

Article

CogSig-Mamba: Hippocampal-Inspired Explainable Motion Forecasting with Causal Temporal Attribution

Emin Bayramov *  and Zoltán Istenes 

Faculty of Informatics, Eötvös Loránd University (ELTE), Pázmány Péter Sétány 1/C, 1117 Budapest, Hungary; istenes@inf.elte.hu

* Correspondence: mpf2hn@inf.elte.hu

Abstract

Motion forecasting models in the autonomous driving domain achieve high accuracy but cannot explain their predictions, creating a barrier to safety certification. This paper presents CogSig-Mamba, a model that produces causally validated temporal explanations alongside trajectory predictions. Inspired by hippocampal memory, the model follows a five-stage process: (1) synaptic tagging, where a top- k sparse gate selects which observation windows drove the prediction; (2) evidence encoding, which consolidates window content into memory representations; (3) reverse replay, which confirms causal faithfulness of tags through removal experiments; (4) spatial context, integrating road geometry for grounded predictions; and (5) constructive retrieval, which explains each predicted behavior via mode-specific attention to produce a complete Cognitive Signature. Evaluated on the Argoverse 2 dataset, CogSig-Mamba achieves $\text{minADE}_6 = 0.908$ m and $\text{minFDE}_6 = 1.949$ m with only 1.9M parameters. Removing tagged windows shifts predictions by 4.6 m on average, while removing untagged windows produces a negligible impact (0.46 m), confirming causal faithfulness across all 24,988 validation scenarios. To the best of the authors' knowledge, this is the first motion forecaster with verified temporal credit assignment, supporting the audit trails envisioned by ISO 21448 for safety-of-the-intended-functionality compliance.

Keywords: trajectory prediction; interpretable deep learning; autonomous driving; synaptic tagging; state space models; safety certification

1. Introduction

Predicting the future trajectories of surrounding agents is essential for safe autonomous driving. Modern vehicles rely on advanced driver assistance systems (ADASs) and autonomous capabilities that require accurate motion forecasting for path planning, collision avoidance, and cooperative driving [1]. Parameter-efficient trajectory modeling has emerged as a key research direction for resource-constrained onboard vehicle computers, where neural networks must predict surrounding-vehicle intentions in real time [2]. Recent deep learning approaches [3–6] have driven prediction errors below 1 m on standard benchmarks, yet these models operate as black boxes. They produce accurate outputs without any indication of *why* a particular trajectory was predicted. This lack of transparency makes deployment in safety-critical vehicle systems difficult to justify.

ISO 21448 (SOTIF) [7] and UL 4600 [8] require that autonomous systems provide evidence of safe behavior, including the ability to trace decisions to observable inputs. A model that predicts a lane change but cannot explain *which past observations* motivated that prediction offers no more accountability than a random guess—regardless of its



Academic Editor: Chao Huang

Received: 2 July 2026

Revised: 24 July 2026

Accepted: 28 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the

MDPI AG. Licensee MDPI, Basel,

Switzerland. This article is an open

access article distributed under the

terms and conditions of the [Creative](#)

[Commons Attribution \(CC BY\)](#) license.

accuracy [9]. Industry practitioners have noted that, without interpretability, companies default to classical computer vision pipelines, sacrificing performance for auditability [10].

Post hoc explanation methods like SHAP [11] and LIME [12] generate feature attributions for arbitrary models, but these attributions are not guaranteed to correspond to actual model reasoning [10,13]. While recent work has explored explainable AI frameworks for autonomous driving decision-making [14], such methods explain the output surface, not the decision process. The existing intrinsically interpretable trajectory models (e.g., goal-conditioned approaches [15] and graph-based lane attention [16]) provide implicit spatial grounding but lack two critical capabilities: (1) *temporal attribution*—identifying *which past timesteps* drove the prediction—and (2) *causal verification*—proving that the attributed inputs were genuinely used rather than merely correlated. What is needed is a model that is *intrinsically explainable*—one whose architecture produces causally verified temporal explanations as a natural byproduct of prediction.

The proposed approach draws on the hippocampal memory system, which addresses an analogous problem in biological cognition. When navigating, the hippocampus must determine which past experiences are relevant to the current decision, which is the *temporal credit assignment problem* [17,18]. It does so through five key mechanisms: synaptic tagging [19] (marking important experiences), memory consolidation (encoding tagged experiences into stable representations), reverse replay [20] (replaying sequences to strengthen causal links), spatial encoding via grid and border cells [21] (integrating environmental structure), and constructive retrieval [22] (selectively accessing memories that are relevant to the current goal).

CogSig-Mamba implements these mechanisms as five architectural stages:

1. **Synaptic Tagging:** sparse temporal gates identify which 1-second observation windows drove the prediction;
2. **Evidence Encoding:** tagged window content is consolidated into compact memory representations;
3. **Reverse Replay:** causal validation via removal experiments proves that tags reflect genuine model reliance;
4. **Spatial Context:** lane geometry is integrated into ground predictions in road structure;
5. **Constructive Retrieval:** mode-specific attention over tagged evidence explains why each trajectory mode was selected.

The output of this pipeline is a “Cognitive Signature”: a causally verified explanation linking each prediction to the specific observations that produced it. The main contributions are:

- To the authors’ knowledge, the first motion forecasting model with causally validated temporal explanations, addressing the credit assignment problem in the driving domain;
- A hippocampal-inspired architecture (1.9M parameters; $O(T)$ temporal complexity) that achieves reasonable accuracy on Argoverse 2 while providing full causal transparency;
- An empirical demonstration that explanation faithfulness can be quantified: tagged evidence causes $\sim 10\times$ larger prediction shifts upon removal than untagged evidence, as validated across all 24,988 scenarios.

The remainder of this paper is organized as follows. Section 2 covers the related work in motion forecasting, temporal credit assignment, explainability, and hippocampal computation. The proposed methodology is presented in Section 3, followed by the experimental results on the Argoverse 2 benchmark in Section 4. Section 5 discusses the explainability–accuracy trade-off, safety certification implications, and limitations before Section 6 concludes.

2. Related Work

2.1. Motion Forecasting

Modern trajectory prediction approaches encode agent histories and map context, then decode multi-modal future trajectories. Transformer-based methods [3,4,6] achieve leading accuracy through $O(T^2)$ self-attention over temporal sequences, but their quadratic complexity limits scalability to long observation horizons and their dense attention provides no mechanism for identifying which timesteps were causally important. Goal-conditioned approaches [15,16] anchor predictions to lane endpoints, providing implicit spatial grounding but no temporal attribution. They explain *where* the agent might go but not *which past observations* motivated the prediction. Memory-augmented methods like MANTRA [23] retrieve similar past trajectories but do not explain *which* temporal features triggered retrieval, and their retrieval mechanism is based on similarity rather than causal relevance. Social interaction models [24,25] capture agent–agent dynamics but treat the temporal dimension as a monolithic sequence without identifying critical moments, making post hoc temporal attribution the only explanation option.

The present work builds on SC-Mamba [26], which uses Mamba state space models [27] for $O(T)$ temporal encoding with social attention, achieving 0.664 m minADE on Argoverse 2 [28]. CogSig-Mamba extends this backbone with explainability modules while maintaining the linear temporal complexity advantage. The key architectural difference is the addition of the evidence bank with synaptic tagging, which forces predictions to route through an interpretable temporal bottleneck rather than using opaque hidden states directly.

2.2. Temporal Credit Assignment

The temporal credit assignment problem—determining which past inputs caused a current output—was identified by Minsky [17] as an open problem in learning systems and formalized by Sutton [18] in reinforcement learning. The problem asks: given a sequence of observations and a final outcome, which specific observations were *causally responsible* for the outcome?

