

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

Sirui Wang*, Zhou Guan*, Bingxi Zhao, Tongjia Gu, Jie Liu[†]

Beijing Jiaotong University
{siruiwang, zhouguan, bingxizhao, gut2gu, jieliu}@bjtu.edu.cn

Abstract

Accurate prediction of driving intention is key to enhancing the safety and interactive efficiency of human-machine co-driving systems. It serves as a cornerstone for achieving high-level autonomous driving. However, current approaches remain inadequate for accurately modeling the complex spatiotemporal interdependencies and the unpredictable variability of human driving behavior. To address these challenges, we propose CaTFormer, a causal Temporal Transformer that explicitly models causal interactions between driver behavior and environmental context for robust intention prediction. Specifically, CaTFormer introduces a novel Reciprocal Delayed Fusion (RDF) mechanism for precise temporal alignment of interior and exterior feature streams, a Counterfactual Residual Encoding (CRE) module that systematically eliminates spurious correlations to reveal authentic causal dependencies, and an innovative Feature Synthesis Network (FSN) that adaptively synthesizes these purified representations into coherent temporal representations. Experimental results demonstrate that CaTFormer attains state-of-the-art performance on the Brain4Cars dataset. It effectively captures complex causal temporal dependencies and enhances both the accuracy and transparency of driving intention prediction.

Code — <https://github.com/srwang0506/CaTFormer>

Introduction

Driver intention prediction is crucial for autonomous driving systems, as it effectively mitigates risks and enhances driving safety. By forecasting potential outcomes several seconds in advance, the system can proactively alert the driver or initiate evasive maneuvers, significantly improving its safety capabilities.

Initially, intention prediction primarily relied on the extraction and fusion of visual features for basic predictions (Huang et al. 2022). However, advancements in sensor technology have enabled the incorporation of multi-modal information, such as GPS coordinates, vehicle speed, map data, and driver head pose (Hu et al. 2021; Mo et al. 2023; Li, Zhao, and Wang 2022; Wu et al. 2023). By leveraging complex models for feature extraction, fusion, and prediction,

*These authors contributed equally.

[†]Corresponding Author: jieliu@bjtu.edu.cn.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

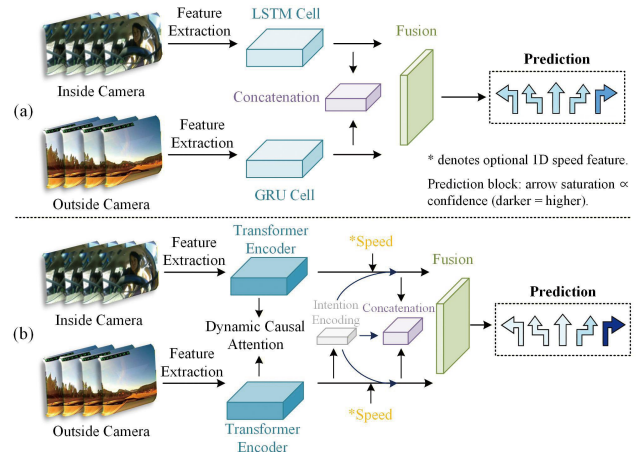


Figure 1: Comparison between previous driving intention prediction methods and ours. (a) is an LSTM-GRU framework processing interior and exterior streams independently before concatenation. (b) is our CaTFormer, a Transformer-based model enabling dynamic causal fusion of dual streams with integrated intention priors. Through joint modeling of global-local dependencies and cross-stream interactions, our approach outperforms existing methods.

the performance of driver intention prediction has significantly improved (Sui et al. 2021; Guo et al. 2023; Liu, Wu, and Wang 2023; Gao et al. 2023).

Despite advancements in driver intention prediction, the rich multi-modal data remains underutilized. Most existing studies simply concatenate or linearly aggregate this information, as shown in Fig. 1 (a). However, given that the driver controls the vehicle, changes in their state directly influence the vehicle’s driving status, indicating a strong dependency. Therefore, we propose to explicitly model the causal relationship between the driver and the environment, highlighting this causality’s decisive impact on the prediction task, as illustrated in Fig. 1 (b).

Specifically, we adopt a Transformer-based architecture and introduce three sequential components to enhance the causal modeling, encode the driver’s intention, and fuse the multi-dimensional features. First, we introduce a Reciprocal

Delayed Fusion (RDF) module that cross-fuses interior and exterior features through a shifting mechanism, explicitly establishing a temporal dependency between the two feature streams. As the fused features may contain considerable causal noise, we further devise a Counterfactual Residual Encoding (CRE) module to filter out such noise to obtain a more explicit causal representation. Finally, we utilize a Feature Synthesis Network (FSN) that employs a gating mechanism to integrate the interior, exterior, and interaction representations, enabling the modeling of both local and global causal structures. Our main contributions are as follows:

- We propose CaTFormer, an efficient Transformer-based framework for driving intention prediction that embeds causal spatio-temporal reasoning with adaptive multi-view fusion in a unified end-to-end architecture.
- Through dual-stream reciprocal delayed fusion, CaTFormer explicitly captures dependencies across interior and exterior streams, isolates genuine causal effects via counterfactual attention subtraction, and adaptively integrates complementary visual cues, effectively enhancing robustness under complex driving conditions.
- Extensive evaluation on the Brain4Cars dataset demonstrates that CaTFormer demonstrates superior performance in driving intention prediction in both highway and urban scenarios.

