TP-DID: Temporal Prediction of Driver Interpretation Distributions

Hao Tang[1,2], Yu Guo[1,2], Chaoyi Yang[2], Yilin Li[2], Chang Leng[2], Ying Qiao[2,(✉)], Hongan Wang[2]

[1]University of Chinese Academy of Sciences
[2]Institute of Software, Chinese Academy of Sciences

✉ Corresponding author: qiaoying@iscas.ac.cn

Abstract

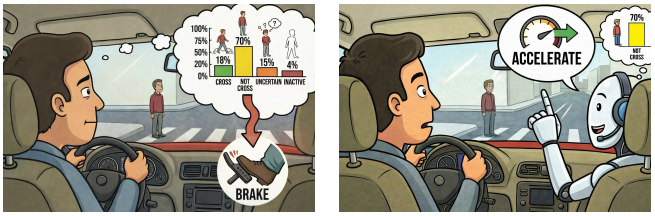
Driving assistance systems require both system-side estimation of a pedestrian's crossing intent and driver-side modeling of how that intent is interpreted. Existing pedestrian intent prediction methods often rely on one-hot labels or binary crossing probabilities, which cannot capture observer variation or intermediate interpretation states. We introduce *Temporal Prediction of Driver-side Interpretation Distributions* (TP-DID), a causal frame-level task for predicting empirical distributions of driver-side interpretations towards pedestrian intent. To capture temporal distribution changes, we propose Input-Modulated Mamba (IM-Mamba), a causal state-space model that reweights current visual evidence before temporal state prediction. Experiments on PSI show that IM-Mamba improves distribution prediction, especially for observer variation, awareness-related states, and temporal changes.

Keywords: Driver-side interpretation, pedestrian intent, empirical interpretation distribution, causal temporal prediction

Introduction

Vehicle-Pedestrian interaction is a safety-critical problem in driving assistance systems (WorldHealthOrganization, 2023). To provide timely and accurate warnings or interventions for the driver, an intelligent driving assistance system must anticipate a pedestrian's crossing intent before the interaction becomes critical. However, crossing intent is not directly observable and must be inferred from ego-view pedestrian states. Thus, pedestrian crossing-intent prediction is a fundamental perception task for driving assistance systems.

Existing methods mainly perform system-side crossing-intent estimation from ego-view observations (Azarmi et al., 2025; Qu et al., 2025; Sharma et al., 2022). Although useful for behavior anticipation, system-side estimation alone is insufficient for human-centered driver-in-the-loop assistance. Unlike an isolated autonomous planner, an assistance system should coordinate its interpretation with the driver's online interpretation of the same pedestrian. The same system-side estimate may require different assistance depending on whether the driver has noticed the pedestrian, remains uncertain, or has formed a cautious judgment. An alert may be redundant when the driver has recognized the risk, but necessary when the driver is unaware of the pedestrian. Ignoring driver-side interpretation can therefore cause redundant or conflicting assistance, weakening human-machine coordination. Since driver recognition of pedestrian crossing intent is important in pedestrian interaction (Chen et al., 2019), and recognition-level assistance can improve operator intervention and vehicle stability (Kuribayashi et al., 2023), this paper studies driver-side interpretation prediction towards pedestrian crossing intent.



(a) Driver Interpretation distribution over pedestrian.

(b) Conflict between human driver and assistance system

Figure 1: Driver Interpretation and Human–Machine Coordination in Ambiguous Crosswalk Scenarios

However, existing label representations in pedestrian intent prediction are inadequate for driver-side interpretation prediction. One-hot labels collapse observer responses into a single category and discard inter-observer variation, although such variation may reflect scene ambiguity and perception difficulty rather than annotation noise (Zhang et al., 2023). Binary crossing probabilities are softer, but still reduce interpretation to the crossing-versus-not-crossing formulation (Landry & Akhloufi, 2025; Moreno et al., 2023; Yao et al., 2021). They cannot distinguish a confident non-crossing judgment from an unformed judgment, nor represent the case where the observer has not noticed the target pedestrian. Thus, driver-side interpretation should be modeled as a distribution over multiple states rather than as a one-hot label or a binary probability.