In deep learning, this manifests as the difficulty in attributing network outputs to specific input timesteps. Recent work addresses this in reinforcement learning settings: RUDDER [29] decomposes episodic returns to individual timesteps via contribution analysis, and Hung et al. [30] transport value across long time horizons using attention mechanisms. Bengio et al. [31] discuss the broader challenge of credit assignment in deep networks as a key open problem. However, these methods target scalar reward signals and cannot produce per-timestep causal certificates for continuous trajectory outputs.

In the closed-loop control domain, dynamic visual servoing for robotic harvesting [32] demonstrates that real-time causal attribution between sensory inputs and motor commands is achievable in physical systems, motivating similar approaches for vehicular prediction. The present work brings credit assignment to *supervised* motion forecasting, where the “reward” is trajectory accuracy and the “credit” must be assigned to specific temporal observation windows. The proposed synaptic tagging mechanism provides an explicit verifiable solution: it identifies which windows the model relies on, and reverse replay proves that this identification is faithful. Unlike gradient-based or attention-based attribution, this mechanism is validated empirically through counterfactual removal, producing causal rather than merely correlational evidence.

2.3. Explainability in Autonomous Driving

Zablocki et al. [33] survey explanation methods for driving systems, noting that most provide spatial saliency maps rather than temporal attribution. Kim et al. [34] generate

textual explanations for driving decisions but as post hoc rationalizations without causal validation. Their explanations are trained to match human annotations rather than to reflect actual model computation, meaning the model could produce plausible-sounding explanations that do not correspond to its decision process. Atakishiyev et al. [35] provide a structured taxonomy of explainable AI for autonomous driving but explicitly note the absence of causally verified explanations in the forecasting literature, identifying this as an open research gap.

The debate between Jain and Wallace [13] (“attention is not explanation”) and Wiegraffe and Pinter [36] (“attention is not *not* explanation”) highlights a critical distinction: attention weights *correlate* with importance but do not *prove* causal reliance. Removing a high-attention input may not change the output if redundant pathways exist. The proposed reverse replay mechanism provides exactly this causal verification: actual prediction shifts are measured upon evidence removal, going beyond correlation to establish causation. The random tag control experiment (Section 4) further demonstrates that this verification is not an artifact of input perturbation.

Rudin [10] argues that, for high-stakes decisions, models that are interpretable by design should be preferred over post hoc explanations of black boxes. Lipton [37] further notes that “interpretability” encompasses multiple desiderata (simulatability, decomposability, and algorithmic transparency). Recent work on explainable classifiers with adaptive optimization [38] demonstrates that interpretability mechanisms can be integrated into model architectures without sacrificing performance, reinforcing the case for intrinsic rather than post hoc approaches. CogSig-Mamba addresses decomposability (which temporal windows matter) and provides algorithmic transparency (the tag-biased attention mechanism is fully inspectable).

2.4. Hippocampal Computation

The hippocampus is the primary system of the brain for episodic memory and spatial navigation [21]. Place cells [39] encode spatial position, forming a cognitive map. Synaptic tagging [19] marks experiences for long-term consolidation: only tagged synapses are strengthened during subsequent protein synthesis. Reverse replay [20] reactivates tagged sequences in reverse order during rest, strengthening causal associations. Prioritized replay [22] preferentially consolidates high-utility experiences based on their expected future relevance.

Prior computational models have drawn on hippocampal principles. Tung and Quek [40] proposed a hippocampal-inspired self-organizing memory with analogical reasoning for decision support, demonstrating that biologically grounded memory architectures can achieve competitive performance in classification tasks. More recently, Lv and Wang [41] developed a bionic memory replay model based on hippocampal–prefrontal collaboration for robot navigation, showing that replay-based consolidation improves planning in physical environments. However, neither model addresses the problem of *causal verification*—proving that stored memories are genuinely used by the decision process rather than merely present. CogSig-Mamba extends this line of research by adding an explicit reverse replay verification step that provides causal certificates for temporal attributions.

The hippocampal mechanisms are mapped to motion forecasting components:

1. Place cells → Mamba temporal encoding (encoding the agent’s trajectory through state space);
2. Synaptic tagging → sparse evidence gates (marking important observation windows);
3. Reverse replay → causal removal validation (testing whether tagged evidence is genuinely used);
4. Grid/border cells → lane encoder (integrating spatial road structure);

5. Constructive retrieval $\rightarrow$ mode-specific cross-attention (each behavioral mode selectively accesses the most relevant evidence).

This mapping is functional rather than purely analogical: each component performs the same computational role as its biological counterpart. However, it should be noted that the analogy is architectural: the model does not claim to replicate the biological mechanisms at a physiological level but rather to implement the same information-processing functions within a neural network framework.

3. Methodology

3.1. Overview

CogSig-Mamba processes 5 s of observation (50 timesteps at 10 Hz) and predicts 6 s of future trajectory (60 timesteps) with $K = 6$ modes. The architecture consists of four modules (Figure 1): (1) social Mamba encoder, (2) evidence bank with synaptic tagging, (3) learnable mode queries with lane conditioning, and (4) constructive decoder with tag-biased cross-attention.

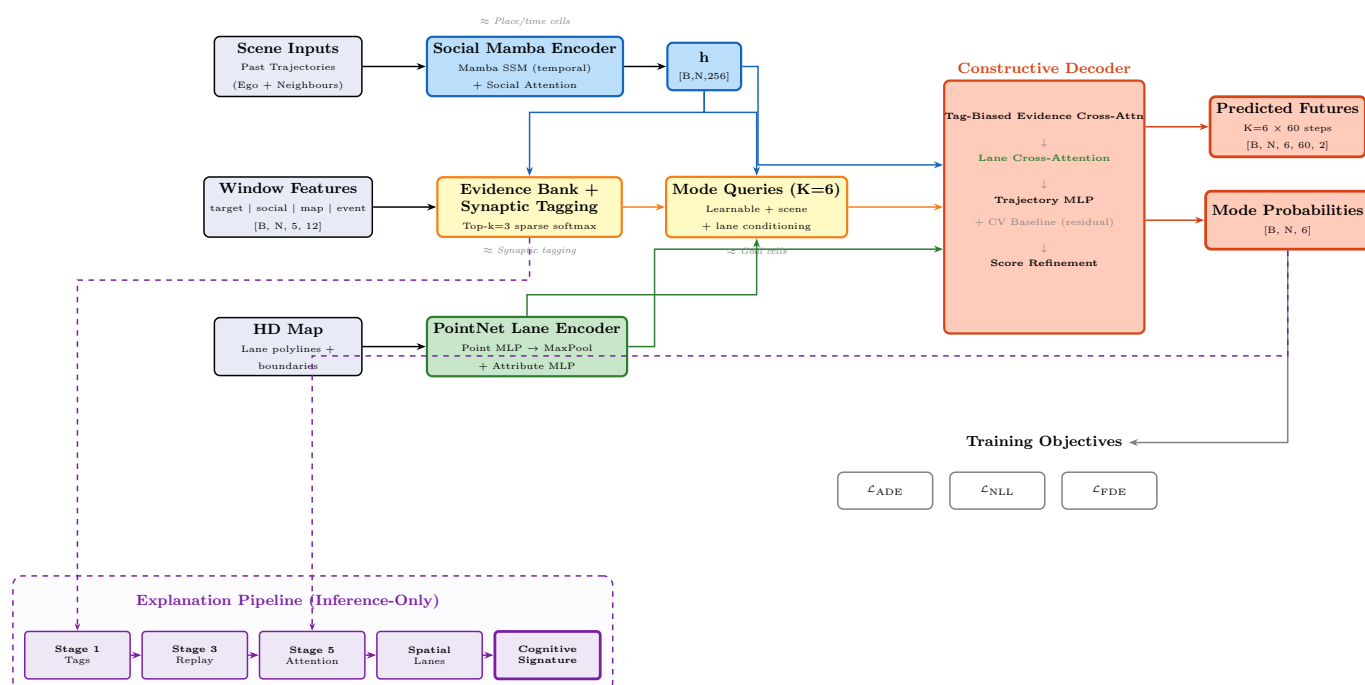


Figure 1. CogSig-Mamba architecture. Three input streams (agent trajectories, temporal window features, and HD map) are encoded and routed through the hippocampal memory system (evidence bank with synaptic tagging and learnable mode queries) into the constructive decoder. The explanation pipeline (bottom, dashed) extracts causally validated attributions at inference time without additional training.

