Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pedestrian Crossing Intention Prediction via Semantic Motion Representation and Cognitive Interaction Modeling

Yuemei Xi
Shijiazhuang Tiedao University
Regan Thomas Potangaroa
Auckland University of Technology
Kelvin Xi ZUO
Massey University
Chenchen Song
Affiliated Hospital of Hebei University
Wensheng Zhang
Shijiazhuang Tiedao University
Kun Liu
Shijiazhuang Tiedao University
Yan Yang
yanyang@stdu.edu.cn
Shijiazhuang Tiedao University

Research Article

Keywords: Pedestrian crossing intention prediction, Autonomous driving, Interpretable prediction, Semantic motion representation, Cognitive interaction modeling, Laban Movement Analysis

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11033052/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Pedestrian Crossing Intention Prediction via Semantic Motion Representation and Cognitive Interaction Modeling

Abstract

Accurate pedestrian crossing intention prediction is essential for autonomous-driving decision support because vehicles must anticipate crossing behavior before explicit stepping actions or displacement occur. Existing methods mainly learn associations between observable features and crossing labels from trajectories, skeleton dynamics, and context, while paying limited attention to the cognitive process underlying pedestrian crossing intention. This neglect weakens interpretability and limits the reliability of early prediction in safety-critical scenarios. To address this, this paper proposes Psychology-informed Semantic Motion and Cognitive Interaction Decision (PsyMIND), an explainable early crossing-intention prediction framework for autonomous driving. The framework infers latent decision logic from observable behavior and pedestrian-related contextual cues. First, Laban Movement Analysis transforms pedestrian skeleton sequences into psychology-informed semantic motion representations along four dimensions: Body, Shape, Effort, and Space. Second, pedestrian-related contextual cues, including traffic constraints, individual attributes, and communicative signals, describe the decision conditions under which crossing judgments occur. Furthermore, a cognitive interaction modeling method is used to represent the internal decision logic underlying the formation of pedestrian crossing intention, including rule compliance, risk acceptance, waiting cost, and safety-efficiency trade-offs. Finally, semantic motion features, contextual features, and cognitive interaction decision features are integrated into a unified crossing-intention representation, and a support vector machine is adopted for classification. Experiments on the Joint Attention in Autonomous Driving dataset show that PsyMIND achieves an accuracy of 0.76, an area under the receiver operating characteristic curve of 0.83, and a recall of 0.93. These results indicate that PsyMIND provides an interpretable and safety-oriented basis for early prediction.

Keywords: Pedestrian crossing intention prediction, Autonomous driving, Interpretable prediction, Semantic motion representation, Cognitive interaction modeling, Laban Movement Analysis

1 Introduction

Accurate pedestrian crossing intention prediction is a critical step for autonomous-driving systems to move from environmental perception to decision control. Intelligent transportation systems increasingly use AI-driven prediction and decision support to improve mobility and traffic safety (Delavari et al. 2026). Before making a crossing decision, pedestrians usually perceive contextual information such as vehicle motion, road environment, and traffic signals, and then combine it with individual attributes and their own motion states to cognitively evaluate crossing opportunities, risk levels, rule constraints, and crossing efficiency, thereby forming a short-term decision tendency regarding whether to cross. However, for autonomous-driving systems, it remains challenging to infer pedestrians' latent decision logic from observable behavior and pedestrian-related contextual cues, and to understand the cognitive process behind it, so as to improve the reliability and interpretability of decision support.

Current pedestrian intention prediction methods have evolved along multiple directions and can be grouped into three categories: trajectory-based methods, skeleton geometry-based methods, and multimodal fusion-based methods. Trajectory-based approaches simplify pedestrians as moving points in space and extrapolate future motion from historical trajectories. Representative trajectory-based studies include the contextual path prediction model of Kooij et al. (2014) and the probabilistic short-term crossing prediction method of Keller et al. (2011). Skeleton geometry-based methods exploit skeletal geometric information and attempt to infer pose semantics from joint distributions. Fang and López (2018) performed intention recognition using 2D pose estimation, while Gesnouin et al. (2021) proposed the TrouSPI-Net model, which incorporates a spatiotemporal attention module to extract skeletal motion features more accurately.

An increasing number of studies have adopted multimodal fusion strategies that integrate pedestrian pose with external cues, such as traffic signal states, vehicle motion trends, and road geometry, to construct a more comprehensive predictive context. Chen et al. (2024) proposed a prediction framework based on a cross-modal Transformer and uncertainty-aware multitask learning. Chen et al. (2025) proposed a multimodal fusion framework that combines pedestrian bounding boxes, pose skeletons, and ego-vehicle speed for pedestrian crossing intention prediction.

Although the above studies have made progress in feature modeling, information fusion, and predictive performance, their overall research paradigm still largely focuses on optimizing observable feature combinations and improving prediction accuracy. However, insufficient attention has been paid to pedestrians' psychological decision-making process. Pedestrian crossing is not merely driven by trajectory variation or pose change, but is a dynamic cognitive outcome shaped by external environmental stimuli, individual risk perception, mobility demand, and social norms. Without explicitly modeling this intention formation mechanism, existing methods may have limited ability to provide interpretable and reliable early prediction in complex human–vehicle interaction scenarios. Such limitations mainly arise because motion representation is often insufficiently semantic, while decision logic remains largely implicit. Therefore, pedestrian crossing intention prediction needs to move beyond observable feature

aggregation and explicitly model the psychological decision-making process underlying pedestrian crossing intention.

This paper therefore proposes **PsyMIND** (Psychology-informed Semantic Motion and Cognitive Interaction Decision), a pedestrian crossing intention prediction framework that integrates fine-grained motion quality with high-level cognitive decision-making. The proposed framework models pedestrian crossing intention formation from two complementary perspectives: psychology-informed semantic motion representation and cognitive interaction-driven decision modeling. Specifically, a Psychological Semantic Motion (**PSM**) Encoder is developed to elevate raw skeletal motion into psychology-informed semantic action features by incorporating Laban Movement Analysis (LMA), thereby characterizing pedestrians' crossing tendencies and motion quality. In parallel, temporal environmental context features (**TECF**) are constructed to describe traffic constraints, pedestrian attributes, and communicative cues. A Cognitive Interaction Decision (**CID**) Module is then introduced to capture the latent cognitive decision-making mechanism by modeling the trade-off between safety and crossing efficiency. Finally, action semantics, environmental constraints, and cognitive decision features are fused for pedestrian crossing intention prediction. This hierarchical design improves early intention recognition while making the model's decision process more transparent and interpretable. The main contributions of this paper are summarized as follows:

1. A psychological semantic motion representation method is proposed for early pedestrian crossing intention prediction, and the PSM Encoder is designed accordingly. By introducing LMA, raw skeleton sequences are mapped into an action semantic space with four dimensions, namely Body, Shape, Effort, and Space (BESS). This representation enables the model to capture interpretable pre-crossing behavioral cues, such as posture variation, movement intensity, spatial tendency, and hesitation.

2. A cognitive interaction modeling strategy is proposed, and the CID Module is designed to represent the internal decision logic underlying the formation of pedestrian crossing intention. Unlike conventional feature fusion methods that treat contextual information only as auxiliary input, the CID Module incorporates traffic context, individual attributes, and motion states to characterize pedestrians' trade-offs among rule compliance, risk acceptance, waiting cost, and safety–efficiency relationships during short-term crossing judgment. This allows the model to interpret pedestrian crossing choices from the intention formation process rather than merely from externally observable behavioral outcomes.

3. A unified and interpretable prediction framework, named PsyMIND, is constructed for autonomous-driving decision support. PsyMIND integrates psychological semantic motion features, pedestrian-related contextual features, and cognitive interaction decision features into a unified crossing-intention representation for early pedestrian crossing intention prediction. Experimental results on the Joint Attention in Autonomous Driving (JAAD) dataset show that PsyMIND achieves an accuracy of 0.76, an area under the receiver operating characteristic curve (AUC) of 0.83, and a recall of 0.93, demonstrating its effectiveness in safety-oriented early prediction.

2 Related Work

Pedestrian crossing intention prediction is a fundamental precursor to autonomous driving decision-making, as it aims to anticipate whether a pedestrian is likely to initiate a crossing action from historical behavior and surrounding scene context, thereby providing reliable support for downstream planning and control. In dynamic multi-agent environments, predicted motion intention and trajectory uncertainty have also been incorporated into spatiotemporal risk representations for proactive collision avoidance and downstream path planning (Qiu and Liu 2026). To position this study within the existing literature and to motivate the necessity of a psychology-oriented modeling perspective, this section reviews prior work from three complementary aspects. It first summarizes pedestrian intention prediction methods based on trajectories, skeletal representations, and multimodal fusion. It then discusses the evolution of skeleton-based modeling from geometric configuration description to motion-quality-aware representation. Finally, it examines the decision-level interaction mechanisms underlying pedestrian crossing decisions and reviews relevant modeling strategies for capturing such decision-making processes.

2.1 Overview of Pedestrian Intention Prediction Research

Early studies mainly relied on trajectory modeling, where pedestrians were represented as moving points and future motion or crossing likelihood was inferred from historical trajectories. Saleh et al. (2017, 2018) introduced stacked Long Short-Term Memory (LSTM) networks for pedestrian trajectory modeling and formulated intention prediction as a temporal classification task, while Shen et al. (2018) developed a transferable motion prediction model for intersection scenarios. Later work improved trajectory-based prediction by incorporating visual cues (Landry and Akhloufi 2026), interaction modeling (Yao et al. 2021), and trajectory-conditioned relation learning (Zhou et al. 2022). Beyond pedestrian-specific studies, driving intention has been used as a semantic prior for multi-model trajectory prediction (Wang et al. 2026). This direction further indicates that intention-level representations can complement low-level kinematic observations. However, since these methods mainly rely on positional variation, they may be less effective at early stages when crossing intention has begun to form but observable displacement has not yet occurred.