We formulate *Temporal Prediction of Driver Interpretation Distributions* (TP-DID), a causal frame-level distribution prediction task. Given ego-view observations up to the current frame, the model predicts the current *Empirical Interpretation Distribution* over four states: *Cross*, *Not Cross*, *Not Sure*, and *Not Aware*. *Cross* and *Not Cross* denote formed judgments about pedestrian crossing intent, whereas *Not Sure* and *Not Aware* denote an unformed judgment and lack of awareness of the target pedestrian, respectively. We instantiate TP-DID on PSI (Jing et al., 2025). Because PSI annotations are collected from offline observers rather than instrumented in-vehicle drivers, we use the empirical observer distribution as a measurable proxy for visually induced driver-side interpretation in normal ego-view pedestrian scenes.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

TP-DID requires causal temporal modeling of driver-side interpretation distributions. Driver-side interpretation towards pedestrian intent depends on temporally evolving pedestrian states, such as position, velocity, and acceleration, which provide compact but informative cues for crossing-intent estimation (Rasouli et al., 2017; Schmidt & Faerber, 2009; Yao et al., 2021). Unlike continuous physical states, however, human interpretation is observed through discrete frame-level responses. Thus, the Empirical Interpretation Distribution may evolve smoothly when pedestrian motion changes gradually, but may shift abruptly when newly observed motion evidence changes observer judgments. A suitable model should therefore preserve historical context while remaining sensitive to current-frame evidence.

To this end, we propose Input-Modulated Mamba (IM-Mamba), a causal state-space model for TP-DID. IM-Mamba modulates current visual evidence before temporal state propagation, emphasizing newly emerging intent-related cues while preserving efficient causal sequence modeling. This design enables IM-Mamba to capture both temporal continuity and abrupt transitions in driver-side interpretation distributions.

Our contributions are:

1. We formulate TP-DID, a causal frame-level task for predicting driver-side interpretation towards pedestrian intent as an Empirical Interpretation Distribution.
2. We propose IM-Mamba, an input-modulated state-space model for temporal Empirical Interpretation Distribution prediction.
3. We evaluate IM-Mamba on PSI and show improved prediction of observer variation, awareness-related states, and temporal distribution changes.

Task Description

Dataset

We use the PSI dataset (Jing et al., 2025), which provides causal, frame-level observer interpretations of pedestrian crossing intent from ego-view videos. Each timestamp is annotated as an empirical distribution y_t over four categories: *Cross*, *Not Cross*, *Not Sure*, and *Not Aware* (Figure 2). These labels capture temporal observer variation and support driver-side interpretation distribution prediction beyond binary crossing intent.

Existing benchmarks are less suitable for this formulation. JAAD (Kotseruba et al., 2016) provides deterministic labels for future pedestrian actions, while PIE (Rasouli et al., 2019) mainly represents uncertainty as a binary crossing probability. Both settings collapse intermediate and awareness-related interpretation categories, such as *Not Sure* and *Not Aware*, that are important for human-centered assistance.

Limitations and Scope. PSI annotations are collected from offline observers rather than instrumented in-vehicle drivers, and therefore do not fully capture cognitive load, control responsibility, or multi-tasking in active driving. We treat the PSI empirical observer distribution as a structured offline proxy for visually induced driver-side interpretation, rather than a direct measurement of in-vehicle driver interpretation.

Problem Formulation

We define *Temporal Prediction of Driver Interpretation Distributions* (TP-DID) as a causal streaming inference problem. At each time step $t \in \{1, \dots, T\}$, the model predicts the observer population's interpretation distribution over pedestrian crossing intent using only currently available observations.

Streaming Input/Output. At each time step t, the model receives perception outputs for the target pedestrian, represented as a geometric feature vector $x_t \in \mathbb{R}^{d_x}$. This vector comprises the tracked 2D bounding box and its temporal derivatives, including geometry, center position, area as a distance proxy, velocity, and acceleration. While human interpretation fundamentally relies on rich social and environmental contexts, we intentionally restrict x_t to these low-dimensional kinematic features to isolate spatio-temporal reasoning and mitigate spurious background correlations.

Let

$$X_{1:t} = \{x_1, x_2, \dots, x_t\} \tag{1}$$

denote the causal input prefix. Under strict causality, the prediction depends exclusively on $X_{1:t}$. The model outputs a categorical interpretation distribution over m intent hypotheses,

$$\hat{y}_t = \mathcal{M}(X_{1:t}) \in \Delta^{m-1}, \tag{2}$$

which acts as the best-available estimate of the observer population's interpretation state.

Empirical Interpretation Distribution. At each time step t, the target is an empirical interpretation distribution $y_t \in \Delta^{m-1}$. Rather than a deterministic ground-truth action label, y_t represents how an observer population interprets the pedestrian's situated crossing intent from the currently available visual evidence.