Related Work

In the early stages of driving intention prediction research, studies primarily focused on learning spatiotemporal representations directly from raw video, often employing 3D CNN-LSTM architectures for maneuver prediction (Gebert et al. 2019). While some later work attempted to enrich context by fusing interior and exterior streams via convolutional LSTMs for a more comprehensive decision-making basis (Rong, Akata, and Kasneci 2020), both approaches struggled with limited capacity for modeling long-range, non-consecutive dependencies.

Inspired by human cognitive processes, TIFN (Guo et al. 2023) introduced a state update unit (STU) to integrate environmental context into driver state modeling and extract semantic segmentation features as attention cues. Similarly, another study framed intention prediction as a sequence-labeling task, combining bidirectional LSTMs with a conditional random field to capture the contextual dependencies of driving behaviors (Zhou et al. 2021). Although these methods incorporate multi-source information, they typically learn inter-modal feature correlations implicitly, lacking explicit disentanglement and reasoning of their interactions. Moreover, to enhance model generalization, existing studies have developed personalized prediction models using techniques like domain-adversarial RNN (Tonutti et al. 2019), inverse reinforcement learning (Liu et al. 2025), and a federated learning framework (Zhu et al. 2024).

The Transformer architecture excels at capturing long-range dependencies and global interaction information in temporal data through its unique attention mechanism (Liu et al. 2024). Building on this, CemFormer (Ma et al. 2023) integrates data from both interior and exterior cameras,

learning a unified cross-view representation via a spatio-temporal Transformer to infer driver intention directly from their behavior. However, these methods primarily rely on the Transformer’s inherent structure to interpret dependencies. Meanwhile, a non-autoregressive Transformer with hybrid attention has been employed to simultaneously capture the temporal dynamics of a single vehicle and interactions among multiple vehicles (Jiang et al. 2024). DriveTransformer (Jia et al. 2025) further establishes a unified end-to-end framework to handle perception, prediction, and planning tasks in parallel. While this approach treats intention prediction as an integrated component of a scalable system, it may dilute the model’s focus on specific driver-intention features.

Besides, some early works also integrated causal inference methods into intention prediction. For instance, one study built a causal model to capture temporal relationships within invariant representations from driving data, aiming for domain generalization (Hu et al. 2022). Another adopted a driver-centric approach, framing risk object identification as a causal inference problem and introducing a two-stage causal framework (Li, Chan, and Chen 2020). However, these methods typically cover a relatively limited scope of scenarios and conditions.

Building on prior work, we propose a dual-stream Transformer architecture that leverages a learned intention embedding to explicitly capture both local and global causal dependencies between in-cabin and external modalities. Through systematic fusion of these multi-dimensional feature representations, the model achieves highly accurate intention prediction in complex driving environments.

Method

As illustrated in Fig. 2, **CaTFormer** processes a bi-stream image sequence $\mathcal{I} = \{(I_{b,t}^{\text{out}}, I_{b,t}^{\text{in}})\}_{b=1,t=1}^{B,T}$ of B synchronized frame pairs over T time steps, where b and t index the sample and temporal frame, respectively. Feature extractors encode each stream into features $F^{\text{out}}, F^{\text{in}} \in \mathbb{R}^{B \times T \times D}$, where D denotes the dimensionality of each encoded feature vector. The features pass through a Reciprocal Delayed Fusion (RDF) module for temporal causality modeling via our proposed bi-stream attention, followed by a Counterfactual Residual Encoding (CRE) module to inject learnable causal representations. Finally, visual features ($z_{\text{in}}, z_{\text{out}}$ and interaction features z_{ctx}) are adaptively fused by a Feature Synthesis Network (FSN) into a joint prediction ℓ_{joint} .

Reciprocal Delayed Fusion (RDF)

On multi-lane highways, cameras concurrently record the exterior traffic scene and the driver’s interior state. Motivated by bidirectional dependencies between environmental context and driving behavior, our dual-stream architecture explicitly models their interactions by jointly processing both feature streams.

To model inter-frame temporal precedence, we introduce a temporal delay mechanism in the Key and Value sequences. Specifically, at time step t , the attention mechanism accesses only information from the preceding frame $t - 1$.