With the development of pose estimation and graph learning, increasing attention has shifted to skeleton-based methods, which aim to extract finer-grained behavioral cues from the spatial configuration and temporal evolution of human joints. Fang and López (2018) introduced 2D pose estimation for pedestrian intention recognition by exploiting explicit geometric features, such as distances and angles between keypoints. Cadena et al. (2019, 2022) further applied Graph Convolutional Networks (GCNs) to model spatial dependencies among joints and enhanced performance through learnable adjacency matrices and multimodal fusion. Gesnouin et al. (2020, 2021) proposed Skeleton-based Pedestrian Intention Network (SPI-Net) and TrouSPI-Net, which used multi-branch structures and spatiotemporal attention to better characterize skeletal dynamics. Subsequent studies further verified the effectiveness of skeleton information from the perspectives of spatiotemporal graph modeling and traffic-light-controlled

scenarios (Zhang et al. 2022; Ghori et al. 2018; Zhang et al. 2022). Pedestrian Spatial-Temporal Attention Graph Convolutional Network (PedAST-GCN) (Ling et al. 2024) improved prediction efficiency through spatiotemporal attention-based graph convolution, while Li et al. (2024) explored the interpretability of skeleton features. Although skeleton-based methods provide more detailed body-level cues than trajectory-based approaches, most of them still describe pedestrian motion mainly through geometric relations among joints.

More recently, multimodal fusion has become the dominant paradigm. By combining pedestrian pose with traffic signal states, vehicle motion, road geometry, appearance, and other external cues, these methods aim to construct richer contextual representations. Head orientation can provide an additional nonverbal cue for interpreting pedestrian attention and communicative behavior, although reliable estimation must account for large rotations, occlusion, and landmark-detection failures (Algabri et al. 2025). Rasouli et al. (2019) released the Pedestrian Intention Estimation (PIE) dataset and established an important multimodal benchmark by integrating dynamic, contextual, and appearance information. Rasouli et al. (2019) subsequently proposed a stacked gated recurrent unit (GRU) model (SF-GRU), which progressively fused appearance, scene context, pose, trajectory, and vehicle speed. Kotseruba et al. (2021) introduced Pedestrian Crossing Prediction with Attention (PCPA), a unified benchmark that jointly modeled local scenes, pose keypoints, and vehicle speed using 3D convolution, GRU, and attention mechanisms. Rasouli and Kotseruba (2023) further proposed PedFormer, which enhanced behavior prediction through cross-modal attention and gated multitask learning. Later studies, including the dual-branch fusion framework (Yang et al. 2022), Graph-Embedded Transformer with Positional Decoupling (GTransPDM) (Xie et al. 2025), cross-modal Transformer (Chen et al. 2024), Crossing Intention Prediction Network based on Feature Fusion modules (CIPF) (Ham et al. 2023), Attention-Guided Cross-Modal Interaction Transformer (ACIT) (Li and Müller 2025), and the pre-normalized Transformer-based fusion model (Chen et al. 2025), explored increasingly sophisticated cross-modal interaction strategies. Other baseline models considered in the experiments include Deep Virtual-to-Real Distillation for Pedestrian Crossing Prediction (VR-PCP) (Bai et al. 2022), Progressive Interaction Transformer (PIT) (Zhou et al. 2023), Dual-Channel Pedestrian Crossing Intention Anticipation Network (DPCIAN) (Yang et al. 2024), and Gating Syn-to-Real Knowledge for Pedestrian Crossing Prediction (Gated-S2R-PCP) (Bai et al. 2025). Recent work in *Engineering Applications of Artificial Intelligence* further addressed obstruction scenarios by combining graph-based feature reconstruction with multi-scale self-attentive encoding (Liu et al. 2025). Other efforts incorporated graph convolution for visual reasoning (Chen et al. 2021) or leveraged urban contextual information for intention recognition (Yang et al. 2022). However, most multimodal methods still treat contextual cues as additional input features and focus on feature-level fusion, with limited consideration of how pedestrians internally form crossing decisions under traffic constraints.

Overall, trajectory-based, skeleton-based, and multimodal fusion-based methods have advanced pedestrian crossing intention prediction from different perspectives. Trajectory-based methods capture macroscopic movement trends but are limited

in early stages where displacement is weak. Skeleton-based methods provide more detailed body-level cues but often remain at the level of geometric pose representation, with insufficient attention to motion quality and its psychological implications. Multimodal fusion methods incorporate richer environmental context, but they usually organize observable cues through data-driven fusion rather than explicitly modeling the cognitive process through which crossing intention is formed. Therefore, a remaining challenge is how to move beyond observable feature aggregation and construct a unified representation that can jointly describe semantic motion quality, contextual constraints, and cognitive decision logic. This motivates the proposed PsyMIND framework, which introduces psychologically semantic motion representation and cognitive interaction-driven decision modeling to provide a more interpretable and reliable basis for early pedestrian crossing intention prediction.

2.2 From Geometric Configuration to Motion Quality: The Evolution of Skeleton Representation

In skeleton-based representation learning, early studies mainly focused on geometric configuration, where pedestrian body shape and motion tendency were described through the spatial distribution of keypoints, their relative positions, and temporal variations. Reliable pedestrian detection supports skeleton-based motion analysis, while lightweight multi-scale models improve performance in complex scenes (Tang et al. 2024). Cao et al. (2017) proposed Part Affinity Fields, which laid the foundation for multi-person 2D pose estimation in complex scenes. Sun et al. (2019) improved the capture of subtle pose variations through high-resolution representation learning (HRNet), while Xiao et al. (2018) further enhanced the stability of pose estimation and tracking with an efficient baseline framework. In the 3D domain, Zhou et al. (2017) investigated 3D skeleton estimation in real-world scenes, and Lin et al. (2021) further achieved end-to-end human mesh reconstruction using a Transformer-based architecture. These upstream advances transformed pedestrians from ambiguous image regions into structured sets of semantically meaningful keypoints, providing the basis for subsequent skeleton-based behavior modeling.

Once skeleton sequences became available, the central problem shifted to how to extract spatiotemporal dynamics from them. Yan et al. (2018) proposed the Spatio-Temporal Graph Convolutional Network (ST-GCN), a seminal framework that modeled skeleton sequences as spatiotemporal graphs and learned joint dependencies through graph convolution. Shi et al. (2019) improved adjacency learning with adaptive graph convolution, while Li et al. (2019) further explored structured motion patterns. Cheng et al. (2020) combined convolutional neural networks (CNNs) and GCNs to better capture subtle actions, and Si et al. (2019) introduced an attention-enhanced graph convolutional LSTM to model long-range temporal dependencies. More recent hypergraph-based methods have extended pairwise skeletal connections to higher-order relationships among multiple joints and have jointly modeled position, velocity, and bone features for spatiotemporal action representation (Shao et al. 2024). Directional joint variations and temporal dynamics are also important for distinguishing fine-grained actions. Multi-scale central-difference graph convolution has been used to emphasize local directional motion cues while jointly integrating spatial

and temporal skeletal features (Zong et al. 2025). Together, these studies advanced skeleton modeling from static pose description to spatiotemporal dynamic analysis, demonstrating that skeleton motion contains rich behavioral information.

Despite these advances, most existing skeleton-based methods still take geometric change as the primary representation target, focusing on how joints move and how poses evolve, while paying limited attention to the quality with which motion is performed. Another related direction combines semantic body-part parsing with skeleton representations. For example, fine-grained multimodal gait modeling has demonstrated that body-part semantics and skeletal topology provide complementary information for describing human appearance and motion patterns (Xi et al. 2025). However, physical displacement is not equivalent to psychological intention. Similar motion amplitudes, when accompanied by different velocity profiles, force patterns, or spatial expansion tendencies, may reflect entirely different behavioral inclinations. For example, subtle body contraction, forward weight shift, or motion hesitation before stepping off often provides stronger cues to crossing intention than positional variation alone. To address this limitation, this study introduces LMA and develops a PSM Encoder, thereby extending skeleton representation from geometric configuration toward psychologically semantic motion modeling.

2.3 Pedestrian Crossing Decision From a Cognitive Trade-off Perspective: Balancing Safety and Efficiency

Pedestrian crossing decisions are not merely passive responses to external constraints, but rather complex cognitive processes involving risk perception, attention allocation, waiting patience, and subjective utility trade-offs. Cœugnet et al. (2019) pointed out that emotion and social cognition serve as deep psychological drivers of crossing behavior. Gorrini et al. (2019) summarized this process as a dynamic three-stage sequence of approach–appraisal–crossing, highlighting its staged nature. Building on this view, Cambon de Lavalette et al. (2009) showed that pedestrians do not mechanically follow traffic rules, but instead make decisions through a joint evaluation of environmental conditions, task demands, and psychological costs, reflecting an underlying cognitive trade-offs. These studies collectively point to the Theory of Planned Behavior (TPB) as a core theoretical framework for understanding pedestrian decision-making.

Traffic psychology research further provides rich empirical support for this perspective. Evans and Norman (1998) were among the first to apply TPB to pedestrian crossing behavior, showing that crossing intention is significantly influenced by attitude, subjective norm, and perceived behavioral control. Subsequent studies deepened this understanding. Dommes and Cavallo (2011) found that differences in cognitive ability affect how individuals process environmental cues. Osei et al. (2024) demonstrated that risk-compensation psychology dynamically alters decision tendencies. Taken together, these studies suggest that pedestrian crossing decisions arise from a dynamic trade-off between traffic efficiency and collision risk avoidance. As waiting time increases, patience gradually decays, which may lower the threshold for risk acceptance and induce unsafe crossing impulses.

However, these psychological decision mechanisms remain insufficiently incorporated into autonomous-driving-oriented pedestrian intention prediction. Existing

models often treat traffic signals, vehicle motion, pedestrian attention, and contextual attributes as observable inputs, while paying limited attention to how these cues are cognitively evaluated during the formation of crossing intention. This gap motivates the CID Module in PsyMIND, which connects motion-context cues with the latent decision process of pedestrians and supports more interpretable early prediction in complex human–vehicle interaction scenarios.

