



# CaTFormer: Causal Temporal Transformer with Dynamic Contextual Fusion for Driving Intention Prediction

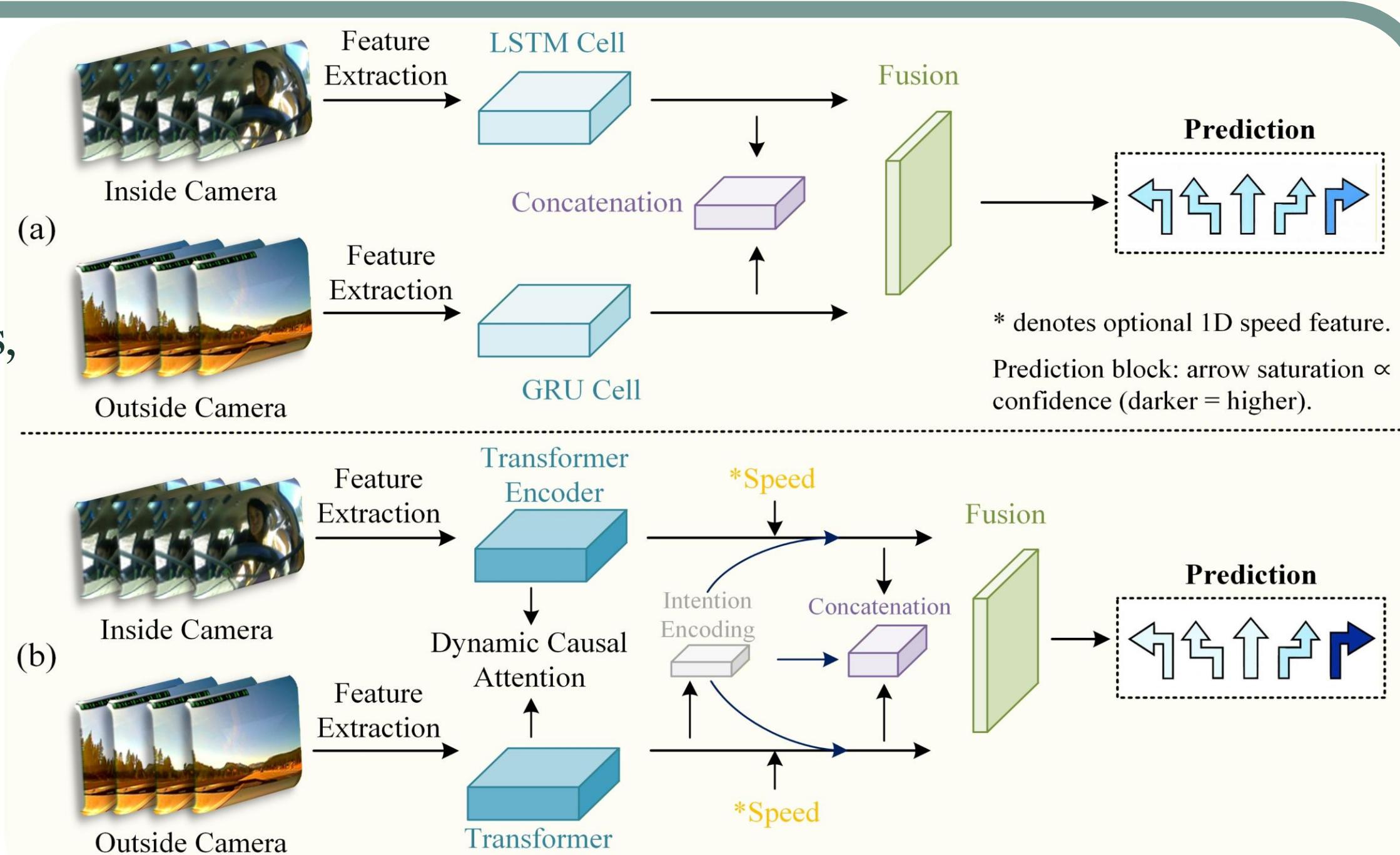
Sirui Wang<sup>†</sup>, Zhou Guant, Bingxi Zhao, Tongjia Gu, Jie Liu\*