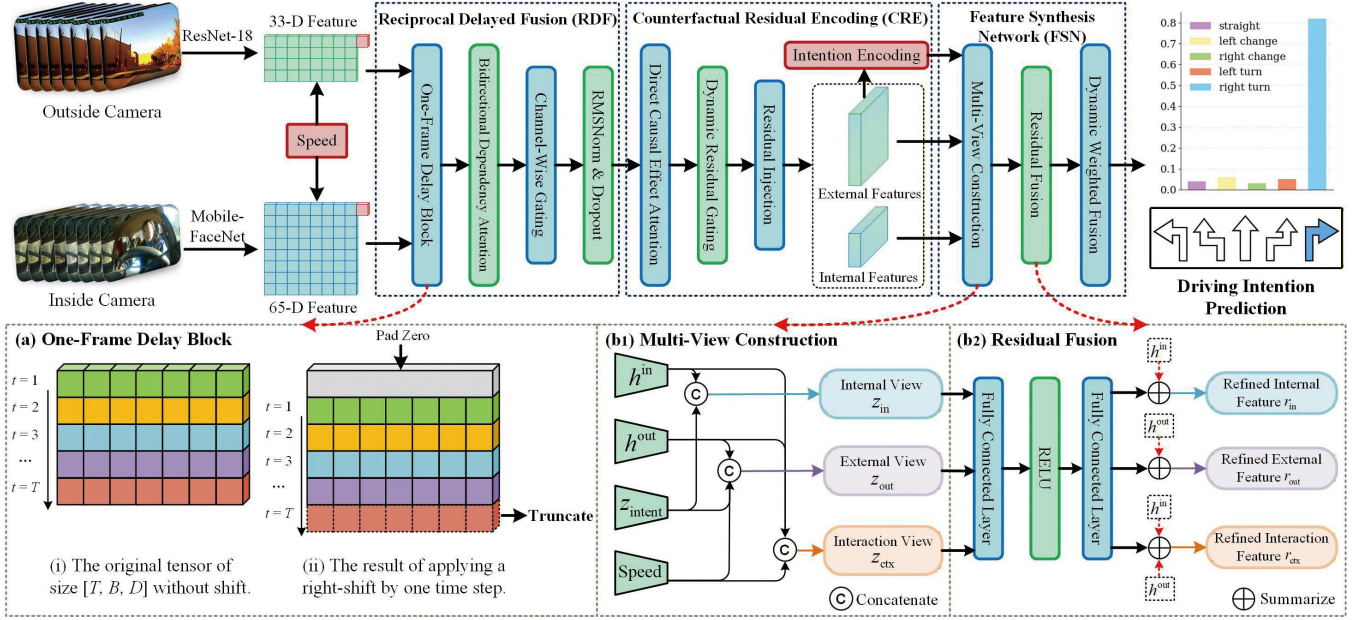


Figure 2: Overview of the **CaTFormer** pipeline. After data preprocessing, exterior optical flow is encoded by ResNet-18 and interior images by MobileFaceNet to produce dual-stream feature sequences. These are then fed into three core modules: **(1) Reciprocal Delayed Fusion (RDF)** for temporal feature integration; **(2) Counterfactual Residual Encoding (CRE)** for causal enhancement and intention embedding; and **(3) Feature Synthesis Network (FSN)** for dynamic fusion of complementary interior, exterior and interaction views to yield the final driving intention prediction.

Concretely, we define the delayed feature

$$\hat{F}_{b,t} = F_{b,t-1} \mathbf{1}_{\{t>1\}}, \quad \mathbf{1}_{\{t>1\}} = \begin{cases} 1, & t > 1, \\ 0, & t = 1. \end{cases} \quad (1)$$

applied separately to F^{out} and F^{in} .

Bidirectional Dependency Attention (BDA). Under a strict single-frame delay constraint, BDA enriches each frame’s representation by fusing interior and exterior contexts from the immediately preceding timestep. The current interior and exterior features attend bidirectionally to their one-frame delayed counterparts, capturing both temporal coherence and cross-stream coupling. To model diverse associations efficiently, we project into H parallel attention heads (in our experiments, $H = 8$) and aggregate their outputs through concatenation and a final linear mapping:

$$\text{BDA}(Q, K, V) = \left[\text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \right]_{i=1}^H W^O, \quad (2)$$

where d_k is the key dimension, $[\cdot]_{i=1}^H$ denotes concatenation across heads, and W^O restores the original feature size. Fig. 3 shows the bidirectional query-key-value fusion, highlighting how interior and exterior streams are jointly updated.

Channel-wise gating. Although BDA generates a fused representation $H_{b,t}$ at each spatial-temporal point, some channels may still carry noise or irrelevant signals. Thus, we apply two channel-wise gating layers to adaptively enhance

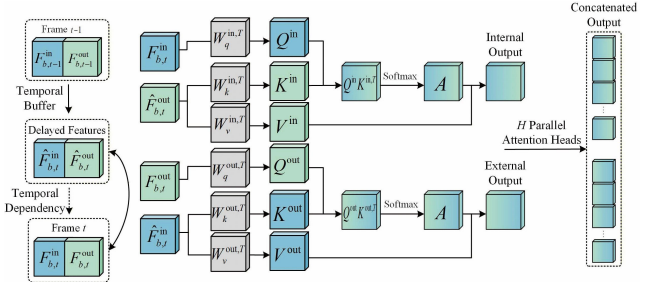


Figure 3: Illustration of Bidirectional Dependency Attention (BDA), where buffered interior and exterior features cross-attend to enhance current-frame representations.

informative features and suppress spurious ones:

$$g_{b,t} = \sigma(W_2(\text{ReLU}(W_1 H_{b,t} + b_1)) + b_2), \quad (3)$$

$$\tilde{H}_{b,t} = g_{b,t} \odot H_{b,t},$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and \odot denotes the Hadamard product between vectors.

Normalization and regularization. For numerical stability and to guard against overfitting, the following operations are applied to the gated outputs $\tilde{H}_{b,t}$:

$$R(x) = x \odot \frac{s}{\sqrt{\frac{1}{D} \sum_{d=1}^D x_d^2 + \epsilon}}, \quad (4)$$

$$X_{b,t} = \text{Dropout}(R(\tilde{H}_{b,t})), \quad (5)$$

where $R(\cdot)$ denotes the Root-mean-square normalization function, s is a learnable scaling parameter, ϵ is a small constant to ensure numerical stability, and D is the dimensionality of the feature vector x , yielding channel-calibrated representations across both interior and exterior feature streams.