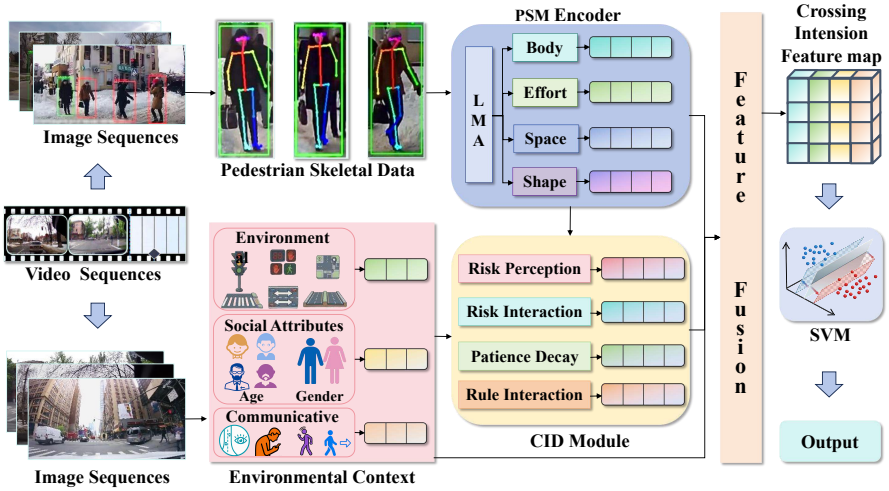


Fig. 1 Architecture of the proposed PsyMIND framework. Given an input video sequence, pedestrian skeletons are first extracted and encoded by the PSM Encoder, which maps raw pose coordinates into psychologically semantic motion features under the Body, Shape, Effort, and Space dimensions based on LMA. In parallel, temporal environmental context features are constructed from environmental constraints, social attributes, and communicative signals. These features are further integrated in the CID Module to model latent cognitive mechanisms, including risk perception, risk interaction, patience decay, and rule interaction. Finally, the psychologically semantic motion features, temporal environmental context features, and cognitive interaction decision features are fused into a unified crossing-intention feature map and fed into a support vector machine (SVM) classifier for final prediction.

3 Methodology

This paper proposes PsyMIND, a hierarchical and progressive framework for pedestrian crossing intention prediction. Built upon an analysis of the psychological mechanisms underlying pedestrian intention formation, the proposed framework aims to establish a closed-loop modeling pathway from fine-grained motion representation, to external contextual constraints, and further to high-level cognitive decision-making. As illustrated in Fig. 1, pedestrian skeletal keypoints are first extracted from the input video sequence, and the resulting skeleton sequence is then fed into the PSM Encoder. By incorporating LMA, this module maps the raw skeletal coordinate system into psychologically semantic motion features along four dimensions, namely Body, Shape, Effort, and Space, which respectively characterize body configuration and posture

variation, motion-form adaptation, dynamic motion quality, and spatial activeness. For example, the forward inclination angle of the torso is used to quantify crossing tendency reflected by center-of-mass transfer, while jerk is employed to characterize hesitation under high-risk interaction scenarios. In this way, low-level geometric displacement is elevated into high-level psychologically semantic motion quality. Meanwhile, the framework constructs temporal environmental context features through dataset-based scene analysis by integrating explicit semantic cues, including traffic signal states, vehicle motion information, pedestrian attributes, and interaction gestures, thereby injecting dynamic contextual constraints and social interaction background into behavior modeling. Subsequently, a CID Module is designed to further couple action semantic features under external contextual constraints and social interaction contexts, so as to explicitly simulate the pedestrian's risk trade-off between safety and crossing efficiency. This module also characterizes latent cognitive mechanisms, including risk perception, risk interaction, patience decay, and rule interaction, thereby enabling deeper inference into the formation process of crossing intention. Finally, the psychologically semantic motion features (PSMF) produced by the PSM Encoder, the temporal environmental context features (TECF), and the cognitive interaction decision features (CIDF) generated by the CID Module are jointly fed into a feature fusion module to construct a unified crossing-intention feature map, which is then passed to an SVM classifier for final intention prediction. This hierarchical modeling paradigm improves early recognition of pedestrian crossing intention and enhances the transparency and interpretability of the model's decision-making process.

3.1 Task Formulation

3.1.1 Problem Definition

Pedestrian crossing intention prediction aims to infer a pedestrian's short-term motion decision by analyzing information within a limited temporal window. To this end, this study formulates the problem as a binary classification task based on time-series features.

Let P_t denote the target pedestrian at time step t in a video sequence. The corresponding intention prediction task is defined as follows: given the historical sequence information of pedestrian P within an observation window of length T_{obs} , i.e., $X = \{x_{t-T_{obs}}, ..., x_t\}$, the goal is to predict the pedestrian crossing intention at future time $t + T_{pred}$, denoted as $Y_{t+T_{pred}} \in \{0, 1\}$, where 0 indicates no crossing intention and 1 indicates crossing intention.

The task parameters are defined as follows:

T_{obs} denotes the observation duration. In this study, it is set to 30 frames (approximately 1 s), which is used to capture the key motion evolution and posture variation prior to crossing.

T_{pred} denotes the prediction horizon. In this study, it is set to 15 frames (approximately 0.5 s), which is used to evaluate the model's capability for early intention recognition before the pedestrian actually initiates a step.

Under this formulation, pedestrian intention prediction can be abstracted as learning an optimal mapping function $f : X \to Y$, so that the model can achieve early and

accurate recognition of the pedestrian's latent decision intention under the guidance of multi-source information.

3.1.2 Temporal Statistical Feature Aggregation

To capture the dynamic evolution of motion, single-frame physical descriptors are further transformed into temporal statistical features. For any continuous physical quantity sequence $S = \{s_{t-T_{obs}}, ..., s_t\}$ within the observation window T_{obs}, such as joint velocity or inter-limb distance, we compute its mean (μ), standard deviation (σ), and current instantaneous value (s_t), yielding $Agg(S) = [\mu_S, \sigma_S, s_t]$. Here, μ_S characterizes the long-term average state, σ_S reflects the fluctuation intensity of the motion, and s_t represents the most recent temporal trend.

3.2 Psychological Semantic Motion Representation Based on LMA

Within the PsyMIND framework, pedestrian skeleton sequences serve as the primary carrier for describing body posture, motion tendency, and pre-crossing behavioral cues. However, raw joint coordinates or simple geometric relations only provide low-level descriptions of how body keypoints are spatially arranged, whereas pedestrian crossing intention is often reflected in more subtle motion qualities, such as forward body inclination, preparatory contraction, hesitation, and changes in movement intensity. LMA is introduced as the theoretical basis for organizing these motion patterns into a more semantically meaningful representation, and the Psychological Semantic Motion Encoder, termed the PSM Encoder, is developed accordingly. Taking raw skeleton sequences as input, this module transforms observable motion into more interpretable semantic representations through a mapping mechanism from physical displacement features to psychologically semantic motion features. As illustrated in Fig. 2, the PSM Encoder first extracts kinematic descriptors from the input skeleton sequence and then organizes them into a BESS-based psychological semantic motion space, thereby establishing a structured transition from low-level motion cues to high-level semantic motion attributes.

Specifically, the PSM Encoder projects raw skeletal motion into a BESS action semantic space composed of four dimensions: Body, Shape, Effort, and Space. Body characterizes body configuration and posture variation, Shape describes morphological adaptation and body expansion patterns, Effort captures the dynamic quality of movement, and Space reflects spatial activeness and directional tendency. This representation describes both the spatial evolution of pedestrian motion and the qualitative manner in which the movement is performed. For example, the forward trunk inclination angle is used to quantify the crossing tendency induced by center-of-mass transfer, while jerk is employed to characterize hesitation under high-risk interaction scenarios. In this way, low-level geometric displacement is elevated into high-level psychologically semantic motion quality, providing a semantically grounded motion basis for subsequent contextual modeling and cognitive decision inference.

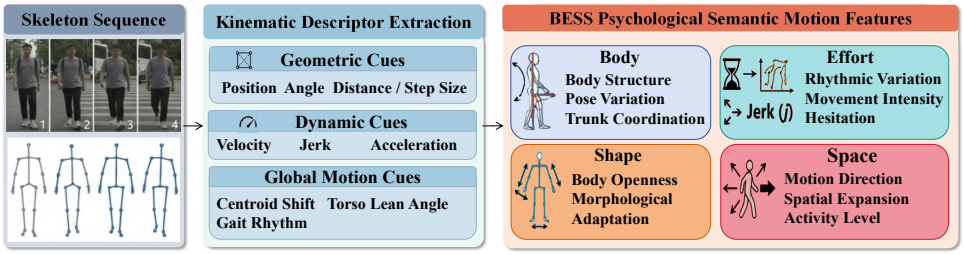


Fig. 2 Structure of the proposed PSM Encoder. Given an input skeleton sequence, the PSM Encoder first extracts kinematic descriptors from three aspects, including geometric cues, dynamic cues, and global motion cues. These low-level motion descriptors are then mapped into a BESS-based psychological semantic motion space composed of Body, Shape, Effort, and Space. Specifically, Body encodes body structure and posture variation, Shape describes body openness and morphological adaptation, Effort captures rhythmic variation, movement intensity, and hesitation-related dynamics, and Space reflects motion direction, spatial expansion, and activity level. Through this design, the PSM Encoder transforms raw skeletal motion into psychologically semantic motion features for subsequent contextual modeling and cognitive decision inference.

Scale variation caused by camera distance and pedestrian body height is reduced through trunk-length normalization. Specifically, all raw coordinates p_{raw} are normalized by the trunk length L of the skeleton to obtain p_{norm}, which provides a robust basis for LMA-based motion mapping. Within this normalized space, three groups of basic motion descriptors are extracted as the core computational components.

Distance measure (D): The normalized Euclidean distance between keypoint pair (i, j) is computed as:

$$D_{i,j}(t) = ||p_i(t) - p_j(t)||_2 \tag{1}$$

Joint angle (A): the angle between limb vectors, or between a limb vector and the vertical axis, is defined as $\theta(t)$.