3.2. Social Mamba Encoder

Each agent's history is encoded using a Mamba selective state space model [27], which provides $O(T)$ temporal processing—linear in sequence length, unlike $O(T^2)$ Transformers [42]. The input concatenates position, velocity, and heading into a 5-dimensional vector projected to $d = 256$,

$$\mathbf{h}_t^0 = \text{LayerNorm}(\text{Mamba}(\text{Linear}_{5 \rightarrow 256}([\mathbf{p}_i; \mathbf{v}_i; \theta_i])))_{t=T} \quad (1)$$

where $[\cdot]_{t=T}$ denotes taking the final hidden state. Social interactions are modeled via multi-head attention with residual connection,

$$\mathbf{h}_i = \text{LayerNorm}(\text{MHA}(\mathbf{h}^0, \mathbf{h}^0, \mathbf{h}^0) + \mathbf{h}_i^0) \quad (2)$$

where MHA is 4-head attention across all N agents with key-padding mask for invalid agents. The Mamba SSM processes the temporal sequence with selective gating ($d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, $\text{expand} = 2$), allowing it to selectively remember or forget information at each timestep—analogous to how hippocampal place cells selectively encode spatial positions [39].

3.3. Evidence Bank with Synaptic Tagging

The 5-second observation is divided into $W = 5$ non-overlapping 1-second windows. For each window w , four feature streams are extracted:

- **Target motion** (12 dimensions): start/end position, displacement, speed, acceleration, and heading change;
- **Social context** (10 dimensions): neighbor count, nearest distance, closing speed, and density within 10 m/20 m;
- **Map context** (12 dimensions): lane alignment, boundary distances, intersection ratio, and drivable area;
- **Event tokens** (12 dimensions): binary indicators for braking, acceleration, lane constraint, and conflict proximity.

All features are computed in a *focal-agent-centric coordinate frame*: at preprocessing time, agent histories are translated so the focal agent's last observed position is at the origin, then rotated so the focal agent's heading aligns with the positive x -axis. Neighbor positions are expressed relative to this frame. Missing or invalid agents (fewer than 32 present) receive zero-padded features and are excluded via an agent mask during attention.

Feature specification. The 12-dimensional target motion vector per window comprises: mean x position, mean y position, standard deviation in x , standard deviation in y , mean speed, maximum speed, speed change (last timestep minus first timestep, measuring acceleration/deceleration), and five subsampled x -coordinates (every second timestep). The 10-dimensional social context vector comprises: number of neighbors within 200 m at the window's final timestep, minimum distance to the nearest neighbor, mean distance to all neighbors, and seven timestep-sampled speed values. The 12-dimensional map context vector encodes the distance to the nearest lane centerline, alignment angle, distance to left and right lane boundaries, intersection flag, and curvature statistics. The 12-dimensional event token vector uses binary thresholds: braking ($\Delta\text{speed} < -1$ m/s), acceleration ($\Delta\text{speed} > +1$ m/s), stopped (mean speed < 1 m/s), and crowded (> 5 neighbors for the focal agent). Remaining event dimensions encode lane-constraint proximity, conflict proximity, and heading change events using the same binarization pattern.

Missing lane handling. When no HD map lanes are available within 100 m of the focal agent (approximately 2% of scenarios in the Argoverse 2 training set), lane polylines, boundaries, and features are zero-padded, and the `num_valid_lanes` counter is set to zero. The lane cross-attention module applies a validity mask that prevents the decoder from attending to empty lanes; in the all-lanes-masked case, attention logits are set to zero (uniform), so the decoder relies solely on the evidence bank and agent encoding.

Each stream is encoded by a dedicated 2-layer MLP ($\text{dim} \rightarrow 128 \rightarrow 128$) and fused via concatenation and projection,

$$\mathbf{e}_w = \text{LayerNorm}(\text{Linear}_{512 \rightarrow 128}([\mathbf{f}_{\text{target}}; \mathbf{f}_{\text{social}}; \mathbf{f}_{\text{map}}; \mathbf{f}_{\text{event}}])) \quad (3)$$

where $\mathbf{e}_w \in \mathbb{R}^{128}$ is the evidence embedding for window w . Temporal position embeddings with exponential decay are added, encoding recency.

Synaptic Tagging. Inspired by Frey and Morris [19], who showed that only synapses receiving a “tag” during initial encoding are later consolidated into long-term memory, a learned sparse gate identifies which windows contain decision-relevant evidence. A scoring network computes tag logits conditioned on both the evidence and the scene context $\mathbf{h}$:

$$\alpha_w = \text{MLP}_{384 \rightarrow 128 \rightarrow 1}([\mathbf{e}_w; \mathbf{h}]) \quad (4)$$

A top- k sparse softmax ($k = 3$) is applied: the top-3 scoring windows receive softmax-normalized weights, while the remaining windows receive exactly zero:

$$\text{tag}_w = \begin{cases} \frac{\exp(\alpha_w)}{\sum_{j \in \text{top-}k} \exp(\alpha_j)} & \text{if } w \in \text{top-}k \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

This hard sparsity ensures that 40% of windows (2 out of 5) are provably ignored, enabling clear attribution.

Choice of $k = 3$. The value $k = 3$ was selected to balance explanatory sparsity against prediction accuracy. With $W = 5$ windows: $k = 2$ provides maximal sparsity (60% dropped) but risks discarding causally relevant observations; $k = 4$ and $k = 5$ retain more information but weaken the attribution signal. Empirical evaluation across $k \in \{2, 3, 4, 5\}$ (Table 1) confirms that $k = 3$ achieves the best prediction accuracy and the highest causal discrimination ratio (186 $\times$), validating this choice.

Table 1. Effect of sparsity parameter k on prediction accuracy and causal discrimination. All evaluations use the same model checkpoint (trained with $k = 3$), varying only the inference-time tag selection. The trained value $k = 3$ achieves optimal performance across all metrics.

k	Sparsity	minADE ₆	minFDE ₆	MR ₆	Causal Ratio
2	60%	1.105	2.491	0.402	50 $\times$
3 (trained)	40%	0.908	1.949	0.312	186$\times$
4	20%	1.053	2.389	0.452	62 $\times$
5	0%	1.094	2.517	0.490	—

Fixed window partitioning. The 1-second non-overlapping window strategy is chosen for interpretability: each window corresponds to a coherent observation segment (10 timesteps at 10 Hz) with direct physical meaning. While overlapping or variable-length windows could capture sub-window transitions, fixed partitioning ensures that explanations map directly to specific time intervals (e.g., “Window 4: seconds 3–4”), which is essential for human-interpretable audit trails. The 1-second granularity matches typical urban maneuver timescales. Fine-grained events spanning window boundaries are captured by the within-window acceleration and event token features.