Beijing Jiaotong University

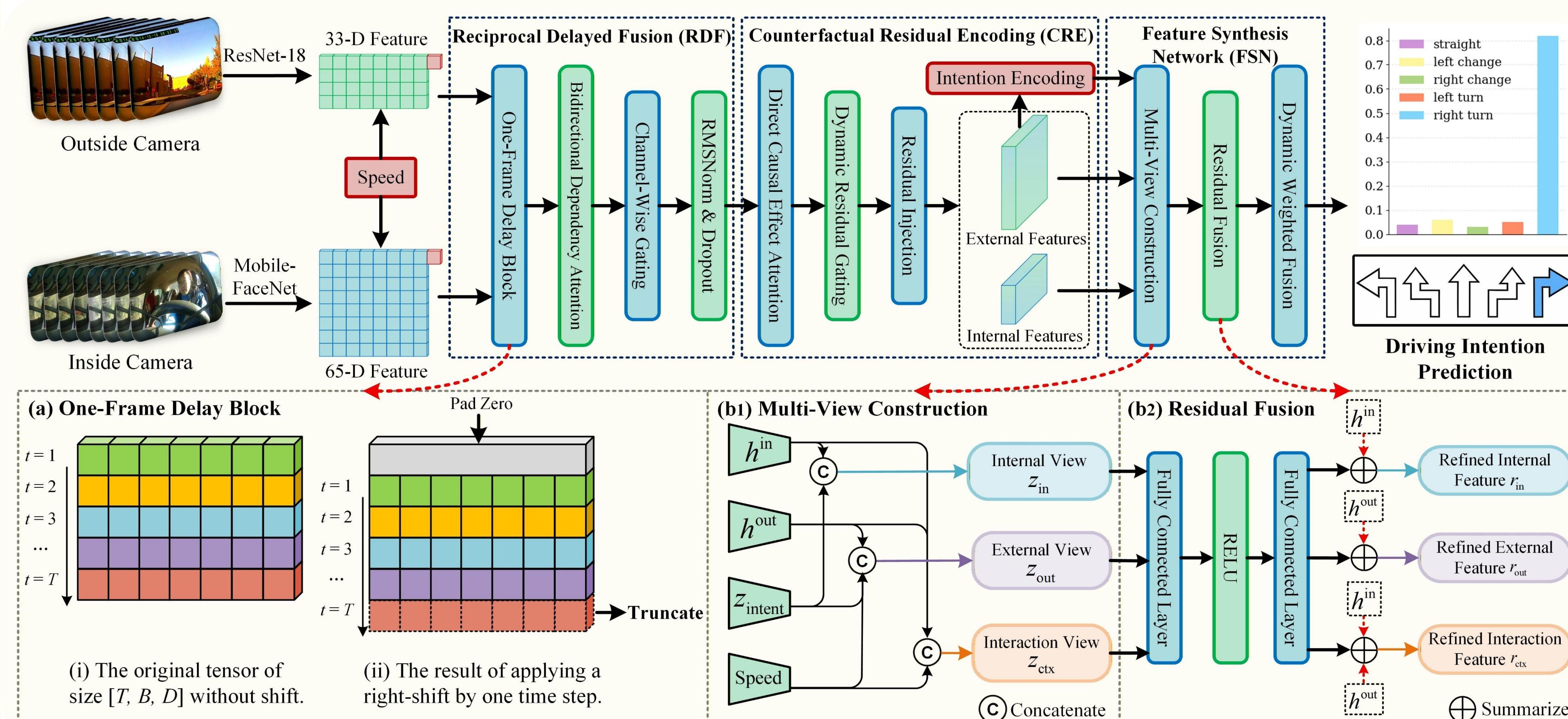


## Introduction

Accurate driving intention prediction is fundamental to enhancing the safety and interactive efficiency of Advanced Driver Assistance Systems (ADAS), yet current approaches often fail to model the complex causal interdependencies between human behavior and dynamic environments. To address this, we propose **CaTFormer**, an efficient Transformer-based framework that embeds causal temporal reasoning with adaptive multi-view fusion within a unified end-to-end architecture. By leveraging **dual-stream reciprocal delayed fusion** to capture explicit dependencies across interior and exterior streams, and employing **counterfactual attention subtraction** to isolate genuine causal effects, our method significantly enhances robustness under complex driving conditions. Extensive evaluations on the **Brain4Cars dataset** demonstrate that CaTFormer delivers superior performance, establishing a new benchmark for intention prediction in both highway and urban scenarios.