Given N annotators providing individual interpretation labels $\alpha_t^{(n)} \in \{1, \dots, m\}$, the empirical target distribution is constructed as

$$y_{t,c} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\left[\alpha_t^{(n)} = c\right], \qquad c \in \{1, \dots, m\}. \tag{3}$$

Our objective is to learn a mapping function $\mathcal{M}$ such that

$$\hat{y}_t = \mathcal{M}(X_{1:t}) \approx P(I_t \mid X_{1:t}), \tag{4}$$

where I_t denotes the pedestrian intention *as interpreted by observers*. Crucially, TP-DID is formulated as an open-loop inference task: it models the natural evolution of subjective beliefs prior to any explicit system intervention, such as warnings or braking, that might retroactively alter driver attention.

A central requirement of this task is *distributional fidelity*: $\hat{y}_t$ must accurately preserve the entropy of the observer group, maintaining structured uncertainty when multiple interpretations remain plausible.

Optimization Objective. We minimize the Kullback–Leibler (KL) divergence between the empirical target distribution y_t

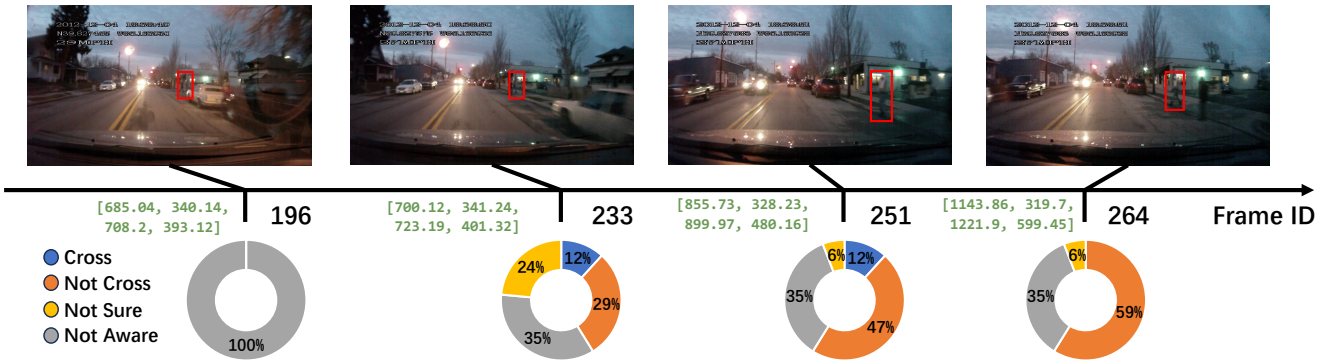


Figure 2: Temporal empirical interpretation distributions of observers toward the target pedestrian in PSI, including the pedestrian bounding box and the proportions of four categories: *Cross*, *Not Cross*, *Not Sure*, and *Not Aware*.

and the predicted distribution $\hat{y}_t$:

$$\mathcal{L}_{\text{step}}(y_t, \hat{y}_t) = D_{\text{KL}}(y_t \parallel \hat{y}_t) = \sum_{c=1}^{m} y_{t,c} \log \frac{y_{t,c}}{\hat{y}_{t,c}}. \quad (5)$$

The convention $0 \log 0 = 0$ is used. Step-wise minimization of this objective enforces causal optimization, ensuring that $\hat{y}_t$ remains a valid online estimate strictly grounded in the observation prefix $X_{1:t}$.

Key Challenges in Cognitive Modeling

The above formulation introduces three key challenges for modeling driver-side interpretation.

Non-linear Updating Dynamics. Human belief updating is highly non-linear, shifting between gradual evidence accumulation in stable environments and rapid reconfiguration during unexpected events, often mediated by neuromodulatory gain control (Frigo, 2022; Kronemyer & Bystritsky, 2014; Nassar et al., 2012; Yu & Dayan, 2005; Zhang et al., 2025). Capturing these dynamics is difficult because a model must infer both the observer's interpretation state and their latent estimate of environmental volatility.

Context-Dependent Integration. Predictive coding frameworks (Gerlach & Poirel, 2018; Huang et al., 2024) suggest humans integrate global priors with local sensory evidence by weighting prediction errors according to context-dependent precision. Models assuming stationary signal-to-noise ratios thus generalize poorly, particularly during rapid environmental transitions where subjective precision fluctuates.

Long-Tailed Cognitive States. Stable, low-entropy interpretations dominate real-world interactions, making volatile cognitive transitions exceptionally rare (Friston et al., 2012; Parr & Friston, 2019). This extreme distributional imbalance causes models trained under standard likelihood objectives to overfit stable states, failing to predict the sparse but safety-critical moments of interpretation restructuring (Gershman et al., 2017).