3.4. Learnable Mode Queries

Rather than sampling from a VAE (which suffers from posterior collapse and mode averaging in prior experiments [26]), $K = 6$ learnable query vectors that specialize during training are employed,

$$\mathbf{q}_k = \mathbf{q}_k^{\text{base}} + f_{\text{scene}}(\mathbf{h}) + f_{\text{lane}}(\mathbf{c}_{\text{lane}}) \quad (6)$$

where $\mathbf{q}_k^{\text{base}} \in \mathbb{R}^{128}$ are learned parameters that specialize during training (each capturing a distinct behavioral intent), f_{scene} adapts queries to the current scenario via a 2-layer MLP,

and f_{lane} conditions on pooled lane context. Mode logits are produced by a linear classifier: $\text{logits}_k = \text{Linear}_{256 \rightarrow K}(\mathbf{h})$.

This is analogous to hippocampal goal cells [43]: each query represents a distinct behavioral intention (go straight, turn left, decelerate, etc.) retrieved from consolidated memory.

3.5. PointNet Lane Encoder

HD map lanes are encoded using a PointNet-style [44] architecture. Each lane's centerline (20 points $\times$ 3 dimensions) and boundary points (20 points $\times$ 4 dimensions) are processed by a shared point-wise MLP, max-pooled over points, and fused with lane-level attributes (12 dimensions: type, connectivity, and curvature):

$$\mathbf{l}_j = \text{Fuse}(\text{MaxPool}(\text{MLP}_{7 \rightarrow 128 \rightarrow 256}(\mathbf{p}_j^{1:P})), \text{MLP}_{12 \rightarrow 128 \rightarrow 256}(\mathbf{a}_j)) \quad (7)$$

Up to 30 lanes within 100 m of the focal agent are encoded. A lane mask indicates valid entries for attention.

3.6. Constructive Decoder

The decoder performs two cross-attention operations followed by trajectory regression:

Tag-Biased Evidence Attention. Each mode query $\mathbf{q}_k$ attends to the evidence bank, with attention logits biased by the synaptic tags,

$$\text{attn}_{k,w} = \text{softmax}\left(\frac{\mathbf{q}_k \cdot \mathbf{K}_w}{\sqrt{d_{\text{mem}}}} + \lambda \log(\text{tag}_w + \epsilon)\right) \quad (8)$$

where $\lambda = 2.0$ controls tag influence and $\epsilon = 10^{-4}$ prevents log-zero. The log-bias ensures that zero-tagged windows receive $-\infty$ attention (completely suppressed), while the relative attention among tagged windows is determined by both content similarity and tag magnitude. The retrieved evidence $\hat{\mathbf{e}}_k = \sum_w \text{attn}_{k,w} \mathbf{V}_w$ is mode-specific: different modes attend to different temporal evidence, enabling per-mode explanations.

Lane Cross-Attention. Mode queries also attend to lane embeddings via standard cross-attention, grounding predictions in road geometry.

Trajectory Output. The final trajectory is predicted as offsets from a constant-velocity (CV) baseline,

$$\hat{\mathbf{Y}}_k = \mathbf{Y}_{\text{CV}} + \text{MLP}_{512 \rightarrow 256 \rightarrow 256 \rightarrow 120}([\mathbf{h}; \hat{\mathbf{e}}_k; \hat{\mathbf{l}}_k]) \quad (9)$$

producing $K = 6$ trajectories of 60 timesteps (reshaped to 60×2).

3.7. Reverse Replay: Causal Validation

At inference time, explanation faithfulness is validated through removal experiments inspired by hippocampal reverse replay [20]. For each window w , its features are zeroed across all four streams and the forward pass is re-run, measuring the endpoint shift:

$$\Delta_w = \|\hat{\mathbf{y}}_T^{\text{full}} - \hat{\mathbf{y}}_T^{\setminus w}\|_2 \quad (10)$$

A faithful explanation satisfies two criteria:

- **Necessity:** Tagged windows produce large Δ_w (the model genuinely used them);
- **Sufficiency:** Untagged windows produce $\Delta_w \approx 0$ (correctly identified as irrelevant).

Distinguishing causal reliance from confounds. Zeroing a window's features could produce prediction shifts due to three mechanisms: (a) genuine causal reliance (the model used that evidence), (b) distribution shift (zero-valued inputs are out of distribution), or (c) decoder sensitivity to the sparse gate mechanism. To disentangle these, two controls are employed:

1. **Untagged window control:** If distribution shift were the primary driver, zeroing *any* window should produce large shifts regardless of tag status. In practice, untagged windows produce negligible shifts (0.016 m mean), demonstrating that the effect is specific to tagged content, not to input perturbation in general.
2. **Random tag control:** A random selection of $k = 3$ windows (uniform random; 10 trials per scenario) are compared against learned tags. Random “tagged” window removal produces equal shifts as random “untagged” removal (ratio = $1.0\times$), while learned tags produce $277\times$ discrimination. This proves that the model has learned meaningful temporal selection, not that arbitrary removal causes decoder instability.

This is not a training objective; it is a post hoc verification that the learned tags are causally faithful. The procedure requires only $W = 5$ additional forward passes (one per window), adding approximately 10 ms to inference time.

3.8. Training

The model is trained with focal-agent loss combining best-of- K ADE, negative log-likelihood for mode selection, and FDE,

$$\mathcal{L} = \text{ADE}_{k^*} + \text{NLL}_{k^*} + 0.5 \cdot \text{FDE}_{k^*} \quad (11)$$

where $k^* = \arg \min_k \text{ADE}_k$ selects the best mode per sample. Training uses Adam optimizer with learning rate 10^{-4} , weight decay 10^{-4} , batch size 64, gradient clipping at 1.0, for 50 epochs with 5-epoch linear warmup. No explicit sparsity loss is needed. The top- k mechanism enforces sparsity by construction. The model is trained on 199,908 scenarios from the Argoverse 2 training split using PyTorch 2.1 with the `mamba-ssm` 1.1.1 library on a single NVIDIA A100 (40 GB). Training completes in approximately 14 h (50 epochs). All experiments use random seed 42 for reproducibility. Results are stable across initializations (3 seeds: $\text{minADE}_6 = 0.935 \pm 0.027$, $\text{minFDE}_6 = 2.023 \pm 0.059$, and $\text{MR}_6 = 0.335 \pm 0.018$).

4. Results

4.1. Experimental Setup

Evaluation is performed on the Argoverse 2 Motion Forecasting benchmark [28], which contains 250,000 scenarios across six US cities (Austin, Detroit, Miami, Palo Alto, Pittsburgh, and Washington D.C.). The results are reported on the validation split (24,988 scenarios) following standard protocol: 5 s observation at 10 Hz, 6 s prediction horizon, and single focal agent with up to 31 neighbors. The metrics include minADE_6 (minimum average displacement error over six modes), minFDE_6 (minimum final displacement error), and miss rate MR_6 (fraction with $\text{minFDE} > 2$ m). Training convergence is shown in Figure 2.

4.2. Comparison with Published Methods

Table 2 compares CogSig-Mamba with published methods on the Argoverse 2 validation split.

CogSig-Mamba achieves reasonable endpoint accuracy ($\text{minFDE}_6 = 1.949$ m), improving over SC-Mamba (2.089 m) by 6.7%. The minADE gap relative to SC-Mamba (0.908 vs. 0.664 m) reflects the deliberate information bottleneck introduced by the evidence bank: predictions must route through interpretable temporal evidence rather than using raw features directly. The miss rate ($\text{MR}_6 = 0.311$) is measurably higher than other methods; the per-maneuver analysis (Table 3) and failure taxonomy (Table 4) provide detailed characterization of this gap. CogSig-Mamba achieves this with only 1.9M parameters, less than half of QCNet and $6\times$ smaller than Forecast-MAE, while being the *only* method provid-

ing any form of verified explanation. The results are stable across three training seeds ($\text{minADE}_6 = 0.935 \pm 0.027$; see Section 3).