Motion differential (K): the velocity $v(t)$, acceleration $a(t)$, and jerk $j(t)$ of keypoints are computed using the central difference scheme:

$$v(t) = (p(t+1) - p(t-1))/(2\Delta t), \quad j(t) = (a(t+1) - a(t-1))/(2\Delta t) \tag{2}$$

3.2.1 Body Component

The Body component quantifies the geometric configuration of the human body and the spatial organization of the limbs, thereby reflecting both static posture and dynamic structural variation of the pedestrian. The following feature set is constructed in this dimension.

Gait size (f_1): the distance between the left and right feet. Gait size is a key indicator of motion state and can effectively distinguish a waiting posture with the feet close together from an advancing state in which the pedestrian is stepping forward to cross.

Upper-limb openness (f_2): the distance between the left and right wrists. This feature measures the openness of the upper body. A larger distance may indicate balancing behavior or warning gestures directed toward approaching vehicles.

Nose-to-wrist distances (f_3, f_4): the distances between the nose and the two wrist keypoints. These features are used to capture specific semantic gestures. For example, hand-to-head contact may imply cognitive hesitation or attention distraction, such as when the pedestrian is using a phone.

Wrist-to-ipsilateral-hip distances (f_5, f_6): the distances between each wrist and the hip on the same side. These features describe the natural swing amplitude of the upper limbs. Periodic variation in these distances reflects a normal walking rhythm, whereas persistently short distances are more likely associated with stationary standing.

Limb flexion and compactness (f_7, f_8, f_9, f_{10}): the straight-line chord lengths from wrist to shoulder and from hip to ankle. These features serve as linear proxies for joint angles and quantify the compactness of the limbs. A shorter distance directly corresponds to joint flexion, indicating a loaded or preparatory state, whereas a larger distance corresponds to full limb extension. By monitoring these indicators, the model can capture subtle preparatory contractions of the limbs before step initiation.

Body expansion (f_{11}): the average distance from the four limb endpoints to the pelvis center. This feature characterizes the overall degree of bodily extension and helps distinguish whether the pedestrian is in an open state associated with readiness to move or in a closed state associated with defensive contraction.

3.2.2 Effort Component

The Effort component quantifies the dynamic quality of motion through motion differentials, revealing how pedestrians regulate their internal driving force during movement. This component corresponds to the four effort-related aspects in LMA, namely Time, Space, Weight, and Flow.

Time: the velocity magnitude of the pelvis center (f_{12}) and its acceleration variation (f_{13}). As motion indicators of the body core, these features provide direct cues about Time Effort. Abrupt velocity changes are associated with a sudden movement tendency, whereas smoother velocity profiles correspond to a more sustained state. In addition, the instantaneous velocities of the wrists and ankles ($f_{14}, f_{15}, f_{16}, f_{17}$) are included to capture localized high-frequency actions, such as rapid hand signaling or tentative foot probing before movement, thereby complementing the limitations of core-velocity-only descriptions.

Space: the trajectory turning angle (f_{18}) and head–path consistency (f_{19}). A direct spatial focus is typically characterized by a small turning angle and high visual-path consistency, indicating a clear target-oriented movement tendency. By contrast, an indirect focus is often manifested as a more curved path, suggesting hesitation or exploratory behavior.

Weight: deceleration (f_{20}) is used to quantify braking force and corresponds to Weight Effort. Strong deceleration commonly occurs during abrupt stopping or evasive responses when pedestrians encounter sudden vehicle threats. The accelerations of limb endpoints ($f_{21}, f_{22}, f_{23}, f_{24}$) are further introduced to characterize instantaneous force release. High acceleration values imply a strong weight effort and often appear in decisive stepping or large-amplitude gestures.

Flow: jerk (f_{25}) is adopted as a core indicator of motion smoothness and corresponds to Flow Effort. A high jerk value indicates discontinuous force control, revealing

a more bound and hesitant motion state, especially under high-risk interaction scenarios.

3.2.3 Shape Component

The Shape component captures morphological adjustments and spatial directional changes that arise as the body adapts to the surrounding environment, thereby reflecting how pedestrians perceive and respond to external context.

Forward trunk inclination (f_{26}): the angle between the vector from the neck to the pelvis center and the vertical gravity line. This feature serves as one of the most informative geometric precursors of crossing intention. A marked increase in forward inclination often occurs before actual foot movement, indicating a forward transfer of body mass and an emerging intention to cross.

Lower-limb flexion angles (f_{27}, f_{28}): the bending angles of the knee joints. These features quantify the preparatory loading state of the legs and help distinguish between a rigid standing posture and a flexed posture associated with imminent step initiation.

Upper-limb flexion angles (f_{29}, f_{30}): the bending angles of the elbow joints. These features characterize changes in upper-limb morphology and are useful for identifying specific behaviors such as phone viewing or arm folding.

Limb-to-trunk opening angles ($f_{31}, f_{32}, f_{33}, f_{34}$): the angles between the four limb vectors and the trunk vector. These features describe the directional shape flow of the limbs relative to the body core and reflect the body's extension tendency toward the intended movement direction.

3.2.4 Space Component

The Space component quantifies the geometric relationship between the pedestrian and the surrounding environment, thereby reflecting the pedestrian's spatial occupancy and displacement intention within the scene.

Cumulative displacement within the observation window (f_{35}): this feature measures the pedestrian's spatial occupancy over the observation interval and directly reflects the level of activeness and movement intention in the current scene. A decrease in this quantity often indicates a pause-and-wait state before crossing.

3.3 Cognitive Interaction Decision Feature Modeling

Pedestrian crossing intention is not determined by isolated body movements alone, but is progressively formed through the interaction between action semantics and traffic-context constraints. The same observable motion may convey different decision meanings under different traffic conditions. For example, a hand wave may indicate an ordinary gesture when no vehicle is nearby, but may become a communicative signal or a risk-taking cue when an approaching vehicle is present. Therefore, motion semantics extracted by the PSM Encoder need to be grounded in environmental context before cognitive decision modeling can be performed. In this study, temporal environmental context features are constructed from three aspects: environmental constraints, social attributes, and communicative signals.

Environmental Constraints: Traffic signals not only serve as hard legal constraints, but also act as key triggers of underlying cognitive trade-off, such as urgency near the end of a green phase or violation impulses under a red signal. In this study, the boundary conditions of crossing behavior are defined by incorporating traffic-signal control condition (f_{36}), traffic-signal state (f_{37}), intersection location (f_{38}), designated crossing area (f_{39}), traffic direction (f_{40}), and road type (f_{41}). In addition, vehicle state (f_{42}), such as acceleration or deceleration, is introduced as a source of dynamic pressure. Together, these factors characterize the physical feasibility and situational constraints of pedestrian decision-making.

Social Attributes: Considering that different demographic groups may exhibit different risk preferences, age (f_{43}) and gender (f_{44}) are incorporated as parameterized prior attributes. These factors directly influence pedestrians' risk tolerance and decision thresholds, enabling the model to adapt to population-specific behavioral tendencies.

Communicative Signals: Traffic awareness (f_{45}), head orientation (f_{46}), gesture (f_{47}), and motion direction (f_{48}) are introduced as explicit symbolic communication cues. These features reflect the real-time feedback acquired by pedestrians during interaction and provide important evidence for contextualized intention inference.

After obtaining the temporal environmental context features, a spatiotemporal alignment mechanism is used to couple contextual semantics with the psychologically semantic motion features generated by the PSM Encoder. The resulting action–environment association representation links what the pedestrian is doing with the traffic conditions under which the action occurs. As illustrated in Fig. 3, this representation is fed into the CID Module, where pedestrian crossing intention is modeled as a cognitively grounded trade-off between safety and crossing efficiency. Instead of treating environmental context as auxiliary input, the CID Module regards contextual cues as conditional factors that modulate the decision meaning of motion semantics. Based on this design, the crossing decision process is decomposed into four interrelated cognitive mechanisms: risk perception, risk interaction, patience decay, and rule interaction. These mechanisms transform coupled motion–context cues into cognitive interaction decision features, enabling the model to infer not only whether a pedestrian is likely to cross, but also how such an intention is formed under traffic constraints.

Risk perception feature (f_{49}): this feature describes the pedestrian's attention allocation to dynamic traffic flow. By combining visual attention (Look) and Vehicle Threat, it simulates the process of gap searching. The feature is formulated as:

$$f_{49} = P_{\text{look}}^{\text{veh}} \times T_{\text{veh}} \tag{3}$$

where $P_{\text{look}}^{\text{veh}}$ is encoded from the pedestrian-looking annotation toward approaching vehicles, and T_{veh} denotes the normalized vehicle threat level derived from vehicle motion cues.

Risk interaction (f_{50}): according to acceptable gap theory, pedestrians assess crossing risk based on the available vehicle gap. When a high-threat vehicle is approaching, if the pedestrian still maintains a forward-leaning posture, this often indicates a strong intention to force a crossing. Therefore, by combining the LMA-based forward body

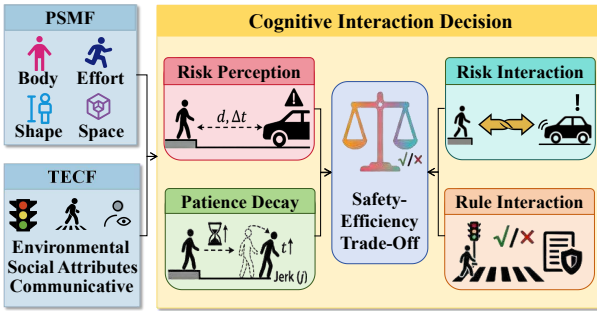


Fig. 3 Illustration of the proposed CID Module. The PSMF and TECF are first aligned to construct an action–environment association representation. Based on this representation, the CID Module explicitly models pedestrian crossing intention as a safety–efficiency trade-off and decomposes the decision process into four interrelated cognitive mechanisms: risk perception, risk interaction, patience decay, and rule interaction. Through this design, the module transforms coupled motion–context cues into cognitive interaction decision features for subsequent crossing intention prediction.

inclination with vehicle threat, this feature quantifies the pedestrian's risk-taking tendency under dynamically hazardous conditions. The corresponding risk interaction strength is defined as:

$$f_{50} = P_{\text{body}} \times T_{\text{veh}} \tag{4}$$

where P_{body} represents the normalized forward trunk inclination extracted from the Shape component, and T_{veh} denotes the normalized vehicle threat level.