Counterfactual Residual Encoding (CRE)

Conventional intention prediction architectures aggregate heterogeneous interior and exterior cues under an implicit correlation assumption and thus often mistake coincidental patterns for proper decision drivers. To overcome this limitation, our CRE module contrasts observed and counterfactual cross-stream attentions to disentangle direct causal contributions. We then selectively amplify only those residuals that genuinely influence driving intention, resulting in improved robustness and generalization in safety-critical scenarios. Specifically, CRE takes the bidirectionally fused interior and exterior features $X_{b,T}^{\text{in}}, X_{b,T}^{\text{out}} \in \mathbb{R}^{T \times \tilde{B} \times D}$ as inputs to perform this causal reasoning process.

Direct causal effect. At each time step, we compute two attention distributions. We first calculate observed dependency attention $A_{\text{in},t}^{\text{obs}}$ using actual exterior features, and then generate counterfactual dependency attention $A_{\text{in},t}^{\text{cf}}$ by replacing all exterior features with their temporal mean $\bar{X}^{\text{out}} = \frac{1}{TB} \sum_{u=1}^T \sum_{b=1}^{\tilde{B}} X_{u,b}^{\text{out}}$, which serves as a neutral baseline that removes environmental variations. The difference Δ_t^{in} between these two distributions quantifies the direct causal attention of exterior context on interior representations. Specifically, for $t = 1, \dots, T$, we obtain

$$\begin{aligned} A_{\text{in},t}^{\text{obs}} &= \text{BDA}(X_t^{\text{in}}, X_{t-1}^{\text{out}}, X_{t-1}^{\text{out}}), \\ A_{\text{in},t}^{\text{cf}} &= \mathcal{A}(X_t^{\text{in}}, \bar{X}^{\text{out}}, \bar{X}^{\text{out}}), \\ \Delta_t^{\text{in}} &= A_{\text{in},t}^{\text{obs}} - A_{\text{in},t}^{\text{cf}}, \end{aligned} \quad (6)$$

where $\mathcal{A}(Q, K, V)$ denotes multi-head scaled dot-product attention. Similarly, Δ_t^{out} is defined by interchanging the two streams. We further orthogonalize each causal residual against the global baseline vector \bar{X} to ensure that the identified causal patterns reflect true intention-relevant dependencies rather than dataset-specific biases. Formally,

$$\Delta_t^\perp = \Delta_t - \frac{\Delta_t^\top \bar{X}}{\|\bar{X}\|^2 + \epsilon} \cdot \bar{X}, \quad (7)$$

where ϵ ensures numerical stability. The orthogonal projection yields Δ_t^\perp by removing baseline-aligned components. We retain Δ_T^\perp as the decision-relevant signal, guiding downstream fusion toward temporally salient causal cues.

Dynamic residual gating. The causal relevance of residuals differs across driving scenarios, as critical maneuvers merit amplification, while routine patterns warrant attenuation. We use a learnable gating mechanism to adjust residual contributions according to their predictive value for intention inference. Specifically, we derive gating coefficients for the orthogonally filtered residuals $\Delta_T^{\perp, \text{in}}$ and $\Delta_T^{\perp, \text{out}}$ using a linear layer followed by sigmoid activation, and then integrate these gated residuals with the original features:

$$h^{\text{in}} = X_T^{\text{in}} + g_T^{\text{in}} \cdot \Delta_T^{\perp, \text{in}}, \quad h^{\text{out}} = X_T^{\text{out}} + g_T^{\text{out}} \cdot \Delta_T^{\perp, \text{out}}, \quad (8)$$

where g_T^{in} and g_T^{out} are learned gating coefficients that selectively modulate causal signal contributions to enhance prediction robustness.

Adaptive intention encoding. Beyond frame-level causal cues, holistic intention understanding requires global semantic reasoning. We extract a coarse intention distribution from the exterior summary h^{out} through softmax classification over M predefined intention categories (M denotes the total number of classes):

$$\xi = \text{softmax}(W_{\text{int}} h^{\text{out}}) \in \mathbb{R}^M \quad (9)$$

where $W_{\text{int}} \in \mathbb{R}^{M \times D}$ linearly projects the D -dimensional exterior summary onto M logits. This intention distribution is then re-embedded as an intention token $z_{\text{intent}} \in \mathbb{R}^D$ that encodes the driving intention in a continuous representation. The intention token serves as a global semantic anchor, providing top-down guidance across processing streams for consistent interpretation of ambiguous scenarios.