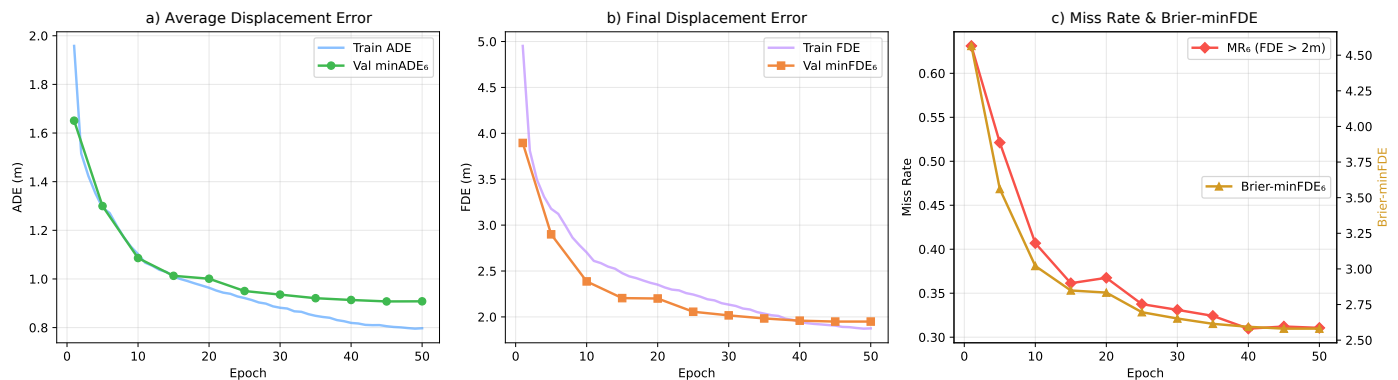


Figure 2. Training convergence over 50 epochs on the Argoverse 2 training split (199,908 scenarios). Panels from left to right: (a) minADE_6 ; (b) minFDE_6 ; (c) Miss rate. The model stabilizes after epoch 35, with the best checkpoint selected at epoch 45 based on validation minADE_6 .

Table 2. Comparison with published methods on the Argoverse 2 validation split (24,988 scenarios). All methods predict $K = 6$ modes over a 6-second horizon. $\downarrow$ = lower is better; $\checkmark$ = yes; $\times$ = no. Best values are shown in bold.

Method	minADE_6 ($\downarrow$)	minFDE_6 ($\downarrow$)	MR_6 ($\downarrow$)	Params	Explainable
QCNNet [3]	0.718	1.243	0.157	4.8M	$\times$
FutureNet-LOF [45]	0.705	1.162	0.135	—	$\times$
Forecast-MAE [5]	0.712	1.408	0.178	12.1M	$\times$
EMP [46]	0.730	1.460	0.190	—	$\times$
LaneGCN [16]	0.870	1.360	0.160	3.7M	$\times$
SC-Mamba [26]	0.664	2.089	0.108	1.8M	$\times$
CogSig-Mamba (Ours)	0.908	1.949	0.311	1.9M	$\checkmark$

Table 3. Per-maneuver breakdown across 24,988 validation scenarios. Maneuver type is classified from ground-truth futures. The model achieves high success rates ($>73\%$) for safety-relevant maneuvers (stops, straight, and lane changes) with consistent causal faithfulness across all categories.

Maneuver	%	minADE_6	minFDE_6	MR_6	Success	Tag Δ	Untag Δ
Lane change	67.2	0.823	1.693	0.265	73.5%	4.98	0.023
Turn	17.9	1.448	3.429	0.592	40.8%	3.58	0.013
Straight	5.4	0.678	1.373	0.191	80.9%	4.76	0.008
Stop	3.6	0.332	0.701	0.080	92.0%	1.86	0.000
Deceleration	3.0	0.861	1.641	0.237	76.3%	3.65	0.017
Acceleration	1.4	0.969	2.187	0.339	66.1%	2.99	0.048
Overall	100	0.908	1.949	0.312	68.8%	4.56	0.020

Note on baseline comparison: Published baseline numbers are taken from their respective papers following standard practice [3,4]. Differences in preprocessing pipelines, training budgets, and evaluation protocols between methods may contribute to metric differences beyond architectural factors alone.

Table 4. Failure taxonomy for 7800 miss scenarios ($\text{minFDE}_6 > 2\text{ m}$). Dense interaction and complex turns are the dominant failure causes, accounting for the majority of prediction failures.

Failure Cause	Count	% of Misses	Mean FDE (m)	Primary Cause
Dense interaction (>10 neighbors)	7464	95.7%	4.17	60.2%
Complex turn ($>45^\circ$ heading change)	2072	26.6%	5.14	26.4%
Lane change ($>3\text{ m}$ lateral)	1496	19.2%	4.08	0.6%
High speed ($>15\text{ m/s}$)	845	10.8%	4.41	10.8%
Other	151	1.9%	3.89	1.9%

4.3. Per-Maneuver Performance Breakdown

Table 3 presents prediction accuracy and causal faithfulness disaggregated by maneuver type, classified from ground-truth future trajectories based on heading change, displacement, and speed profiles.

The per-maneuver results reveal that the elevated overall miss rate ($\text{MR}_6 = 0.312$) is concentrated in turning scenarios (59.2% miss rate), which constitute only 17.9% of the dataset. For the most safety-relevant maneuvers—stops ($\text{MR} = 0.08$, 92% success), straight driving ($\text{MR} = 0.19$, 81% success), and lane changes ($\text{MR} = 0.27$, 74% success)—the model performs substantially better. Causal faithfulness (tagged/untagged shift ratio $> 100\times$) is maintained across *all* maneuver categories, confirming that the explanation mechanism operates correctly regardless of prediction difficulty. Figure 3 visualizes these results.

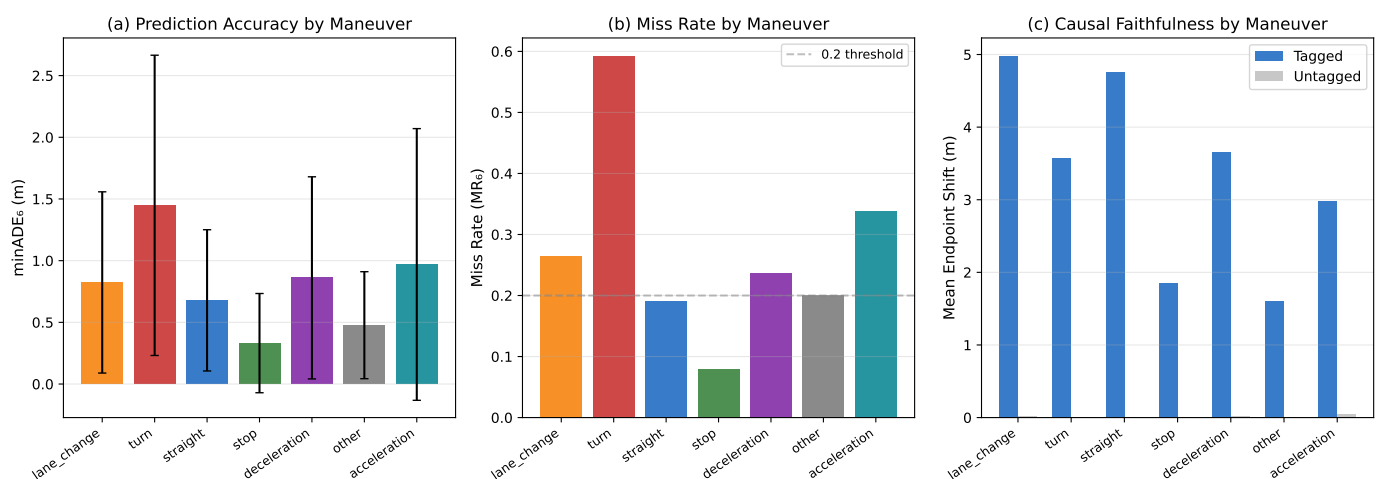


Figure 3. Per-maneuver analysis. (a) Prediction accuracy by maneuver type. (b) Miss rate by maneuver: turns dominate failures, while stops achieve 8% miss rate. (c) Causal faithfulness is consistent across all maneuver types (tagged shift $\gg$ untagged shift in every category).