Patience decay (f_{51}): based on utility theory, pedestrians continuously balance the potential benefit of crossing against the cost of waiting. As waiting time increases, the threshold for risk acceptance gradually decreases, which may induce non-compliant crossing impulses. To capture this effect, the cumulative waiting duration is combined with the jerk term from the LMA Flow Effort dimension, yielding the following feature:

$$f_{51} = (\Delta t_{\text{wait}} / T_{\max}) \times \int_{t-\tau}^{t} \text{Jerk}(u)\, du \tag{5}$$

where Δt_{wait} denotes the cumulative waiting duration, $T_{\max}$ is the maximum waiting duration used for normalization, τ denotes the temporal integration window, and $\text{Jerk}(u)$ represents the jerk feature extracted from the LMA Flow Effort dimension. Rule interaction (f_{52}): although traffic rules, such as traffic signals, are objective physical constraints, their actual influence on pedestrian behavior depends on whether the pedestrian allocates visual attention to activate them cognitively. This feature explicitly combines attention allocation with environmental constraint through feature crossing, transforming the interaction between traffic awareness and signal status from a neutral observation into a directional decision cue that either suppresses or promotes crossing. In this way, the model can better characterize how pedestrians cognitively process right-of-way rules. The rule internalization strength is defined as:

$$f_{52} = P_{\text{look}}^{\text{sig}} \times S_{\text{light}} \tag{6}$$

where $P_{\text{look}}^{\text{sig}}$ denotes whether the pedestrian attends to the traffic signal, and S_{light} is encoded according to the constraining state of the traffic-light signal. After obtaining the psychologically semantic motion features extracted by the PSM Encoder, the temporal environmental context features, and the cognitive interaction decision features generated by the CID Module, the three feature groups are further fused to construct a crossing intention feature map, which is then fed into an SVM classifier to produce the final pedestrian crossing intention prediction.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset

This study evaluates pedestrian crossing intention prediction on the publicly available Joint Attention in Autonomous Driving (JAAD) dataset released by York University, Canada (Rasouli et al. 2017). Collected from real-world naturalistic traffic environments, JAAD is specifically annotated for pedestrian crossing and non-crossing behaviors, making it well suited for intention prediction tasks. The original JAAD dataset contains 346 video clips. In this study, after applying the task-specific filtering criteria, 81 non-crossing and 234 crossing scenarios were used for model evaluation. In terms of scale, JAAD comprises approximately 88,000 frames, with annotations for 2,624 individual pedestrians and about 390,000 pedestrian bounding boxes. In addition, each bounding box is associated with an occlusion label, including approximately 72,000 partially occluded instances (about 18%) and 46,000 heavily occluded instances (about 11%). Beyond pedestrian annotations, the dataset also provides rich scene-level contextual information, such as traffic signage and road width, which offers valuable support for pedestrian intention prediction.

4.1.2 Data Preprocessing

A complete preprocessing pipeline is designed to extract robust features from raw onboard video data. The pipeline consists of four stages: data cleaning and skeleton extraction, spatial normalization and motion derivatives, temporal sample construction, and feature standardization. These stages reduce environmental noise and inter-subject variability while ensuring the construction of reliable training samples.

(1) Data Cleaning and Skeleton Extraction

A strict filtering strategy is first applied to the raw data to ensure reliable feature extraction. According to the dataset annotations, samples heavily occluded by vehicles or surrounding objects are removed, and only pedestrian sequences labeled as Unoccluded are retained to preserve the completeness of skeleton keypoints. After that, OpenPose (Cao et al. 2017) is used to extract the coordinates of 25 body keypoints (x, y, c) from each frame. To address the reduced recognition accuracy of distant pedestrians under onboard camera views, all pedestrian crops are resized to a fixed height of 256 pixels before being fed into OpenPose. In addition, to guarantee temporal continuity, sequence consistency checking is performed to remove clips with severe

frame loss or failed detections, and only complete skeleton sequences with more than 50 consecutive frames are retained.

(2) Spatial Normalization and Motion Derivatives

Since raw pixel coordinates are strongly affected by imaging distance and pedestrian body size, directly using them would lead to poor generalization. To address this issue, a torso-based normalization strategy is adopted. Specifically, for each frame, the Euclidean distance from the Neck keypoint to the MidHip keypoint is computed as the reference length L_{trunk}. All keypoint coordinates are then divided by L_{trunk}, transforming absolute pixel coordinates into relative dimensionless coordinates. This step reduces scale variation and allows the extracted features to reflect body configuration more consistently.

On this basis, to capture the Effort cues in LMA, first-, second-, and third-order derivatives of the normalized coordinate sequences are computed using the central difference method, yielding velocity, acceleration, and jerk, respectively. In addition, to suppress noise caused by slight body sway during stationary waiting, a motion threshold is introduced. When the pedestrian displacement falls below this threshold, directional features such as trajectory turning angle are forcibly set to zero, thereby reducing noise interference in near-static states.

(3) Sliding-Window Sample Construction

To convert continuous temporal streams into supervised learning sample pairs (X, Y), a sliding-window strategy is adopted. The observation window is set to $T_{obs} = 30$ frames (approximately 1 s) to extract historical behavioral features, and the prediction horizon is set to $T_{pred} = 15$ frames (approximately 0.5 s) to define the prediction target. During feature aggregation, for continuous physical features within each window, such as joint angles and velocity sequences, the mean, standard deviation, and last-frame instantaneous value are computed to characterize the average state, fluctuation level, and latest motion trend, respectively. For discrete semantic features, such as traffic-light status, only the value at the last frame is retained to represent the environmental constraint at the current time step. For label assignment, the pedestrian behavior label (crossing / non-crossing) at time T_{pred} is used as the ground truth of the current sample, so as to train the model for anticipatory intention prediction.

(4) Dataset Splitting and Standardization

Model generalization is evaluated under a Leave-One-Video-Out protocol implemented through Group K-Fold, which also prevents data leakage. Instead of using random shuffling, the data are strictly grouped by *video_id*, ensuring that all pedestrian clips from the same video are assigned entirely to either the training set or the test set. This avoids overfitting caused by memorizing background patterns or pedestrian appearance from the same scene. Finally, considering the substantial scale discrepancy between LMA features (typically within a 0–1 range) and cumulative spatial features (with relatively larger numerical values), all features are normalized using Z-score standardization implemented with StandardScaler before being fed into the SVM classifier. Notably, the standardization parameters $(\mu_{train}, \sigma_{train})$ are computed exclusively from the training set and then applied to the test set, thereby strictly preventing information leakage during preprocessing.

4.1.3 Evaluation Metrics

Model performance is assessed using Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1-score (F1), and the area under the receiver operating characteristic (ROC) curve (AUC). The standard confusion-matrix terms used below are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy: Accuracy measures the proportion of correctly classified samples among all test samples and reflects the overall classification capability of the model:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{7}$$

Precision: Precision denotes the proportion of correctly predicted crossing cases among all samples predicted as crossing. In autonomous driving scenarios, a higher precision indicates fewer false positives, which helps reduce unnecessary deceleration or stopping and thus improves traffic efficiency:

$$\text{Precision} = TP/(TP + FP) \tag{8}$$

Recall: Recall denotes the proportion of correctly identified crossing cases among all samples that actually involve crossing behavior. For safety-critical autonomous driving systems, Recall is particularly important because it directly affects whether potential collision risks can be detected effectively, thereby reducing false negatives:

$$\text{Recall} = TP/(TP + FN) \tag{9}$$

F1-score: The F1-score is the harmonic mean of precision and Recall, and is used to provide a balanced evaluation of model robustness, especially under class-imbalanced conditions:

$$F_1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \tag{10}$$

AUC: AUC measures the model's ability to distinguish between positive and negative samples, i.e., crossing and non-crossing cases, by computing the area under the ROC curve. It provides an objective assessment of the robustness of the model in separating crossing from non-crossing intentions and is relatively insensitive to class imbalance. It is computed by integrating the true positive rate (TPR) over the false positive rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}. \tag{11}$$

where TPR and FPR are defined as

$$TPR = TP/(TP + FN), \qquad FPR = FP/(FP + TN) \tag{12}$$

These terms follow the standard confusion-matrix definitions introduced above.

4.2 Overall Performance Comparison

The overall effectiveness of PsyMIND is examined by comparing it with representative pedestrian intention prediction methods, including TrouSPI-Net (Gesnouin et al.

Table 1 Overall Performance Comparison on the JAAD Dataset

Model	Acc.	AUC	F1	Prec.	Rec.
TrouSPI-Net (Gesnouin et al. 2021)	0.64	0.56	0.76	0.66	0.91
PCPA (Kotseruba et al. 2021)	0.50	0.47	0.59	0.61	0.58
SF-GRU (Rasouli et al. 2019)	0.51	0.45	0.63	0.61	0.64
VR-PCP (Bai et al. 2022)	0.64	0.66	0.76	0.40	0.89
FF-STAtt (Yang et al. 2022)	0.62	0.54	0.74	0.63	0.85
Pedestrian Graph+ (Cadena et al. 2022)	0.70	0.70	0.76	**0.77**	0.75
PIT (Zhou et al. 2023)	0.69	0.67	0.77	0.74	0.83
PedAST-GCN (Ling et al. 2024)	0.69	0.66	0.79	0.68	**0.93**
DPCIAN (Yang et al. 2024)	0.71	0.66	0.78	0.73	0.85
Gated-S2R-PCP (Bai et al. 2025)	0.66	0.68	0.77	0.71	0.85
ACIT (Li and Müller 2025)	0.70	0.62	0.79	0.69	**0.93**
TrajFusionNet (Landry and Akhloufi 2026)	0.74	0.68	**0.82**	0.73	**0.93**
PsyMIND (Ours)	**0.76**	**0.83**	0.78	0.67	**0.93**

2021), PCPA (Kotseruba et al. 2021), SF-GRU (Rasouli et al. 2019), VR-PCP (Bai et al. 2022), and other competitive baselines. The comparison results are summarized in Table 1.