Feature Synthesis Network (FSN)

The CRE module provides a set of disentangled feature vectors corresponding to interior and exterior cues and a preliminary intention token. We further introduce the Feature Synthesis Network (FSN), which performs adaptive fusion of these features to construct a superior synthesized representation for predicting driving intention. By selectively emphasizing the most relevant information, the FSN module enhances the robustness of driving intention prediction. Each visual branch undergoes a residual nonlinear transformation via a dual-stage feedforward network with intermediate activation, which, combined with the speed feature s , yields the fused representations for the interior, exterior, and interaction streams:

$$\begin{aligned} r_{\text{in}} &= f_{\text{in}}([h^{\text{in}}, z_{\text{intent}}]) + h^{\text{in}}, \\ r_{\text{out}} &= f_{\text{out}}([h^{\text{out}}, z_{\text{intent}}, s]) + h^{\text{out}}, \\ r_{\text{ctx}} &= f_{\text{ctx}}([h^{\text{in}}, h^{\text{out}}, z_{\text{intent}}, s]) + h^{\text{in}} + h^{\text{out}} \end{aligned} \quad (10)$$

where each f denotes a dual-stage feedforward mapping comprising two fully connected layers separated by a ReLU activation (FC–ReLU–FC). Let $\mathcal{C} = \{\text{in}, \text{out}, \text{ctx}\}$. Each refined feature r_i ($i \in \mathcal{C}$) is mapped to class logits ℓ_i and a corresponding confidence weight w_i , which adaptively controls each branch’s contribution:

$$w_i = \frac{\exp(u_i^\top r_i)}{\sum_{j \in \mathcal{C}} \exp(u_j^\top r_j)}, \quad \ell_{\text{joint}} = \sum_{i \in \mathcal{C}} w_i (W_i r_i). \quad (11)$$

Model Training

To address class imbalance and enhance sensitivity to rare intentions while promoting early prediction, we design a unified loss function that combines the average cross-entropy (CE) across complementary streams with an intention-prediction term:

$$\mathcal{L} = \underbrace{\frac{1}{4} \sum_{i \in \mathcal{H}} \text{CE}(\ell_i, y)}_{\text{main loss}} + \underbrace{\alpha \text{CE}(\ell_{\text{intent}}, y)}_{\text{intention loss}} \quad (12)$$

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)	✓	✓		✓	99.0	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
CaTFormer (Ours)	✓				96.7	98.5	97.6
	✓			✓	98.7	98.5	98.6

Table 1: Comparison of state-of-the-art methods on the Brain4Cars dataset using camera and additional sensor modalities (GPS, Map, Speed). The best results are highlighted in bold.

where $\mathcal{H} = \{\text{in, out, ctx, joint}\}$ denotes the four stream-level heads, and α controls the weight of the intention supervision term. This unified objective integrates class-imbalance mitigation, multi-view fusion, and intention supervision within a cohesive framework.

Experiments

Data Preparation

Brain4Cars: The Brain4Cars dataset (Jain et al. 2016a) comprises exterior (480×720) and interior (1088×1920) videos of up to 5-second segments, refined to 594 valid events after excluding incomplete or unsynchronized samples. Each video is uniformly sampled to 150 frames, extracting the 5-second segment preceding the maneuver. Interior frames are cropped to 900×800 , resized to 112×112 , and encoded by a MobileFaceNet yielding 64-D features. Exterior frames undergo Farneback optical flow computation, are resized to 144×96 , and then processed by ResNet-18 to produce 32-D features. Appending a smoothed speed signal yields 65-D (interior) and 33-D (exterior) vectors. These vectors are linearly projected, positionally encoded, and passed through a Transformer encoder to obtain temporal representations for CaTFormer. The dataset spans highway and urban settings with five maneuver classes: straight, left turn, right turn, left lane change, and right lane change.

Implementation Details

Our proposed CaTFormer was implemented by PyTorch, and experiments were performed on a server with six NVIDIA RTX 2080 Ti GPUs. The model was trained end-to-end on Brain4Cars using the Adam optimizer (initial learning rate 1×10^{-3}) for 160 epochs with a batch size

Method	In	Out	F1 (%)	Param. (M)
Gebert (Gebert et al. 2019)	✓		81.7	85.3+162
	✓	✓	43.4	85.3+162
Rong (Rong, Akata, and Kasneci 2020)	✓		75.5	46.2+162
	✓	✓	66.4	5.4+162
TIFN (Guo et al. 2023)	✓	✓	87.9	12.3+5.3
IDIPN (Liu et al. 2025)	✓	✓	94.5	11.75+5.3
CaTFormer (Ours)	✓	✓	98.6	14.53+5.3

Table 2: Comparison of our CaTFormer against other end-to-end methods on the Brain4Cars dataset, using interior and exterior streams, with F1-score (%) and parameters (M).

of 16. During training, each input comprised a chunk of frames randomly sampled from the 5-second pre-maneuver segment. When testing, chunks were obtained via uniform sampling. The weight α in the unified loss was empirically set to 0.1. Model performance was evaluated using 5-fold cross-validation.

Evaluation Protocols

In driving intention prediction, straight driving is considered background, and only turns and lane changes are treated as target events. To evaluate our CaTFormer model, we define the following prediction-based metrics: true positives

Method	F1-score (%)				
	[-5,0]	[-5,-1]	[-5,-2]	[-5,-3]	[-5,-4]
Rong (Rong, Akata, and Kasneci 2020) (in)	75.7	73.1	68.6	58.5	48.2
Rong (Rong, Akata, and Kasneci 2020) (out)	66.4	62.4	47.0	38.8	38.9
Rong (Rong, Akata, and Kasneci 2020) (both)	84.3	78.9	70.6	60.3	53.4
TIFN (Guo et al. 2023)	87.9	80.9	71.0	55.0	44.6
IDIPN (Liu et al. 2025)	94.5	84.1	74.2	62.0	55.4
CaTFormer (Ours)	98.6	97.4	90.1	78.4	63.7

Table 3: F1-scores on the Brain4Cars dataset for evaluation on video segments truncated 1–4 s before action onset.