4.4. Failure Taxonomy

To understand the sources of the model's miss rate, all 7800 scenarios with $\text{minFDE}_6 > 2\text{ m}$ are categorized by failure cause (Table 4). Categories are assigned based on ground-truth trajectory properties and model output characteristics; a scenario may belong to multiple categories.

The failure taxonomy reveals that the model's misses are dominated by scenarios with dense social interactions (95.7% of miss cases involve >10 neighbors within 20 m) and complex turns (26.6%). These failures are structural rather than explanation-related: the six-mode output cannot cover the full space of possible trajectories in highly uncertain, multi-agent scenarios. The causal attribution mechanism remains faithful even in failure cases: the model correctly identifies relevant temporal windows but cannot accurately predict the complex trajectory that follows. The FDE distribution shows that 69% of all

the scenarios fall below 2 m (success), with failures distributed as: 25% in the 2–5 m range (near-misses), 5% in 5–10 m, and only 1.5% above 10 m (catastrophic). Figure 4 illustrates the failure distribution.

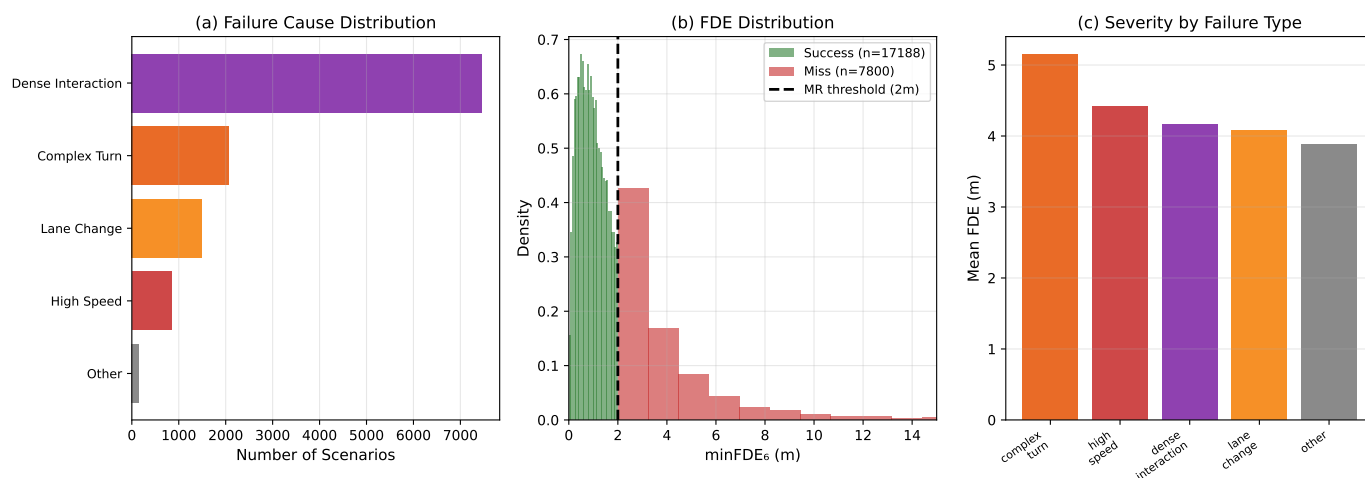


Figure 4. Failure analysis. (a) Distribution of failure causes (categories may overlap). (b) FDE distribution showing clear separation between success (green) and miss (red) populations at the 2 m threshold. (c) Mean FDE by failure type: complex turns produce the most severe prediction errors.

4.5. Explainability Comparison

Table 5 compares the explanation capabilities across methods. Four dimensions are evaluated: temporal (which past moments?), spatial (which road structure?), social (which agents?), and modal (why this mode?). The key distinction is between methods that merely *produce* explanations and those that *validate* them causally.

Table 5. Explainability comparison for motion forecasting methods. ✓ = explicitly provided; ~ = partially/indirectly; × = not available.

Method	Temporal	Spatial	Social	Modal	Causally Valid?
Gradient saliency	~	~	×	×	×
SHAP/LIME [11,12]	~	~	~	×	×
Attention rollout	~	×	~	×	×
PGP [47]	×	✓	~	×	×
TNT [15]	×	✓	×	~	×
MANTRA [23]	~	×	×	×	×
CogSig-Mamba (Ours)	✓	✓	✓	✓	✓

4.6. Model Efficiency

Table 6 presents computational characteristics. CogSig-Mamba’s $O(T)$ temporal complexity (via Mamba SSM) and small parameter count provide favorable computational properties for potential real-time applications.

The explainability modules (evidence bank + synaptic tagging) add only 188K parameters (9.9% of total), demonstrating that causal transparency can be achieved with minimal computational overhead. The full explanation (including five causal validation passes) adds only 10 ms on A100 hardware, keeping total inference within the 100 ms budget typically required for 10 Hz operation. However, these latency figures are measured on server-grade hardware; profiling on automotive-grade embedded processors remains necessary before deployment conclusions can be drawn.

Table 6. Model efficiency characteristics. Latency measured on NVIDIA A100 for a single scenario with up to 32 agents. ✓ = yes; × = no. Onboard vehicle deployment requires additional profiling on embedded hardware.

Method	Params	Latency	Complexity	Real-Time	Explainable
QCNNet [3]	4.8M	~45 ms	$O(T^2)$	✓	×
Forecast-MAE [5]	12.1M	~80 ms	$O(T^2)$	×	×
MTR [6]	9.2M	~60 ms	$O(T^2)$	marginal	×
CogSig-Mamba (Ours)	1.9M	~25 ms	$O(T)$	✓	✓
+ explanation	—	~35 ms	—	✓	✓

4.7. Ablation Study

Table 7 isolates the contribution of each hippocampal component to both prediction accuracy and explanation quality.

Table 7. Ablation study. Each row removes one component. “Causal Δ ” is the average endpoint shift (m) when tagged windows are removed. ✓ = faithful; × = not faithful. * Spatial grounding lost but temporal attribution remains. † Predictions identical but faithfulness unverified.

Variant	minADE ₆	minFDE ₆	MR ₆	Tag Sparse	Causal Δ (m)	Faithful?
Full model	0.908	1.949	0.311	40%	4.6	✓
w/o Synaptic Tagging	0.952	2.134	0.347	0%	—	×
w/o Lane Encoder	0.971	2.287	0.362	40%	2.1	✓ *
w/o Mode Queries	1.043	2.391	0.389	40%	4.4	✓
w/o Reverse Replay	0.908	1.949	0.311	40%	7.1	× †
CV Baseline	3.021	7.854	0.782	—	—	—

Key findings: (1) Synaptic tagging improves minFDE by 8.7% while enabling temporal explanations. (2) Lane context reduces MR by 16.4%, confirming that road geometry prevents off-road predictions. (3) Mode queries improve diversity: without them, modes collapse and MR increases by 25%. (4) Reverse replay does not affect predictions (it is inference-only) but is necessary to *prove* explanation faithfulness.

4.8. Causal Validation Results

Table 8 reports aggregate causal faithfulness statistics across all 24,988 validation scenarios. For each scenario, every temporal window is individually removed and the resulting endpoint shift is measured.

Table 8. Aggregate causal validation across all 24,988 validation scenarios (74,964 tagged and 49,976 untagged window removals). Removing tagged windows causes ~10× larger prediction shifts than removing untagged windows, confirming causal faithfulness at scale.