As shown in Table 1, PsyMIND achieves the best Accuracy and AUC on the JAAD dataset, reaching 0.76 and 0.83, respectively, while its Recall reaches 0.93, which is on par with the highest reported value among the compared methods. This comparison shows that PsyMIND provides stronger class discrimination and safety-oriented detection capability, particularly in terms of AUC and Recall. In particular, compared with the strong baseline TrajFusionNet, PsyMIND improves the AUC from 0.68 to 0.83, with an absolute gain of 0.15, suggesting a substantially stronger ability to distinguish crossing from non-crossing intentions in a robust manner.

Furthermore, compared with TrajFusionNet, PsyMIND improves Accuracy from 0.74 to 0.76 while maintaining a high Recall of 0.93. The gain can be attributed to the ability of psychologically semantic motion features to capture subtle pre-crossing motion quality, while the CID features further encode decision-related cues under contextual constraints. As a result, the proposed framework provides a more expressive representation of the pedestrian intention formation process.

It should also be noted that PsyMIND does not achieve the best Precision or F1-score, and its Precision of 0.67 is lower than that of several competing methods. This indicates that the proposed method tends to favor the detection of potentially crossing pedestrians, trading a certain increase in false positives for a higher Recall. For pedestrian intention prediction in autonomous driving, such a safety-oriented tendency is practically meaningful, since a high Recall is critical for reducing the risk of missing hazardous crossing behaviors. Overall, the performance profile of PsyMIND, characterized by the highest Accuracy and AUC and a high Recall, demonstrates its suitability for early pedestrian crossing intention recognition in safety-critical scenarios.

4.3 Early Prediction Capability Analysis

Since early recognition is critical for autonomous driving safety, the anticipatory capability of PsyMIND is further analyzed under different prediction horizons. Specifically, the observation window is fixed at 30 frames, while the prediction horizon is set to 15, 30, 45, and 60 frames, corresponding to 0.5 s, 1.0 s, 1.5 s, and 2.0 s, respectively. The results are presented in Fig. 4.

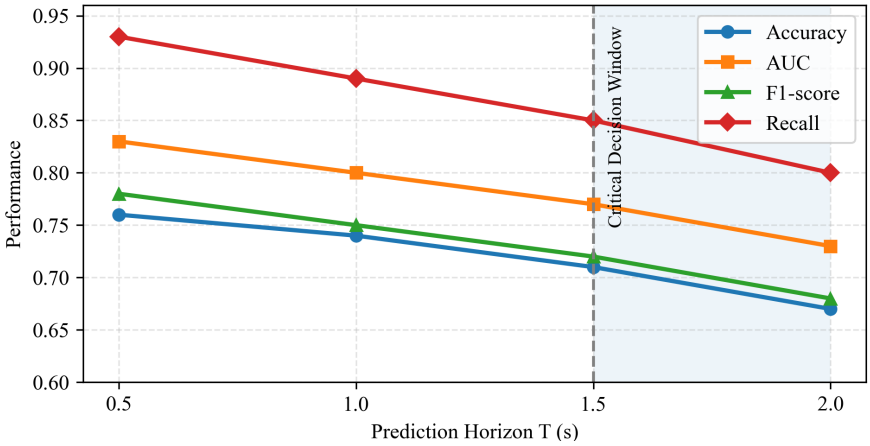


Fig. 4 Early prediction performance of the proposed PsyMIND under different prediction horizons. The observation window is fixed at 30 frames, while the prediction horizon is varied from 15 to 60 frames (0.5 s to 2.0 s). A vertical dashed line is plotted at T=1.5s to mark the Critical Decision Window, showing that the proposed framework maintains reliable prediction performance even when pedestrian intention must be recognized well in advance.

Overall, the performance gradually decreases as the prediction horizon becomes longer, indicating that pedestrian crossing intention prediction becomes more challenging when the model is required to anticipate behavior further in advance. This trend is consistent with the nature of pedestrian decision formation, since observable motion cues are less explicit at earlier stages and the relationship between environmental stimuli and internal decision tendency is more implicit.

Despite this increased difficulty, PsyMIND still maintains strong predictive capability at longer horizons. In particular, at T = 1.5 s, the model preserves relatively high AUC, F1-score, and Recall, indicating that the proposed psychologically semantic motion features and cognitive interaction decision features can capture latent crossing tendencies within the critical decision window before observable displacement occurs.

Furthermore, the relatively smooth degradation of AUC across increasing horizons demonstrates that the proposed framework retains stable class discrimination ability under different anticipation requirements. The consistently high Recall further indicates strong sensitivity to potentially crossing pedestrians, which is particularly desirable in safety-critical autonomous driving scenarios, where missed detections should be minimized. Overall, these results show that PsyMIND achieves a favorable

balance between anticipatory prediction and recognition robustness, highlighting its potential for early risk-aware decision support in autonomous driving.

4.4 Ablation Study

The contribution of each feature group is investigated through an ablation study, where psychologically semantic motion features, temporal environmental context features, and cognitive interaction decision features are progressively combined. The results are reported in Table 2. When only PSMF is used, the model shows relatively weak performance, with Accuracy, AUC, and F1 of 0.47, 0.58, and 0.45, respectively. This limited performance shows that motion semantics alone cannot fully determine crossing intention without contextual grounding. Although PSMF captures motion quality and potential behavioral tendency, it lacks explicit modeling of external traffic constraints and social interaction context, which limits the model's ability to interpret the meaning of pedestrian motion.

Table 2 Ablation Analysis of Different Feature Groups

Feature	Acc.	AUC	F1	Prec.	Rec.
PSMF	0.47	0.58	0.45	0.38	0.60
PSMF+TECF	0.73	0.83	0.66	0.65	0.75
PSMF+CIDF	0.55	0.59	0.48	0.44	0.59
PSMF+TECF+CIDF	0.76	0.83	0.78	0.67	0.93

After introducing TECF, the performance improves markedly, with Accuracy, AUC, and F1 increasing to 0.73, 0.83, and 0.66, respectively. The improvement after introducing TECF confirms the importance of situational grounding for interpreting pedestrian motion semantics.

In contrast, the combination of PSMF and CIDF yields only limited gains, achieving Accuracy, AUC, and F1 of 0.55, 0.59, and 0.48, respectively. The limited gain of PSMF and CIDF implies that cognitive decision features require sufficient contextual support to become discriminative. Without environmental context, the high-level cognitive mechanisms related to risk, rules, and interaction cannot be effectively grounded, and the resulting discriminative benefit remains limited.

When all three feature groups are jointly integrated, the model achieves competitive overall performance. Specifically, Accuracy reaches 0.76, F1 reaches 0.78, and Precision and Recall reach 0.67 and 0.93, respectively, while AUC remains at the highest value of 0.83. In particular, compared with PSMF+TECF, the introduction of CIDF further improves Recall from 0.75 to 0.93 and F1 from 0.66 to 0.78. The improvement in Recall and F1 further confirms that CIDF contributes mainly to identifying potentially crossing pedestrians when supported by both motion semantics and environmental context.

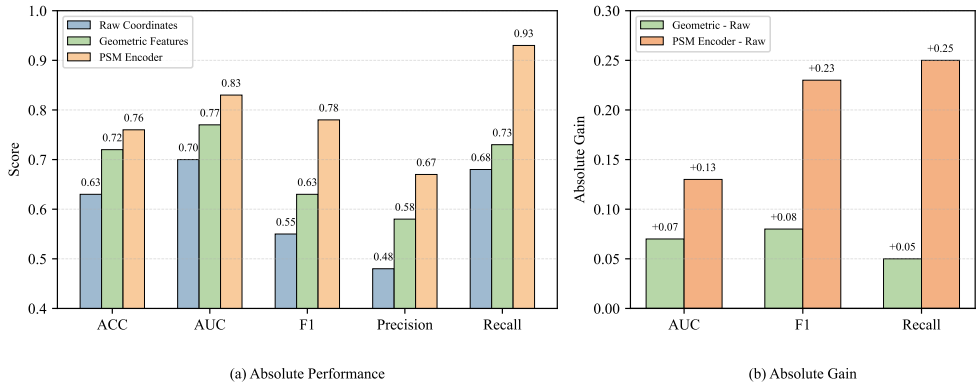


Fig. 5 Comparison of different skeleton representation methods. (a) Absolute performance of Raw Coordinates, Geometric Features, and the proposed PSM Encoder. (b) Absolute gains between adjacent representation levels. The proposed PSM Encoder achieves the strongest improvements, especially in F1 and Recall.

Overall, TECF provides the necessary situational grounding for action semantics, while CIDF further models the cognitive trade-off between safety and crossing efficiency on this basis. The three feature groups are therefore highly complementary and together form a more complete representation of pedestrian crossing intention.

4.5 Comparison of Skeleton Representation Methods

Since the PSM Encoder is designed to move beyond low-level geometric descriptions, three skeleton representation schemes with different abstraction levels are compared, including Raw Coordinates, Geometric Features, and the proposed PSM Encoder. As shown in Fig. 5(a), the performance improves consistently as the representation evolves from low-level coordinates to explicit geometric modeling and further to the proposed psychologically semantic motion representation. When only Raw Coordinates are used, the model achieves relatively limited performance, with Accuracy = 0.63, AUC = 0.70, and F1 = 0.55, indicating that raw keypoint coordinates alone are insufficient for reliable pedestrian intention recognition in complex traffic scenarios. After replacing them with Geometric Features, the corresponding values increase to 0.72, 0.77, and 0.63, respectively, suggesting that explicit modeling of joint distances, limb angles, and local geometric relations is more effective in capturing posture variation and motion tendency. On this basis, the proposed PSM Encoder achieves the best overall performance, with Accuracy = 0.76, AUC = 0.83, F1 = 0.78, Precision = 0.67, and Recall = 0.93. In particular, as illustrated in Fig. 5(b), the performance gains from Geometric Features to PSM Encoder are especially pronounced in F1 and Recall, with absolute improvements of 0.15 and 0.20, respectively, while AUC is further improved by 0.06. Overall, the progressive improvement from raw coordinates to geometric descriptors and further to the PSM Encoder demonstrates that psychologically semantic motion representation provides more discriminative and behaviorally meaningful features for early pedestrian crossing intention prediction.