Methodology

Input-Modulated Mamba

The first two TP-DID challenges, non-linear updating and context-dependent integration, require a causal sequence backbone that maintains interpretation memory while adapting to evolving geometric evidence. Mamba (Dao & Gu, 2024; Gu & Dao, 2024) provides such an efficient state-space backbone, but a single input pathway may still over-accumulate frequent stable cues and respond insufficiently to interpretation-changing evidence.

We propose Input-Modulated Mamba (IM-Mamba), which augments the Mamba prediction backbone with an explicit input modulator. IM-Mamba contains two causal Mamba modules: Mamba-M generates an input-dependent channel-wise modulation signal, while Mamba-P performs recurrent prediction from the modulated evidence. The final distribution is produced only from the Mamba-P state:

$$\hat{y}_t = \psi(h_t^P), \qquad \psi : \mathbb{R}^{d_h} \to \Delta^{m-1}. \quad (6)$$

Input Representation

Let $q_t \in \mathbb{R}^{d_q}$ denote the bbox-level descriptor of the target pedestrian. We construct position, velocity, and acceleration cues as

$$x_t = [q_t; \Delta q_t; \Delta^2 q_t] \in \mathbb{R}^{d_x}, \qquad d_x = 3d_q, \quad (7)$$

where

$$\Delta q_t = q_t - q_{t-1}, \qquad \Delta^2 q_t = q_t - 2q_{t-1} + q_{t-2}. \quad (8)$$

Boundary terms are handled by causal padding. The embedded evidence is

$$z_t = \mathcal{E}(x_t) \in \mathbb{R}^{d_z}. \quad (9)$$

Explicit Input Modulation

Given the embedded prefix $z_{1:t}$, IM-Mamba defines a causal modulation function

$$a_t = \phi_\theta(z_{1:t}) \in \mathbb{R}^{d_z}, \quad (10)$$

where ϕ_θ is instantiated by Mamba-M with a lightweight projection head. The channel-wise gate is

$$g_t = |\alpha|\,\sigma(a_t - \tau), \qquad 0 < g_{t,i} < |\alpha|, \tag{11}$$

where $\sigma(\cdot)$ is applied element-wise, τ is a fixed offset, and α is a learnable scale initialized to 2.0. In implementation, ϕ_θ consists of the causal Mamba-M branch followed by a linear-normalization-Softplus projection.

The modulated evidence is

$$u_t = g_t \odot z_t, \qquad u_t \in \mathbb{R}^{d_z}. \tag{12}$$

Mamba-P updates the prediction state as

$$h_t^P = F_t^P(h_{t-1}^P, u_t), \qquad \hat{y}_t = \psi(h_t^P). \tag{13}$$

Thus, Mamba-M affects the output only through the input gate. Channels with $g_{t,i} < 1$ are attenuated, while channels with $g_{t,i} > 1$ are amplified before entering Mamba-P.

In implementation, Mamba-P is further initialized with fast and slow heads through its time-step bias: fast heads favor larger update rates, whereas slow heads favor smoother temporal integration. This initialization is applied to the prediction branch.

Local Response Analysis

We analyze the direct input-path effect of the realized gate. This analysis explains the response mechanism of IM-Mamba, rather than providing a performance guarantee over a single Mamba baseline.

Lemma 1 (First-order gated response)**.** *Assume F_t^P is differentiable with respect to its input around u_t. Let*

$$D_t = \mathrm{Diag}(g_t) \in \mathbb{R}^{d_z \times d_z}, J_t^P = \left.\frac{\partial F_t^P(h_{t-1}^P, u)}{\partial u}\right|_{u=u_t} \in \mathbb{R}^{d_h \times d_z}. \tag{14}$$

Conditioned on the realized gate g_t, a local perturbation $\delta z_t \in \mathbb{R}^{d_z}$ induces

$$\delta h_t^{P,\mathrm{IM}} = J_t^P D_t \delta z_t + o(\|\delta z_t\|_2). \tag{15}$$

Proof. Since $u_t = g_t \odot z_t = D_t z_t$, fixing g_t gives

$$\delta u_t = D_t \delta z_t. \tag{16}$$

The result follows from the first-order expansion of F_t^P with respect to its input. □

Proposition 1 (Gate-controlled immediate response)**.** *Consider a locally matched ungated path using the same prediction transition and state, but with $u_t = z_t$. For any perturbation satisfying $\|J_t^P \delta z_t\|_2 > 0$, the leading-order response ratio is*

$$A_t(\delta z_t) = \frac{\|J_t^P D_t \delta z_t\|_2}{\|J_t^P \delta z_t\|_2}. \tag{17}$$

For a channel-local cue $\delta z_t = \eta e_i$, where $\eta \neq 0$ and $J_t^P e_i \neq 0$,

$$A_t(\eta e_i) = g_{t,i}. \tag{18}$$

Hence, $g_{t,i} > 1$ yields a larger immediate response than the matched ungated path, while $g_{t,i} < 1$ attenuates it.