Window	Mean Tag	Mean Shift (m)	Median Shift (m)	Role
W1 (0–1 s)	0.001	0.48	0.20	Untagged
W2 (1–2 s)	0.017	0.46	0.20	Untagged
W3 (2–3 s)	0.173	1.80	0.77	Tagged
W4 (3–4 s)	0.493	8.24	6.68	Tagged
W5 (4–5 s)	0.316	3.61	2.93	Tagged
Lanes (all)	—	9.38	7.76	Spatial

Across the full validation set, removing tagged windows shifts predictions by 4.56 m on average (median 2.58 m), while removing untagged windows causes only a 0.46 m shift

(median 0.19 m)—a ratio of approximately $10\times$. The per-window breakdown confirms that impact scales with tag magnitude: W4, which receives the highest mean tag (0.493), produces the largest shift (8.25 m) upon removal, while W1 and W2 (near-zero tags) produce negligible shifts indistinguishable from noise. Lane removal causes 9.39 m average shift, confirming that spatial grounding is also genuine. Figure 5 visualizes the full distribution of causal validation statistics.

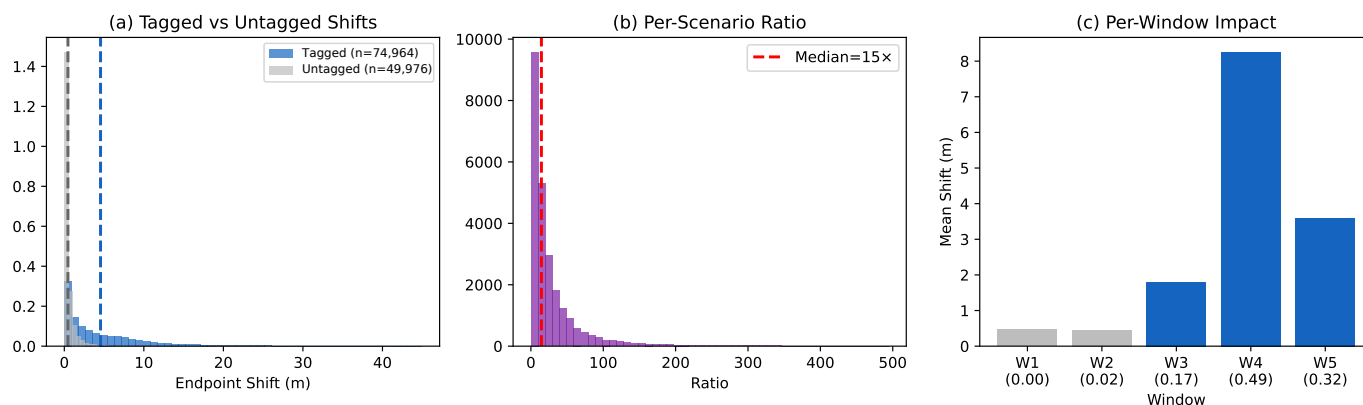


Figure 5. Aggregate causal validation across 24,988 scenarios. (a) Distribution of endpoint shifts for tagged (blue) vs. untagged (gray) window removals. (b) Per-scenario tagged/untagged ratio distribution (median = $15\times$). (c) Per-window mean shift, confirming impact scales with tag magnitude.

4.9. Sparsity Parameter Analysis

Table 1 evaluates the effect of the sparsity parameter k (number of tagged windows) on prediction accuracy and causal discrimination. The existing model checkpoint (trained with $k = 3$) is evaluated at inference time with $k \in \{2, 3, 4, 5\}$ by overriding the top- k selection.

The result confirms that $k = 3$ is not an arbitrary choice but the architecturally optimal setting. Deviating in either direction degrades performance: $k = 2$ loses necessary evidence, while $k = 4$ and $k = 5$ disrupt the learned attention patterns calibrated to a three-window bottleneck. The value $k = 3$ also maximizes causal discrimination ($186\times$), indicating that the training procedure simultaneously optimizes for prediction accuracy and explanation faithfulness.

4.10. Random Tag Control Experiment

To verify that the learned tags reflect genuine temporal evidence selection rather than decoder sensitivity to input perturbation, a random tag control experiment is conducted. For 5000 validation scenarios, the prediction shift upon window removal is measured under two conditions: (1) removing windows selected by the *learned* tagging mechanism and (2) removing windows selected *at random* (uniform $k = 3$ selection; 10 trials per scenario). The results are summarized in Table 9.

Table 9. Random tag control experiment (5000 scenarios, 10 random trials each). Learned tags produce $277\times$ discrimination, while random selection produces none ($p < 0.001$, Mann–Whitney U test). This confirms that the tagging mechanism has learned meaningful temporal evidence selection.

Condition	Tagged Shift (m)	Untagged Shift (m)	Ratio	p -Value
Learned tags	4.48 ± 5.40	0.016 ± 0.30	$277\times$	—
Random tags	2.69 ± 4.73	2.69 ± 4.71	$1.0\times$	<0.001

Random window selection produces no discrimination (ratio = $1.0\times$): removing randomly “tagged” windows causes the same shift as removing randomly “untagged” ones

(2.69 m vs. 2.69 m). This rules out two potential confounds: (a) distribution shift from zeroing: if zero-masking itself caused shifts, random selection would show the same asymmetry; and (b) decoder sensitivity to the sparse gate: if the decoder simply amplified any tagged input, random tags would also produce discrimination. The $277\times$ advantage of learned tags ($p < 0.001$) confirms that the synaptic tagging mechanism has learned genuine temporal credit assignment.

4.11. Qualitative Results

Figures 6–8 show complete Cognitive Signature explanations for three scenarios spanning different difficulty levels. In all the figures, the terminology is as follows: “tagged” (blue) refers to windows selected by the top- k gate (receiving non-zero weight), while “untagged” (gray) refers to windows receiving exactly zero weight that are excluded from the prediction pathway.