## Methodology



### Reciprocal Delayed Fusion (RDF):

**Causal Alignment:** Enforces temporal precedence via one-frame delay to ensure strict causality:

$$\hat{F}_{b,t} = F_{b,t-1} \mathbf{1}_{\{t > 1\}}$$

**Cross-Stream Fusion:** Captures interior-exterior dependencies via multi-head attention:

$$BDA(Q, K, V) = \left[ \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \right]_{i=1}^H W^O$$

**Adaptive Gating:** Filters non-causal noise using channel-wise activations:

$$\tilde{H}_{b,t} = \sigma(W_2 (\text{ReLU}(W_1 H_{b,t} + b_1)) + b_2) \odot H_{b,t}$$

### Counterfactual Residual Encoding (CRE):

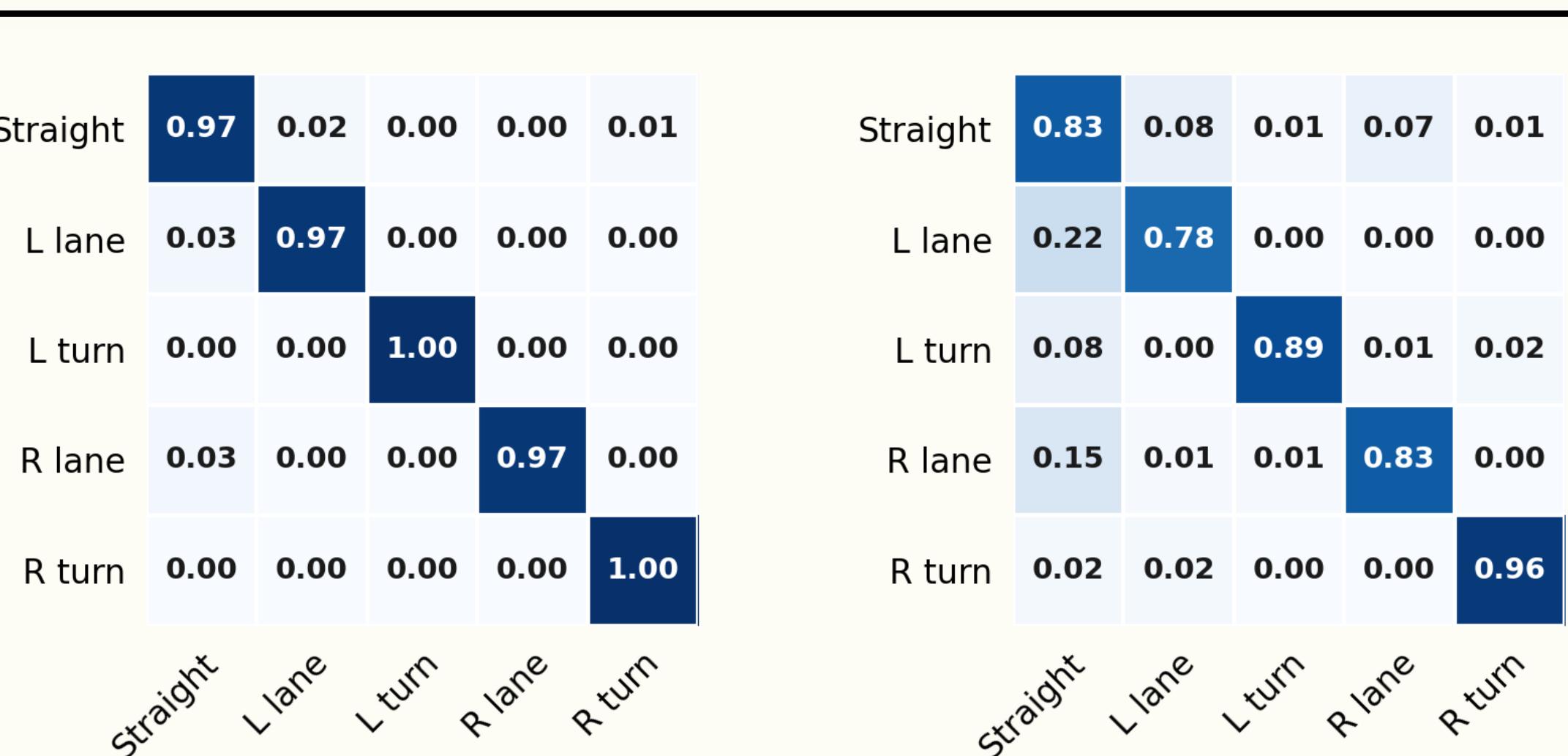
- Causal Disentanglement:** Isolates genuine dependencies by subtracting counterfactual attention (mean baseline) from observed attention:  $\Delta_t^{\text{in}} = A_{\text{in},t}^{\text{obs}} - \mathcal{A}(X_t^{\text{in}}, \bar{X}^{\text{out}}, \bar{X}^{\text{out}})$
- Bias Orthogonalization:** Removes spurious dataset correlations by projecting residuals orthogonally to the global baseline  $\bar{X}$ :  $\Delta_t^{\perp} = \Delta_t - \frac{\Delta_t^{\perp} \bar{X}}{|\bar{X}|^2 + \varepsilon} \cdot \bar{X}$
- Adaptive Refinement:** Dynamically modulates causal signals via gating and extracts global intention semantics  $\xi$ :  $h^{\text{in}} = X_T^{\text{in}} + g^{\text{in}} \cdot \Delta_T^{\perp, \text{in}}$ ,  $\xi = \text{softmax}(W_{\text{int}} h^{\text{out}})$