Proof. By Lemma 1, the gated response is $J_t^P D_t \delta z_t$. The matched ungated response is $J_t^P \delta z_t$. For $\delta z_t = \eta e_i$,

$$J_t^P D_t \delta z_t = \eta g_{t,i} J_t^P e_i, \qquad J_t^P \delta z_t = \eta J_t^P e_i. \tag{19}$$

Taking the norm ratio gives $A_t(\eta e_i) = g_{t,i}$. □

Therefore, IM-Mamba provides an explicit context-dependent response controller: the causal modulator selects the gate, and the prediction Mamba updates from the gated evidence.

Objective

Given the target distribution $y_t \in \Delta^{m-1}$ and prediction $\hat{y}_t \in \Delta^{m-1}$, we minimize the sequence-averaged KL divergence:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}(y_t \parallel \hat{y}_t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{m} y_{t,c} \log \frac{y_{t,c}}{\hat{y}_{t,c}}. \tag{20}$$

A small numerical constant is used in the logarithm.

Experiment

Experimental Setup

Compared methods. All methods follow the same TP-DID protocol.

- Transformer baseline:We use Gao et al. (2025) as the external baseline, which addresses label distribution imbalance via decoupled distribution learning.
- Mamba (ablation): We use a single-stream Mamba (Dao & Gu, 2024) with the same input and objective as IM-Mamba, but removes the modulation stream:
$$G_t = I, \qquad u_t = z_t. \tag{21}$$
to directly ablate explicit input modulation.
- IM-Mamba: IM-Mamba is the full model with bounded input modulation:
$$u_t = G_t z_t. \tag{22}$$

Evaluation metrics. We evaluate TP-DID dominant-category accuracy, distributional fidelity, and temporal consistency. For category-wise evaluation, samples are grouped by the dominant interpretation category

$$c_t^* = \arg\max_c y_{t,c}. \tag{23}$$

We report both macro- and micro-averages.

1. Dominant-category accuracy: Accuracy is computed by comparing $\arg\max_c \hat{y}_{t,c}$ with c_t^*.
2. Distributional fidelity: We use KL divergence, JS divergence, and Chebyshev distance. Lower values denote better distributional alignment.
3. Direction consistency: Temporal alignment is measured by the cosine similarity of interpretation-distribution transitions:
$$\mathrm{DC} = \frac{1}{T-1}\sum_{t=2}^{T} \frac{\langle \Delta y_t, \Delta \hat{y}_t \rangle}{\|\Delta y_t\|\|\Delta \hat{y}_t\| + \epsilon}, \tag{24}$$
where $\Delta y_t = y_t - y_{t-1}$ and $\Delta \hat{y}_t = \hat{y}_t - \hat{y}_{t-1}$. Higher DC indicates better temporal direction alignment of interpretation-distribution changes.

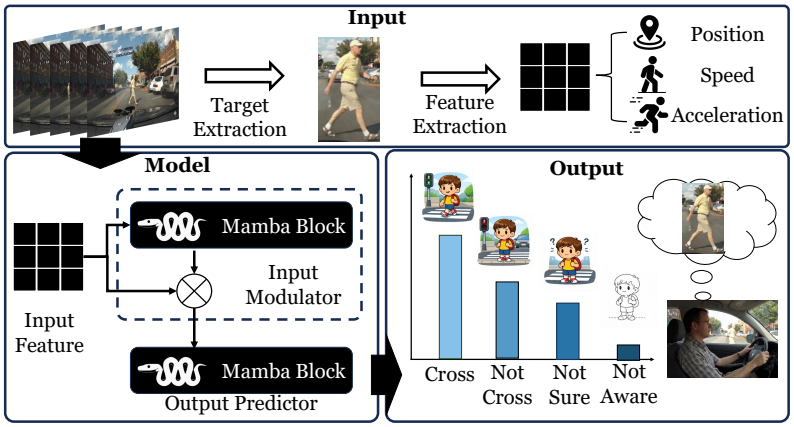


Figure 3: Overview of IM-Mamba. Mamba-M generates an input-dependent gate, while Mamba-P predicts from the gated evidence.