4.6 Classifier Comparison

Finally, the influence of classifier choice is examined by comparing SVM, multilayer perceptron (MLP), residual network (ResNet), and one-dimensional convolutional neural network (1D-CNN) under the same input features, data split, and evaluation protocol. As shown in Fig. 6(a), SVM achieves the best absolute performance across all five evaluation metrics, with Accuracy = 0.76, AUC = 0.83, F1 = 0.78, Precision = 0.67, and Recall = 0.93. In particular, its AUC is higher than that of the best-performing deep baseline, 1D-CNN (0.80), indicating that the proposed feature representation itself already provides strong class discrimination, while SVM is able to exploit this representation more effectively for decision boundary learning. The relative advantage of SVM is further illustrated in Fig. 6(b), where the classifiers are distributed in the AUC–Recall space. As shown in Fig. 6(b), SVM occupies the upper-right region of the AUC–Recall space, reflecting a favorable balance between class discrimination and safety-oriented detection. Although 1D-CNN achieves relatively competitive AUC among the deep baselines, its Recall remains clearly lower than that of SVM. MLP and ResNet perform less favorably overall, indicating that generic deep classifiers do not provide a clear advantage under the current data scale and feature form.

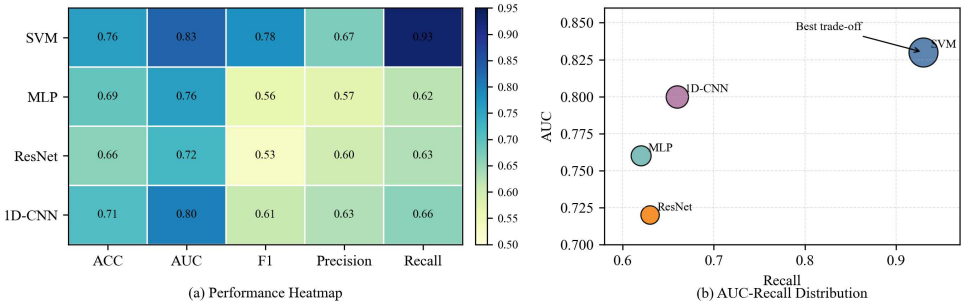


Fig. 6 Comparison of different classifiers. (a) Heatmap visualization of the absolute performance of SVM, MLP, ResNet, and 1D-CNN in terms of Accuracy, AUC, F1, Precision, and Recall. (b) Distribution of different classifiers in the AUC–Recall space, where marker size is proportional to F1-score. The SVM-based solution achieves the most favorable trade-off, indicating stronger class discrimination and better safety-oriented detection capability.

Overall, the experimental results demonstrate that PsyMIND provides an effective and interpretable solution for early pedestrian crossing intention prediction. The overall performance comparison confirms its strong class discrimination and safety-oriented detection capability, especially in terms of AUC and Recall. The early prediction analysis shows that PsyMIND can identify latent crossing tendencies before explicit displacement becomes available. Furthermore, the ablation studies and representation comparisons verify that psychology-informed semantic motion features, pedestrian-related contextual cues, and cognitive interaction decision features make complementary contributions to both prediction performance and interpretability.

5 Conclusion

This study addressed early pedestrian crossing intention prediction from an explainable decision-support perspective for autonomous driving. Rather than treating crossing intention as a label inferred only from observable motion patterns, PsyMIND represents it through the joint modeling of semantic motion quality, pedestrian-related context, and cognitive interaction decision cues. The Psychological Semantic Motion Encoder transforms skeleton dynamics into interpretable motion semantics along the Body, Shape, Effort, and Space dimensions, enabling subtle pre-crossing cues such as posture adjustment, movement intensity, spatial tendency, and hesitation to be represented explicitly. The Cognitive Interaction Decision Module further relates traffic constraints, individual attributes, and motion states to rule compliance, risk acceptance, waiting cost, and safety–efficiency trade-offs. Experiments on the JAAD dataset show that PsyMIND achieves an accuracy of 0.76, an AUC of 0.83, and a recall of 0.93, demonstrating its effectiveness for safety-oriented early prediction.

Beyond the numerical results, the main contribution of this work lies in shifting pedestrian crossing intention prediction from observable feature aggregation toward interpretable intention-formation modeling. The findings suggest that early crossing intention cannot be sufficiently explained by motion displacement or multimodal feature fusion alone; it also depends on how pedestrians perceive traffic situations, regulate motion, and make short-term decisions under contextual constraints. By introducing psychology-informed motion semantics and cognitive interaction modeling, PsyMIND provides a structured way to connect low-level skeletal dynamics with higher-level decision cues. This design helps explain why a pedestrian may cross, wait, or hesitate under similar visual appearances, and it offers a more transparent basis for autonomous-driving decision support. From an engineering application perspective, such interpretable early prediction is meaningful for reducing missed detections of potential crossing pedestrians and for improving public trust in automated vehicles operating in complex urban environments.

Several limitations remain. First, the current evaluation is mainly conducted on the JAAD dataset, and the cross-dataset generalization of PsyMIND still requires further validation. Second, the framework relies on skeleton extraction and context annotations, which may be affected by occlusion, long-distance observation, adverse weather, and night-time scenes. Third, although the cognitive interaction features improve interpretability, they remain a simplified computational approximation of pedestrian decision-making rather than a complete psychological model. Future work will extend PsyMIND to more diverse naturalistic traffic datasets, improve robustness under challenging perception conditions, and explore finer-grained social interaction and individual psychological factors for real-world autonomous-driving deployment.

References

Algabri, R., Shin, H., Abdu, A., Bae, J.-H., Lee, S.: WQuatNet: Wide range quaternion-based head pose estimation. Journal of King Saud University Computer and Information Sciences **37**, 24 (2025) https://doi.org/10.1007/s44443-025-00034-1

Bai, J., Fang, X., Fang, J., Xue, J., Yuan, C.: Deep virtual-to-real distillation for pedestrian crossing prediction. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 1586–1592 (2022). https://doi.org/10.1109/ITSC55140.2022.9921771

Bai, J., Fang, J., Lv, Y., Lv, C., Xue, J., Li, Z.: Gating syn-to-real knowledge for pedestrian crossing prediction in safe driving. IEEE Transactions on Intelligent Transportation Systems **26**(6), 7509–7522 (2025) https://doi.org/10.1109/TITS.2025.3554767

Cœugnet, S., Cahour, B., Kraiem, S.: Risk-taking, emotions and socio-cognitive dynamics of pedestrian street-crossing decision-making in the city. Transportation Research Part F: Traffic Psychology and Behaviour **65**, 141–157 (2019) https://doi.org/10.1016/j.trf.2019.07.011

Chen, Z., Liu, G., Ye, Z.: Multi-modal pedestrian crossing intention prediction model based on pre-normalization Transformer. Application Research of Computers **42**(5), 1378–1384 (2025) https://doi.org/10.19734/j.issn.1001-3695.2024.10.0428

Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Pedestrian graph +: A fast pedestrian crossing prediction model based on graph convolutional networks. IEEE Transactions on Intelligent Transportation Systems **23**(11), 21050–21061 (2022) https://doi.org/10.1109/TITS.2022.3173537

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1302–1310 (2017). https://doi.org/10.1109/CVPR.2017.143

Chen, T., Tian, R., Ding, Z.: Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3096–3102 (2021). https://doi.org/10.1109/ICCVW54120.2021.00345

Cambon de Lavalette, B., Tijus, C., Poitrenaud, S., Leproux, C., Bergeron, J., Thouez, J.-P.: Pedestrian crossing decision-making: A situational and behavioral approach. Safety Science **47**(9), 1248–1253 (2009) https://doi.org/10.1016/j.ssci.2009.03.016

Chen, X., Xu, W., Zhang, S., Cai, Y.: Pedestrian crossing intention prediction via progressive multimodal token fusion for autonomous driving. IEEE Transactions on Intelligent Transportation Systems **26**(9), 12959–12973 (2025) https://doi.org/10.1109/TITS.2025.3567892

Cadena, P.R.G., Yang, M., Qian, Y., Wang, C.: Pedestrian graph: Pedestrian crossing prediction based on 2D pose estimation and graph convolutional networks. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2000–2005 (2019). https://doi.org/10.1109/ITSC.2019.8917118

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 180–189 (2020). https://doi.org/10.1109/CVPR42600.2020.00026

Chen, X., Zhang, S., Li, J., Yang, J.: Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving. IEEE Transactions on Intelligent Transportation Systems **25**(9), 12538–12549 (2024) https://doi.org/10.1109/TITS.2024.3386689

Dommes, A., Cavallo, V.: The role of perceptual, cognitive, and motor abilities in street-crossing decisions of young and older pedestrians. Ophthalmic and Physiological Optics **31**(3), 292–301 (2011) https://doi.org/10.1111/j.1475-1313.2011.00835.x

Delavari, Z., Pflanzner, T., Kertesz, A.: A taxonomy and survey of integrating emerging technologies to intelligent transportation systems. Journal of King Saud University Computer and Information Sciences **38**, 411 (2026) https://doi.org/10.1007/s44443-026-00806-3

Evans, D., Norman, P.: Understanding pedestrians' road crossing decisions: an application of the theory of planned behaviour. Health Education Research **13**(4), 481–489 (1998) https://doi.org/10.1093/her/13.4.481

Fang, Z., López, A.M.: Is the pedestrian going to cross? answering by 2d pose estimation. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1271–1276 (2018). https://doi.org/10.1109/IVS.2018.8500413 . IEEE

Gorrini, A., Crociani, L., Vizzari, G., Bandini, S.: Crossing behaviour of social groups: Insights from observations at non-signalised intersection. In: Hamdar, S.H. (ed.) Traffic and Granular Flow '17, pp. 443–450. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11440-4_48

Ghori, O., Mackowiak, R., Bautista, M., Beuter, N., Drumond, L., Diego, F., Ommer, B.: Learning to forecast pedestrian intention from pose dynamics. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1277–1284 (2018). https://doi.org/10.1109/IVS.2018.8500657