### Feature Synthesis Network (FSN):

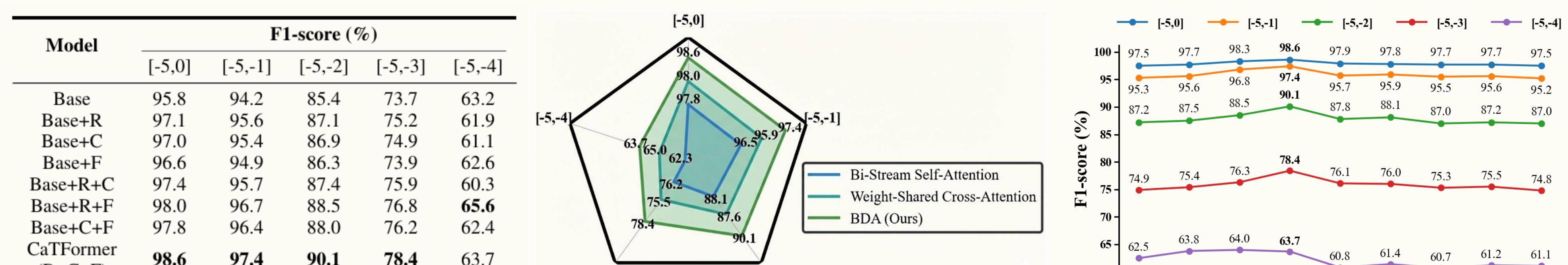
- Multi-View Fusion:** Synthesizes interior, exterior, and speed cues via residual networks:  $r_{\text{ctx}} = f_{\text{ctx}}([h^{\text{in}}, h^{\text{out}}, z_{\text{intent}}, s]) + h^{\text{in}} + h^{\text{out}}$
- Confidence Weighting:** Dynamically evaluates branch reliability  $\mathcal{C} = \{in, out, ctx\}$  using a gating mechanism:  $w_i = \exp(u_i^\top r_i) / \sum_{j \in \mathcal{C}} \exp(u_j^\top r_j)$
- Adaptive Prediction:** Aggregates branch-specific logits weighted by their confidence scores for the final decision:  $\ell_{\text{joint}} = \sum_{i \in \mathcal{C}} w_i (W_i r_i)$