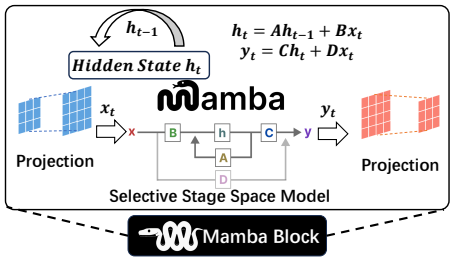


Figure 4: Mamba block.

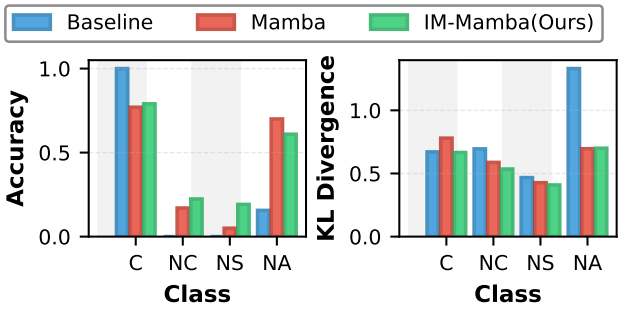


Figure 5: Category-wise comparison of Accuracy and KL-Divergence. {C, NC, NS, NA} represent {*Cross*, *Not Cross*, *Not Sure*, *Not Aware*}.

Results and Ablation Analysis

Table 1 reports the category-wise performance, while Figure 5 visualizes the temporal probability distribution trajectories (both single-sample and dataset-averaged) for each interpretation category. The Transformer baseline suffers from a strong bias toward the majority *Cross* class and fails on minority classes (*Not Cross*, *Not Sure*). The Mamba ablation improves minority-category and temporal behavior through recurrent state-space memory. IM-Mamba further improves most distributional metrics, showing that input modulation benefits category-balanced interpretation-distribution prediction.

Table 2 summarizes overall performance. IM-Mamba achieves the best macro-average scores across all metrics, which is critical under interpretation-category imbalance. Its consistent gain over Mamba isolates the effect of the modulation stream rather than recurrent memory alone. Micro-average results show the same trend, indicating no loss in overall performance.

Table 3 reports Direction Consistency. Mamba substantially improves temporal alignment over the transformer baseline, confirming the benefit of state-space memory. IM-Mamba obtains the highest mean and lowest variation, indicating that modulation further improves the tracking of interpretation-distribution transition directions.

Figure 6 shows temporal trends after sequence normalization. The transformer baseline degrades in later phases, where interpretation distributions are more likely to change. Mamba improves temporal smoothness, while IM-Mamba maintains more balanced category-wise behavior, consistent with its role in reducing recurrent response inertia.

Overall, the transformer baseline provides the external comparison, Mamba isolates recurrent state-space memory, and the gap between Mamba and IM-Mamba verifies the contribution of explicit input modulation.

Related Work

Pedestrian Intent Prediction

Pedestrian intent prediction is typically formulated as binary crossing prediction and supported by benchmarks such as JAAD and PIE (Kotseruba et al., 2016; Rasouli et al., 2019) with standardized protocols (Kotseruba et al., 2021). Recent methods incorporate uncertainty estimation to handle ambiguous observations, including evidential formulations (Zhang et al., 2023), while surveys emphasize epistemic and aleatoric uncertainty for safety-critical prediction (Scaccia et al., 2025). Different from these works, TP-DID models the driver's evolving interpretation distribution over multiple states, rather than

Table 1: Category-wise comparison.

	Cross			Not Cross			Not Sure			Not Aware		
Metric	**Baseline**	**Mamba**	**IM-Mamba**	**Baseline**	**Mamba**	**IM-Mamba**	**Baseline**	**Mamba**	**IM-Mamba**	**Baseline**	**Mamba**	**IM-Mamba**
Accuracy	**0.9978**	0.9054	0.7903	0.0000	0.0194	**0.2237**	0.0000	0.0000	**0.1918**	0.2768	**0.7275**	0.6094
KL Div	0.6743	0.7797	**0.6660**	0.6978	0.5878	**0.5353**	0.4714	0.4229	**0.4088**	1.1680	**0.6961**	0.6999
Chebyshev	0.4643	0.5108	**0.4604**	0.4329	0.3972	**0.3796**	0.3248	0.3064	**0.2961**	0.6402	0.4474	**0.4110**
JS Div	0.2018	0.2265	**0.1947**	0.1946	0.1693	**0.1534**	0.1314	0.1266	**0.1192**	0.3140	0.1997	**0.1897**

Table 2: Macro- and micro-average comparison.