(TP: correctly predicted maneuvers), false positives (FP: maneuvers misclassified as another maneuver), false optimistic predictions (FPP: predicting a maneuver when none occurred), and missing predictions (MP: failing to detect an actual maneuver). Given the set of all behaviors \mathcal{G} and target maneuvers $\mathcal{G}' = \mathcal{G} \setminus \{\text{straight}\}$, Precision (Pr), Recall (Re), and F1-score are computed as follows:

$$\text{Pr} = \frac{1}{|\mathcal{G}'|} \sum_{m \in \mathcal{G}} \frac{TP_m}{TP_m + FP_m + FPP_m},$$

$$\text{Re} = \frac{1}{|\mathcal{G}'|} \sum_{m \in \mathcal{G}} \frac{TP_m}{TP_m + MP_m}, \quad F_1 = \frac{2 * \text{Pr} * \text{Re}}{\text{Pr} + \text{Re}}. \quad (13)$$

Comparison with State-of-the-art Methods

Table 1 provides a systematic comparison of both single- and multi-modal methods on the Brain4Cars dataset. Notably, our camera-only CaTFormer variant achieves an F1-score of 97.6% (precision 96.7%, recall 98.5%), markedly surpassing all previous single-modality methods such as DCNN (92.1%) and CF-LSTM (92.1%). When enriched with speed information, CaTFormer attains a new state-of-the-art F1-score of 98.6% (precision 98.7%, recall 98.5%), outperforming the best prior multi-modal model, FedPRM (95.2% F1), by 3.4%. These results demonstrate that CaTFormer not only establishes a new performance standard but does so with fewer sensor inputs, highlighting its efficiency and robustness in driving intention prediction. Fig. 4 presents the confusion matrices of our CaTFormer and TIFN (Guo et al. 2023). CaTFormer yields a sharper diagonal and substantially fewer off-diagonal entries, demonstrating its superior discrimination of similar maneuvers and reduced false predictions. Detailed comparative results between our method and other end-to-end approaches on full-video inputs appear in Table 2. As the optical-flow algorithm lies outside the core prediction pipeline, its parameters are listed

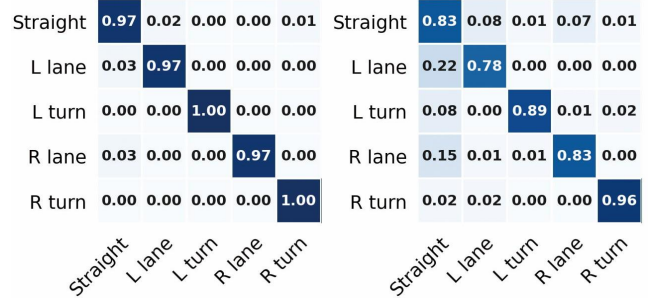


Figure 4: The confusion matrix tested on Brain4cars dataset. Left is ours, right is the result of TIFN (Guo et al. 2023). The color deepens as the value increases.

Model	F1-score (%)				
	[-5,0]	[-5,-1]	[-5,-2]	[-5,-3]	[-5,-4]
Base	95.8	94.2	85.4	73.7	63.2
Base+R	97.1	95.6	87.1	75.2	61.9
Base+C	97.0	95.4	86.9	74.9	61.1
Base+F	96.6	94.9	86.3	73.9	62.6
Base+R+C	97.4	95.7	87.4	75.9	60.3
Base+R+F	98.0	96.7	88.5	76.8	65.6
Base+C+F	97.8	96.4	88.0	76.2	62.4
CaTFormer (R+C+F)	98.6	97.4	90.1	78.4	63.7

Table 4: F1-scores on the Brain4Cars dataset for the dual-stream Transformer baseline (**Base**) and its variants augmented with RDF (**R**), CRE (**C**), and FSN (**F**).

separately. Our model achieves superior recognition performance with only a marginal increase in model size, demonstrating its compact efficiency.

In addition to evaluating F1-scores on complete videos (−5 s to 0 s, where 0 s marks driver action), we assessed early-warning capability by truncating observation windows at −1 s, −2 s, −3 s, and −4 s. As shown in Table 3, prediction accuracy declines nearly linearly with shorter observations, highlighting increased uncertainty at longer forecast horizons. This result reflects the intrinsic trade-off between early intervention and predictive accuracy. Our CaTFormer consistently achieves superior performance across all truncated settings, demonstrating its robustness in driving intention prediction.

Result Visualization

Fig. 5 demonstrates that the model employs a temporal attention mechanism to realize a full reasoning path from dynamic event understanding to static decision attribution. Temporally, (a) and (b) delineate broad, task-relevant event windows (e.g., the lane-change interval), whereas (c) and (d) concentrate on a small set of decisive frames, accentuating discriminative cues and pinpointing instantaneous triggers. This sequence closely mirrors human cognition in which one perceives an event in its entirety before pinpointing its core cause. In the spatial dimension, (e) and

