotive Studies, Tongji University, China, in 2017, and the Ph.D. degree from the Institute of Automotive Engineering, Technical University of Darmstadt, Germany, in 2021. From 2022 to 2024, he was a Research Associate with the University of Edinburgh. He is currently an Assistant Professor with the Heriot-Watt University, Edinburgh. His research interests include explainable AI and safety verification and validation of autonomous vehicles.



Zhuoqi Zeng received the B.Sc. and M.Sc. degrees in mechanical engineering from the University of Duisburg-Essen, Germany, in 2013 and 2015, respectively. From 2016 to 2020, he worked at Bosch (China) Investment Ltd. as a Research Engineer and Product Manager for the Sensor Fusion team. During this time, he also pursued his PhD in Signal Processing at the University of Kaiserslautern. From 2020 to 2022, he served as a Technical Expert in Driver Assistance Systems and Project Manager for Autonomous Driving at Robert Bosch GmbH. From 2022 to 2025, he took on the role of Professor of Mechatronics at Fachhochschule des Mittelstands in Germany. Since 2025, he has held the position of Dean of School of Engineering and Professor at Hainan Bielefeld University of Applied Sciences, China.

Just Accepted



Steffen Müller received the Dipl.-Ing. degree in astronautics and aerospace engineering in 1993 and the Dr.-Ing. degree from Technical University of Berlin in 1998. From 1998 to 2000, he was a Project Manager with ABB Corporate Research Center, Heidelberg, Germany. He finished his post-doctoral research at the University of California at Berkeley in 2001. From 2001 to 2008, he took up different leading positions at the BMW Research and Innovation Centre. From 2008 to 2013, he was the Founder and the Leader of the Chair for Mechatronics in Engineering and Vehicle Technology, Technical University of Kaiserslautern, Germany. He is currently the University Professor and an Einstein Professor with Technical University of Berlin, Germany. He is the Head of the Chair of Automotive Engineering, Faculty of Mechanical Engineering and Transport Systems.

Just Accepted