	Macro-Average			Micro-Average		
Metric	**Baseline**	**Mamba**	**IM-Mamba**	**Baseline**	**Mamba**	**IM-Mamba**
Accuracy	0.3186	0.4130	**0.4538**	0.6299	0.6275	**0.6506**
KL Div	0.7529	0.6226	**0.5775**	0.7861	0.7042	**0.6468**
Chebyshev	0.4655	0.4155	**0.3867**	0.4992	0.4619	**0.4344**
JS Div	0.2105	0.1805	**0.1642**	0.2239	0.2018	**0.1856**

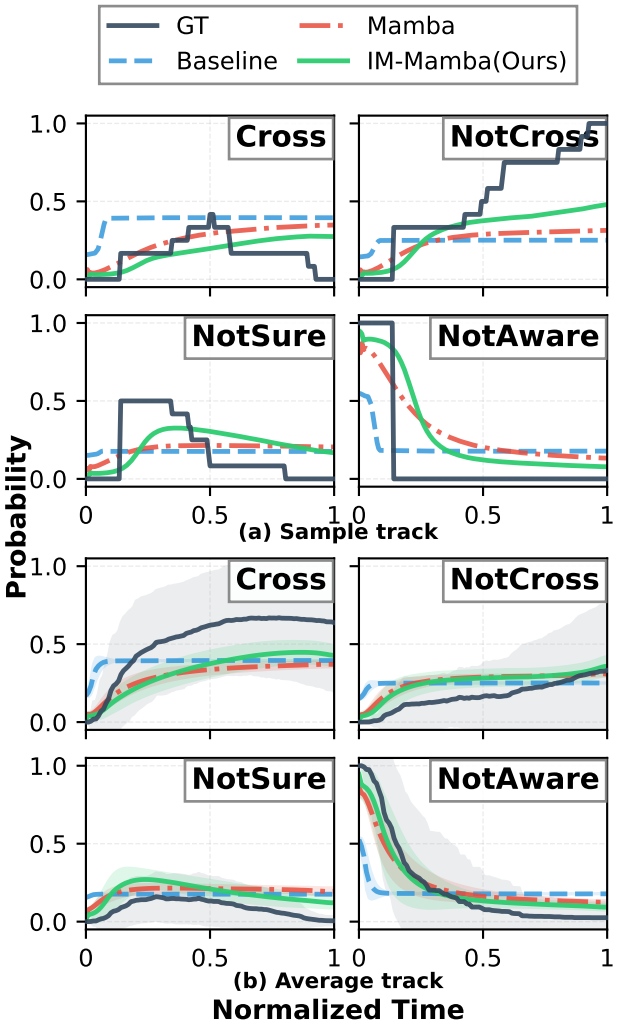


Figure 6: Temporal category-wise interpretation-distribution trends after sequence normalization.

Table 3: Direction Consistency comparison.

Method	**Direction Consistency**
Baseline	0.1667 ± 0.5135
Mamba	0.5207 ± 0.2566
IM-Mamba	**0.5652 ± 0.2422**

system-side uncertainty over binary crossing intent.

Label Distribution Learning

Label Distribution Learning (LDL) supervises each instance with a label distribution to capture ambiguity beyond one-hot labels (Geng, 2016). Recent imbalanced LDL methods improve learning when dominant labels receive most probability mass, using strategies such as representation distribution alignment (Zhao et al., 2023) and dominant/non-dominant decoupling (Gao et al., 2025). TP-DID is related to imbalanced LDL because driver-side interpretation distributions are often dominated by formed judgments, while low-probability states such as *Not Sure* and *Not Aware* remain important for assistance.

Conclusion

Human-centered driving assistance requires modeling not only system-side pedestrian intent estimation, but also the driver's interpretation under ego-view observations. This paper introduced *Temporal Prediction of Driver-side Interpretation Distributions* (TP-DID), a causal frame-level task for predicting empirical distributions over *Cross*, *Not Cross*, *Not Sure*, and *Not Aware*. With Input-Modulated Mamba (IM-Mamba), we emphasize current pedestrian-state evidence while preserving causal temporal memory. Experiments on PSI demonstrate improved distributional fidelity, minority-state prediction, and temporal transition alignment.

Acknowledgments

This work was supported by the CAS Project for Young Scientists in Basic Research (Grant No. YSBR-040) and the Key Project of the Institute of Software Chinese Academy of Sciences (ISCAS-ZD-202401)

References

Azarmi, M., Rezaei, M., & Wang, H. (2025). Pip-net: Pedestrian intention prediction in the wild. *IEEE Transactions on Intelligent Transportation Systems*.