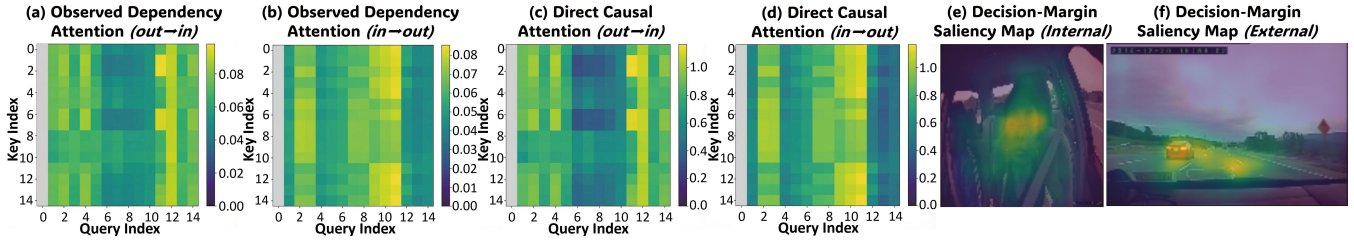


Figure 5: Temporal Attention and Decision-Margin Saliency Visualizations. (a) and (b) display the *Observed Dependency Attention*, where each pixel w_{ij} (row i , column j) represents the attention weight from the j -th query to the i -th key for out \rightarrow in and in \rightarrow out directions, respectively. (c) and (d) demonstrate the *Direct Causal Attention*, highlighting frames with significant causal influence. The first column is gray masked to mark shift padding. (e) and (f) overlay the Decision-Margin Saliency Map on the final interior and exterior frames, each pixel’s intensity defined by $\sum_{t \in T} \alpha_t \left| \partial(z_{c^*} - \frac{1}{g-1} \sum_{c \neq c^*} z_c) / \partial x_t \right|$, where α_t denotes the causal-attention weight for frame t , x_t denotes the input feature vector at that pixel, z_c is the final logit for class c and c^* is the predicted class, highlighting regions most responsible for the model’s final decision. Yellow indicates stronger attention.

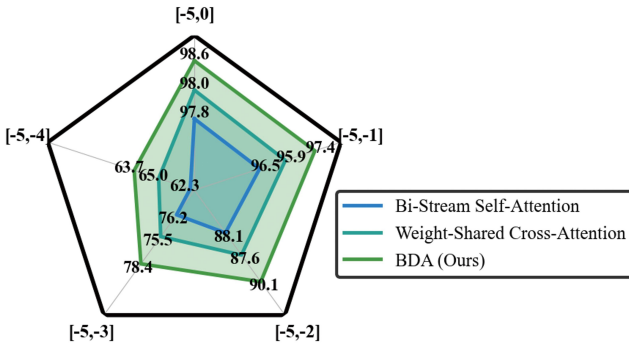


Figure 6: F1-scores (%) of various attention mechanisms on the Brain4Cars dataset, offset by 15 units for visual clarity.

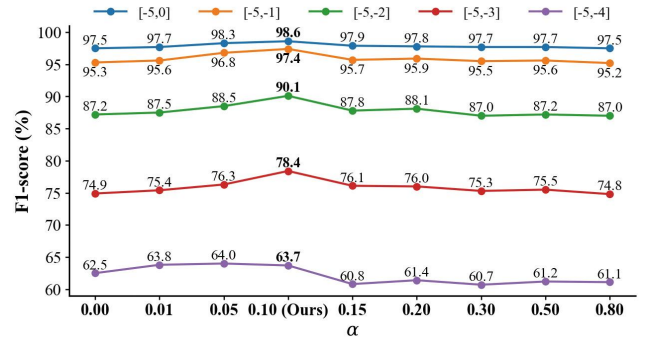


Figure 7: F1-scores (%) for different values of α in the loss function on the Brain4Cars dataset.

(f) anchor the model’s reasoning in semantically relevant regions, including the driver’s facial state inside the cabin and the key road environment outside, entirely consistent with real-world logic. This process vividly demonstrates how the model integrates critical cues to arrive at a reliable judgment, thereby substantiating the soundness of its decision-making.

Ablation Study

Effect of components in CaTFormer. To evaluate the contribution of each component in CaTFormer, we conduct systematic ablation studies, as shown in Table 4. Starting from a dual-stream Transformer baseline (Base) that performs late fusion via feature concatenation, we progressively add RDF, CRE, and FSN to measure their impact on F1-score. We further explored various combinations to explore module interactions and their cumulative effects. Experimental results confirm that each module improves intention prediction, with additional gains from their integration.

Effect of attention design in CaTFormer. To evaluate the effect of different attention mechanisms on driving intention prediction, we conducted ablation experiments on the Brain4Cars dataset comparing three attention schemes, as summarized in Fig. 6. The results indicate that our pro-

posed Bidirectional Dependency Attention (BDA) more effectively captures spatio-temporal correlations between interior and exterior streams while suppressing noise, thereby demonstrating its superiority in isolating dynamic cues and enhancing overall model robustness.

Effect of intention loss. We study the effect of the intention-loss weight α through an ablation analysis. As shown in Fig. 7, $\alpha = 0.10$ achieves the best F1-score, balancing temporal learning with intention supervision and providing informative gradients without destabilizing the main objective. Smaller α under-supervises subtle pre-action cues, whereas larger α induces gradient conflicts that degrade temporal coherence and anticipation accuracy.

Conclusion

In this paper, we introduced CaTFormer, a unified architecture that explicitly models causal interactions between driver behavior and environmental context for accurate intention prediction. Our approach extracts dual-stream features and merges them through a structured fusion pipeline. Extensive experiments on the Brain4Cars dataset confirm that CaTFormer achieves state-of-the-art accuracy, demonstrating its suitability for real-time driver assistance.