Gesnouin, J., Pechberti, S., Bresson, G., Stanciulescu, B., Moutarde, F.: Predicting intentions of pedestrians from 2D skeletal pose sequences with a representation-focused multi-branch deep learning network. Algorithms **13**(12), 331 (2020) https://doi.org/10.3390/a13120331

Gesnouin, J., Pechberti, S., Stanciulescu, B., Moutarde, F.: TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–7 (2021). https://doi.

org/10.1109/FG52635.2021.9666989 . IEEE

Ham, J.-S., Kim, D.H., Jung, N., Moon, J.: CIPF: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3666–3675 (2023). https://doi.org/10.1109/CVPRW59228.2023.00374

Keller, C.G., Hermes, C., Gavrila, D.M.: Will the pedestrian cross? probabilistic path prediction based on learned motion features. In: Joint Pattern Recognition Symposium, pp. 386–395 (2011). https://doi.org/10.1007/978-3-642-23123-0_39 . Springer

Kotseruba, I., Rasouli, A., Tsotsos, J.K.: Benchmark for evaluating pedestrian action prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1258–1268 (2021). https://doi.org/10.1109/WACV48630.2021.00130

Kooij, J.F.P., Schneider, N., Flohr, F., Gavrila, D.M.: Context-based pedestrian path prediction. In: European Conference on Computer Vision, pp. 618–633 (2014). https://doi.org/10.1007/978-3-319-10599-4_40 . Springer

Landry, F.G., Akhloufi, M.A.: TrajFusionNet: Pedestrian crossing intention prediction via fusion of sequential and visual trajectory representations. Computer Modeling in Engineering & Sciences (2026) https://doi.org/10.32604/cmes.2026.086105 . Published online 1 September 2026

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3590–3598 (2019). https://doi.org/10.1109/CVPR.2019.00371

Li, H., Jin, Y., Ren, G.: Interpretable prediction of pedestrian crossing intention: Fusion of human skeletal information in natural driving scenarios. IEEE Transactions on Intelligent Transportation Systems **25**(11), 18153–18170 (2024) https://doi.org/10.1109/TITS.2024.3435991

Li, Y., Müller, S.: ACIT: Attention-guided cross-modal interaction transformer for pedestrian crossing intention prediction. In: 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC), pp. 2018–2024 (2025). https://doi.org/10.1109/ITSC60802.2025.11423621

Ling, Y., Ma, Z., Zhang, Q., Xie, B., Weng, X.: PedAST-GCN: Fast pedestrian crossing intention prediction using spatial-temporal attention graph convolution networks. IEEE Transactions on Intelligent Transportation Systems **25**(10), 13277–13290 (2024) https://doi.org/10.1109/TITS.2024.3398252

Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1954–1963 (2021). https://doi.org/10.1109/CVPR46437.2021.00199

Liu, Z., Xue, Y., Zhu, L.: Pedestrian crossing intention prediction in obstructed scenarios using graph-based feature reconstruction and multi-scale self-attentive encoding. Engineering Applications of Artificial Intelligence **159**, 111748 (2025) https://doi.org/10.1016/j.engappai.2025.111748

Osei, K.K., Obiri-Yeboah, A.A., Adu-Gyamfi, L., Ackaah, W.: Road crossing behavior and preferences among pedestrians: From the lens of the theory of interpersonal behavior. Traffic Injury Prevention **25**(1), 91–100 (2024) https://doi.org/10.1080/15389588.2023.2271107

Qiu, Y., Liu, Y.: Intent-aware spatiotemporal risk field for multi-AGV path planning with heterogeneous dynamic obstacles. Journal of King Saud University Computer and Information Sciences **38**, 570 (2026) https://doi.org/10.1007/s44443-026-00997-9

Rasouli, A., Kotseruba, I.: PedFormer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 9844–9851 (2023). https://doi.org/10.1109/ICRA48891.2023.10161318

Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: 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 (ICCV), pp. 6262–6271 (2019). https://doi.org/10.1109/ICCV.2019.00636

Rasouli, A., Kotseruba, I., Tsotsos, J.K.: 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 (ICCVW), pp. 206–213 (2017). https://doi.org/10.1109/ICCVW.2017.33

Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Pedestrian action anticipation using contextual feature fusion in stacked RNNs. In: Proceedings of the British Machine Vision Conference, pp. 49–14913. BMVA Press, ??? (2019). https://doi.org/10.5244/C.33.49

Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1227–1236 (2019). https://doi.org/10.1109/CVPR.2019.00132

Shen, M., Habibi, G., How, J.P.: Transferable pedestrian motion prediction models at intersections. In: 2018 IEEE/RSJ International Conference on Intelligent Robots

and Systems (IROS), pp. 4547–4553 (2018). https://doi.org/10.1109/IROS.2018.8593783

Saleh, K., Hossny, M., Nahavandi, S.: Intent prediction of vulnerable road users from motion trajectories using stacked LSTM network. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 327–332 (2017). https://doi.org/10.1109/ITSC.2017.8317941

Saleh, K., Hossny, M., Nahavandi, S.: Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks. IEEE Transactions on Intelligent Vehicles **3**(4), 414–424 (2018) https://doi.org/10.1109/TIV.2018.2873901

Shao, Y., Mao, L., Ye, L., Li, J., Yang, P., Ji, C., Wu, Z.: H^2GCN: A hybrid hypergraph convolution network for skeleton-based action recognition. Journal of King Saud University Computer and Information Sciences **36**, 102072 (2024) https://doi.org/10.1016/j.jksuci.2024.102072

Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5686–5696 (2019). https://doi.org/10.1109/CVPR.2019.00584

Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026–12035 (2019). https://doi.org/10.1109/CVPR.2019.01230

Tang, J., Lai, H., Gao, G., Wang, T.: PFEL-Net: A lightweight network to enhance feature for multi-scale pedestrian detection. Journal of King Saud University Computer and Information Sciences **36**(8), 102198 (2024) https://doi.org/10.1016/j.jksuci.2024.102198

Wang, J., Guo, F., Hao, J., Hasan, Z.H.A.: Intention-fused multi-model trajectory prediction for lane-changing. Journal of King Saud University Computer and Information Sciences **38**, 497 (2026) https://doi.org/10.1007/s44443-026-00810-7

Xie, C., Lin, C., Zheng, X., Gong, B., López, A.M.: GTransPDM: A graph-embedded transformer with positional decoupling for pedestrian crossing intention prediction. IEEE Signal Processing Letters **32**, 2109–2113 (2025) https://doi.org/10.1109/LSP.2025.3567249

Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Computer Vision – ECCV 2018, pp. 466–481. Springer, ??? (2018). https://doi.org/10.1007/978-3-030-01231-1_29

Xi, H., Xi, Y., Hu, C., Yahyapour, R.: FMCBGait: Fine-grained multimodal gait recognition with cues embedding and body parts distribution guidance. Journal of

King Saud University Computer and Information Sciences **37**, 311 (2025) https://doi.org/10.1007/s44443-025-00324-8

Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X.: Coupling intent and action for pedestrian crossing behavior prediction. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 1238–1244 (2021). https://doi.org/10.24963/ijcai.2021/171

Yang, B., Wei, Z., Hu, H., Wang, R., Yang, C., Ni, R.: DPCIAN: A novel dual-channel pedestrian crossing intention anticipation network. IEEE Transactions on Intelligent Transportation Systems **25**(6), 6023–6034 (2024) https://doi.org/10.1109/TITS.2023.3333328

Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1), 7444–7452 (2018) https://doi.org/10.1609/aaai.v32i1.12328

Yang, B., Zhan, W., Wang, P., Chan, C., Cai, Y., Wang, N.: Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment. IEEE Transactions on Intelligent Transportation Systems **23**(6), 5338–5349 (2022) https://doi.org/10.1109/TITS.2021.3053031

Yang, D., Zhang, H., Yurtsever, E., Redmill, K.A., Özgüner, Ü.: Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. IEEE Transactions on Intelligent Vehicles **7**(2), 221–230 (2022) https://doi.org/10.1109/TIV.2022.3162719

Zhang, S., Abdel-Aty, M., Wu, Y., Zheng, O.: Pedestrian crossing intention prediction at red-light using pose estimation. IEEE Transactions on Intelligent Transportation Systems **23**(3), 2331–2339 (2022) https://doi.org/10.1109/TITS.2021.3074829

Zhang, X., Angeloudis, P., Demiris, Y.: ST CrossingPose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. IEEE Transactions on Intelligent Transportation Systems **23**(11), 20773–20782 (2022) https://doi.org/10.1109/TITS.2022.3177367

Zhou, C., AlRegib, G., Parchami, A., Singh, K.: Learning trajectory-conditioned relations to predict pedestrian crossing behavior. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 4088–4092 (2022). https://doi.org/10.1109/ICIP46576.2022.9897655

Zong, X., Dong, S., Guo, J.: MSCDGCN: Multi-scale central difference graph convolution network for skeleton-based action recognition. Journal of King Saud University Computer and Information Sciences **37**, 313 (2025) https://doi.org/10.1007/s44443-025-00275-0

Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation

in the wild: A weakly-supervised approach. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 398–407 (2017). https://doi.org/10.1109/ICCV.2017.53

Zhou, Y., Tan, G., Zhong, R., Li, Y., Gou, C.: PIT: Progressive interaction transformer for pedestrian crossing intention prediction. IEEE Transactions on Intelligent Transportation Systems **24**(12), 14213–14225 (2023) https://doi.org/10.1109/TITS.2023.3309309

Statements and Declarations

- Funding
  Funding details have been withheld for double-anonymous review.
- Competing interests
  The authors have no relevant financial or non-financial interests to disclose.
- Data availability
  The data analyzed in this study are publicly available in the Joint Attention in Autonomous Driving (JAAD) dataset. The video data can be accessed from the official JAAD dataset website at https://data.nvision2.eecs.yorku.ca/JAAD_dataset/, and the annotations and data-processing interface are available from the official repository at https://github.com/ykotseruba/JAAD. The processed data generated during the current study are available from the corresponding author upon reasonable request. No new primary data were collected for this study.
- Author contribution
  Author contribution details have been withheld for double-anonymous review.