Chen, W., Zhuang, X., Cui, Z., & Ma, G. (2019). Drivers' recognition of pedestrian road-crossing intentions: Performance and process. *Transportation research part F: traffic psychology and behaviour*, *64*, 552–564.

Dao, T., & Gu, A. (2024). Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

Frigo, V. V. (2022). *An examination of non-normative belief updating behavior in humans (why is it so hard to change minds?)* The University of Wisconsin-Madison.

Friston, K., et al. (2012). Perception and self-organized instability. *Frontiers in Computational Neuroscience*, *6*, 21.

Gao, Y., Sun, X., Ling, M., Tan, C., Zhai, Y., & Lv, G. (2025). Decoupled imbalanced label distribution learning. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Geng, X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*.

Gerlach, C., & Poirel, N. (2018). Navon's classical paradigm concerning local and global processing relates systematically to visual object classification performance. *Scientific reports*, *8*(1), 324.

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife*, *6*, e23763.

Gu, A., & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. *First conference on language modeling*.

Huang, Y. T., Wu, C.-T., Fang, Y.-X. M., Fu, C.-K., Koike, S., & Chao, Z. C. (2024). Crossmodal hierarchical predictive coding for audiovisual sequences in the human brain. *Communications Biology*, *7*(1), 965.

Jing, T., Chen, T., Tian, R., Chen, Y., Domeyer, J., Toyoda, H., Sherony, R., & Ding, Z. (2025). Psi: A benchmark for human interpretation and response in traffic interactions. *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2016). Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*.

Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2021). Benchmark for evaluating pedestrian action prediction. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Kronemyer, D., & Bystritsky, A. (2014). A non-linear dynamical approach to belief revision in cognitive behavioral therapy. *Frontiers in Computational Neuroscience*, *8*, 55.

Kuribayashi, A., Takeuchi, E., Carballo, A., Ishiguro, Y., & Takeda, K. (2023). Recognition assistance interface for human-automation cooperation in pedestrian risk prediction. *SAE International Journal of Connected and Automated Vehicles*, *6*(12-06-03-0023), 345–363.

Landry, F.-G., & Akhloufi, M. A. (2025). Predicting pedestrian crossing intention in autonomous vehicles: A review. *Neurocomputing*, *618*, 129105.

Moreno, E., Denny, P., Ward, E., Horgan, J., Eising, C., Jones, E., Glavin, M., Parsi, A., Mullins, D., & Deegan, B. (2023). Pedestrian crossing intention forecasting at unsignalized intersections using naturalistic trajectories. *Sensors*, *23*(5), 2773.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, *15*(7), 1040–1046.

Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. *Biological cybernetics*, *113*(5), 495–513.

Qu, F., Zhou, P., He, Y., Gao, K., Luo, Y., Feng, X., Liu, Y., & Guo, S. (2025). Efficientpie: Real-time prediction on pedestrian crossing intention with sole observation. *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 1793–1801.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, *3*(1), 61–70.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2019). Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Scaccia, S., Pro, F., & Amerini, I. (2025). Unsupervised pedestrian intention estimation through deep neural embeddings and spatio-temporal graph convolutional networks. *Pattern Analysis and Applications*, *28*(2), 108.

Schmidt, S., & Faerber, B. (2009). Pedestrians at the kerb–recognising the action intentions of humans. *Transportation research part F: traffic psychology and behaviour*, *12*(4), 300–310.

Sharma, N., Dhiman, C., & Indu, S. (2022). Pedestrian intention prediction for autonomous vehicles: A comprehensive survey. *Neurocomputing*, *508*, 120–152.

WorldHealthOrganization. (2023). *Pedestrian safety: A road safety manual for decision-makers and practitioners*. World Health Organization.

Yao, Y., Atkins, E., Roberson, M. J., Vasudevan, R., & Du, X. (2021). Coupling intent and action for pedestrian crossing behavior prediction. *arXiv preprint arXiv:2105.04133*.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/https://doi.org/10.1016/j.neuron.2005.04.026

Zhang, Z., Elahi, M. F., Domeyer, J., & Tian, R. (2025). Driver temporal segmentation of pedestrian crossing intentions during negotiations. *Transportation Research Part F: Traffic Psychology and Behaviour*, *114*, 953–969.

Zhang, Z., Tian, R., & Ding, Z. (2023). Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhao, X., An, Y., Xu, N., Wang, J., & Geng, X. (2023). Imbalanced label distribution learning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.