Acknowledgments

This work was supported in part by the Beijing Jiaotong University Research Fund under Grant KKA309004533.

References

- Gao, K.; Li, X.; Chen, B.; Hu, L.; Liu, J.; Du, R.; and Li, Y. 2023. Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment. *IEEE Transactions on Intelligent Transportation Systems*, 24(6): 6203–6216.
- Gebert, P.; Roitberg, A.; Haurilet, M.; and Stiefelwagen, R. 2019. End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 969–974.
- Guo, C.; Liu, H.; Chen, J.; and Ma, H. 2023. Temporal Information Fusion Network for Driving Behavior Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(9): 9415–9424.
- Hu, Y.; Jia, X.; Tomizuka, M.; and Zhan, W. 2022. Causal-based time series domain generalization for vehicle intention prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, 7806–7813. IEEE.
- Hu, Z.; Lv, C.; Hang, P.; Huang, C.; and Xing, Y. 2021. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE Transactions on Industrial Electronics*, 69(2): 1800–1808.
- Huang, Y.; Du, J.; Yang, Z.; Zhou, Z.; Zhang, L.; and Chen, H. 2022. A survey on trajectory-prediction methods for autonomous driving. *IEEE transactions on intelligent vehicles*, 7(3): 652–674.
- Jain, A.; Koppula, H. S.; Raghavan, B.; Soh, S.; and Saxena, A. 2015. Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models. arXiv:1504.02789.
- Jain, A.; Koppula, H. S.; Soh, S.; Raghavan, B.; Singh, A.; and Saxena, A. 2016a. Brain4Cars: Car That Knows Before You Do via Sensory-Fusion Deep Learning Architecture. arXiv:1601.00740.
- Jain, A.; Singh, A.; Koppula, H. S.; Soh, S.; and Saxena, A. 2016b. Recurrent Neural Networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 3118–3125.
- Jia, X.; You, J.; Zhang, Z.; and Yan, J. 2025. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Jiang, H.; Hu, C.; Niu, Y.; Yang, B.; Chen, H.; and Zhang, X. 2024. Hybrid Attention-based Multi-task Vehicle Motion Prediction Using Non-Autoregressive Transformer and Mixture of Experts. *IEEE Transactions on Intelligent Vehicles*.
- Li, C.; Chan, S. H.; and Chen, Y.-T. 2020. Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10711–10718. IEEE.
- Li, L.; Zhao, W.; and Wang, C. 2022. POMDP motion planning algorithm based on multi-modal driving intention. *IEEE Transactions on Intelligent Vehicles*, 8(2): 1777–1786.
- Liu, H.; Wu, C.; and Wang, H. 2023. Real time object detection using LiDAR and camera fusion for autonomous driving. *Scientific Reports*, 13(1): 8056.
- Liu, M.; Cheng, H.; Chen, L.; Broszio, H.; Li, J.; Zhao, R.; Sester, M.; and Yang, M. Y. 2024. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2039–2049.
- Liu, S.; Li, X.; Chen, J.; Guo, C.; Wu, J.; Luo, Q.; and Ma, H. 2025. Individualized Driving Intention Prediction With Inverse Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 26(6): 8125–8139.
- Ma, Y.; Ye, W.; Cao, X.; Abdelraouf, A.; Han, K.; Gupta, R.; and Wang, Z. 2023. Cemformer: Learning to predict driver intentions from in-cabin and external cameras via spatial-temporal transformers. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 4960–4966. IEEE.
- Mo, X.; Liu, H.; Huang, Z.; Li, X.; and Lv, C. 2023. Map-adaptive multimodal trajectory prediction via intention-aware unimodal trajectory predictors. *IEEE Transactions on Intelligent Transportation Systems*, 25(6): 5651–5663.
- Rekabdar, B.; and Mousas, C. 2018. Dilated Convolutional Neural Network for Predicting Driver’s Activity. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3245–3250.
- Rong, Y.; Akata, Z.; and Kasneci, E. 2020. Driver Intention Anticipation Based on In-Cabin and Driving Scene Monitoring. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–8.
- Sui, Z.; Zhou, Y.; Zhao, X.; Chen, A.; and Ni, Y. 2021. Joint intention and trajectory prediction based on transformer. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7082–7088. IEEE.
- Tonutti, M.; Ruffaldi, E.; Cattaneo, A.; and Avizzano, C. A. 2019. Robust and subject-independent driving manoeuvre anticipation through Domain-Adversarial Recurrent Neural Networks. *Robotics and Autonomous Systems*, 115: 162–173.
- Wu, K.; Zhou, Y.; Shi, H.; Li, X.; and Ran, B. 2023. Graph-based interaction-aware multimodal 2D vehicle trajectory prediction using diffusion graph convolutional networks. *IEEE Transactions on Intelligent Vehicles*, 9(2): 3630–3643.
- Zhou, D.; Liu, H.; Ma, H.; Wang, X.; Zhang, X.; and Dong, Y. 2021. Driving Behavior Prediction Considering Cognitive Prior and Driving Context. *IEEE Transactions on Intelligent Transportation Systems*, 22(5): 2669–2678.
- Zhu, Z.; Zhao, S.; Chu, C.; Wang, C.; Du, A.; and He, B. 2024. FedPRM: Federated Personalized Mixture Representation for Driver Intention Prediction. *IEEE Transactions on Intelligent Vehicles*, 1–14.