## Performance on Brain4Cars Dataset

Method	Camera	GPS	Map	Speed	Pr	Re	F1-score
IOHMM (Jain et al. 2015)	✓	✓	✓	✓	74.2	71.2	72.7
SDAE (Rekabdar and Mousas 2018)	✓			✓	71.9	74.8	73.3
AIO-HMM (Jain et al. 2015)	✓	✓	✓	✓	77.4	71.2	74.2
Deep CNN (Rekabdar and Mousas 2018)	✓			✓	78.0	77.5	77.7
FRNN-UL (Jain et al. 2016b)	✓			✓	82.2	75.9	78.9
FRNN-EL (Jain et al. 2016b)	✓			✓	84.5	77.1	80.6
FRNN-EL w/ 3D head pose (Jain et al. 2016b)	✓			✓	90.5	87.4	88.9
LSTM-GRU (Tonutti et al. 2019)	✓			✓	92.3	90.8	91.3
DCNN (Rekabdar and Mousas 2018)	✓			✓	91.8	92.5	92.1
CF-LSTM (Zhou et al. 2021)	✓			✓	92.0	92.3	92.1
Predictive-Bi-LSTM-CRF (Zhou et al. 2021)	✓			✓	92.4	94.7	93.6
Central (Zhu et al. 2024)	✓		✓		94.4	94.3	94.2
FedPRM (Zhu et al. 2024)	✓	✓	✓	✓	<b>99.0</b>	92.0	95.2
Gebert (Gebert et al. 2019)	✓				-	-	81.7
Rong (Rong, Akata, and Kasneci 2020)	✓				-	-	84.3
CEMFormer (Ma et al. 2023)	✓				-	-	87.1
TIFN (Guo et al. 2023)	✓				89.3	86.4	87.9
IDIPN (Liu et al. 2025)	✓				94.2	94.9	94.5
<b>CaTFormer (Ours)</b>	✓			✓	96.7	<b>98.5</b>	97.6
					98.7	<b>98.5</b>	<b>98.6</b>



## Ablation Studies

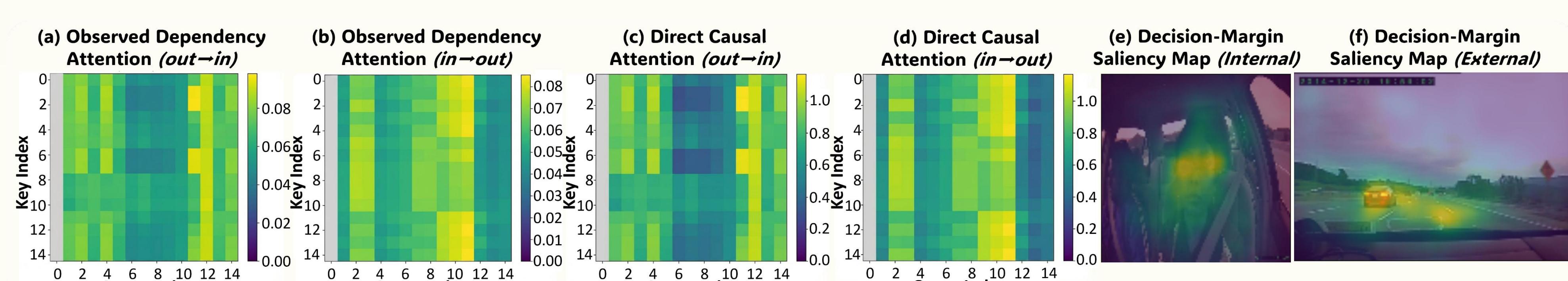


### Module Ablation

### Attention Ablation

### Parameter Ablation

## Visualization of Causal Attention & Saliency



CaTFormer is a novel causal Transformer that explicitly models the **dynamic interactions** between driver behavior and environmental context. Validated by **state-of-the-art** results on the Brain4Cars dataset, it proves to be a robust and transparent solution essential for enhancing the safety and reliability of real-time autonomous driving systems.