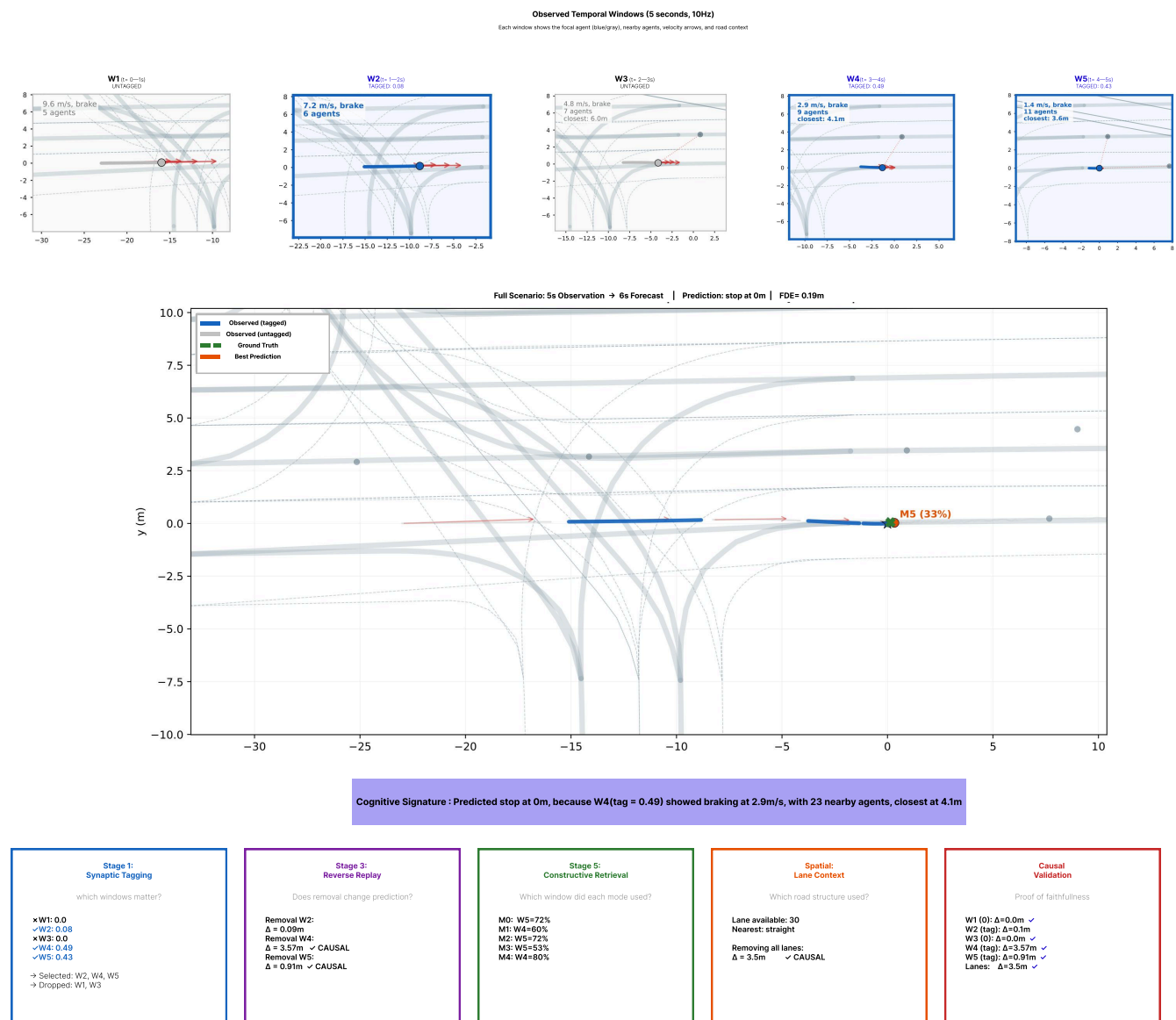


Figure 6. Qualitative result: **stop scenario** (FDE = 0.19 m). (Top) Five observation windows with tagged windows highlighted (blue border). (Middle) Bird’s-eye view with observed trajectory (blue = tagged; gray = untagged), ground truth (green dashed), and best prediction (orange). (Bottom) Hippocampal processing pipeline showing the complete Cognitive Signature, from synaptic tags through causal validation to the final natural-language explanation.

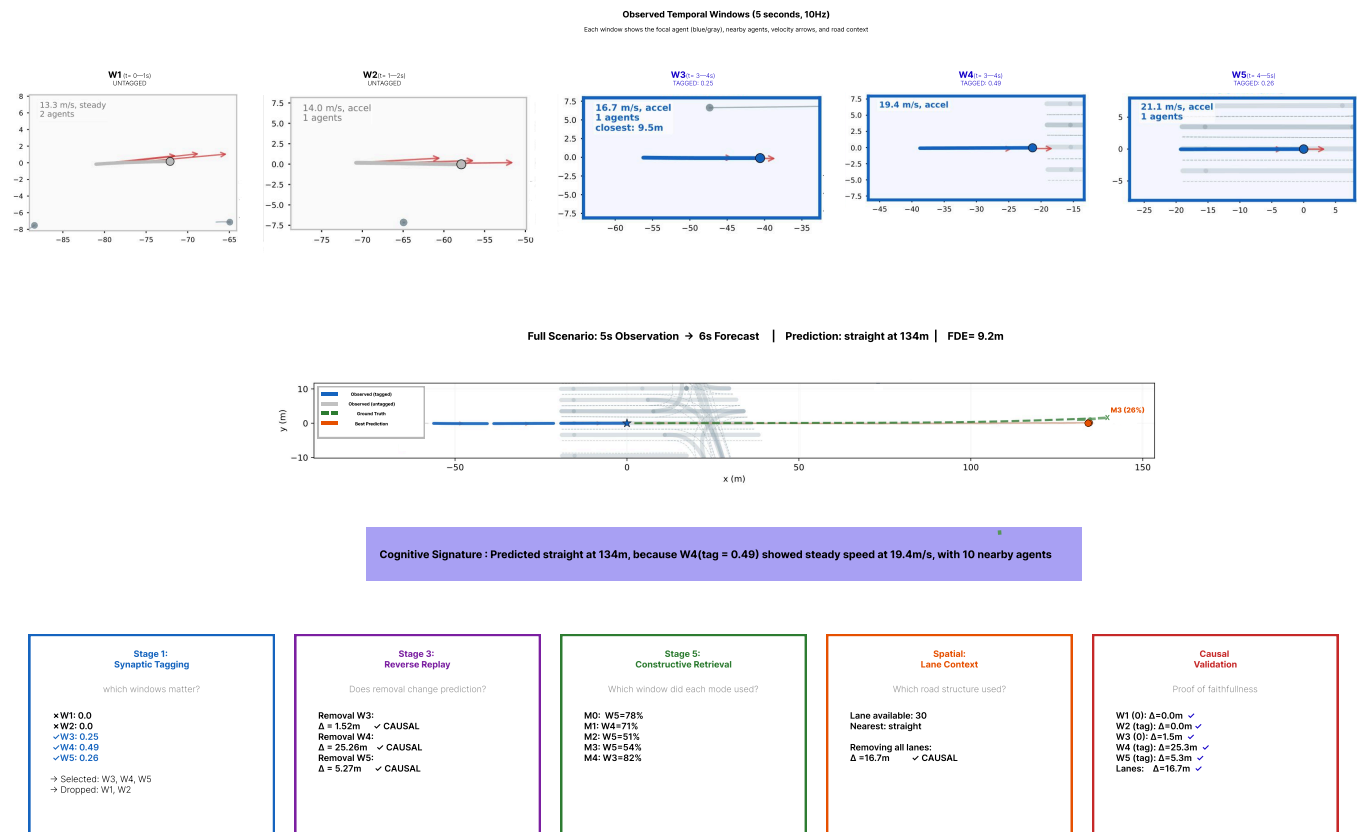


Figure 7. Qualitative result: **straight scenario** (FDE = 9.20 m). The model correctly identifies W4 (tag = 0.49) as the primary evidence window, showing steady high speed at 19.4 m/s. Causal validation confirms: removing W4 shifts the prediction by 25.3 m, while removing untagged W1 causes 0 m shift.

In the stop scenario (Figure 6), the model achieves near-perfect prediction (FDE = 0.19 m) and produces a faithful explanation: W4 (tag = 0.49) captured braking at 2.9 m/s with 23 nearby agents, and removing W4 shifts the prediction by 3.6 m.

The straight scenario (Figure 7) demonstrates a case where the explanation is faithful but the prediction is poor (FDE = 9.20 m). The model correctly identifies W4 as the dominant evidence window (steady high speed at 19.4 m/s), confirmed by a 25.3 m causal shift upon removal. The high FDE indicates that the model fails to predict the *magnitude* of forward travel despite correctly identifying the relevant temporal evidence. This illustrates an important distinction: explanation faithfulness and prediction accuracy are independent properties; the model can correctly explain *why* it made a prediction even when that prediction is inaccurate.

The turn scenario (Figure 8) shows FDE = 12.39 m, representing a typical failure case from the “complex turn” category identified in the failure taxonomy. The model tags W4 (tag = 0.47) and W5 (tag = 0.41), correctly identifying the deceleration and lateral movement preceding the turn. The 15.2 m causal shift confirms genuine reliance on this evidence. However, the 1-second window resolution compresses the turn initiation dynamics into a single vector, losing the fine-grained curvature progression that would be needed for accurate endpoint prediction in such maneuvers.

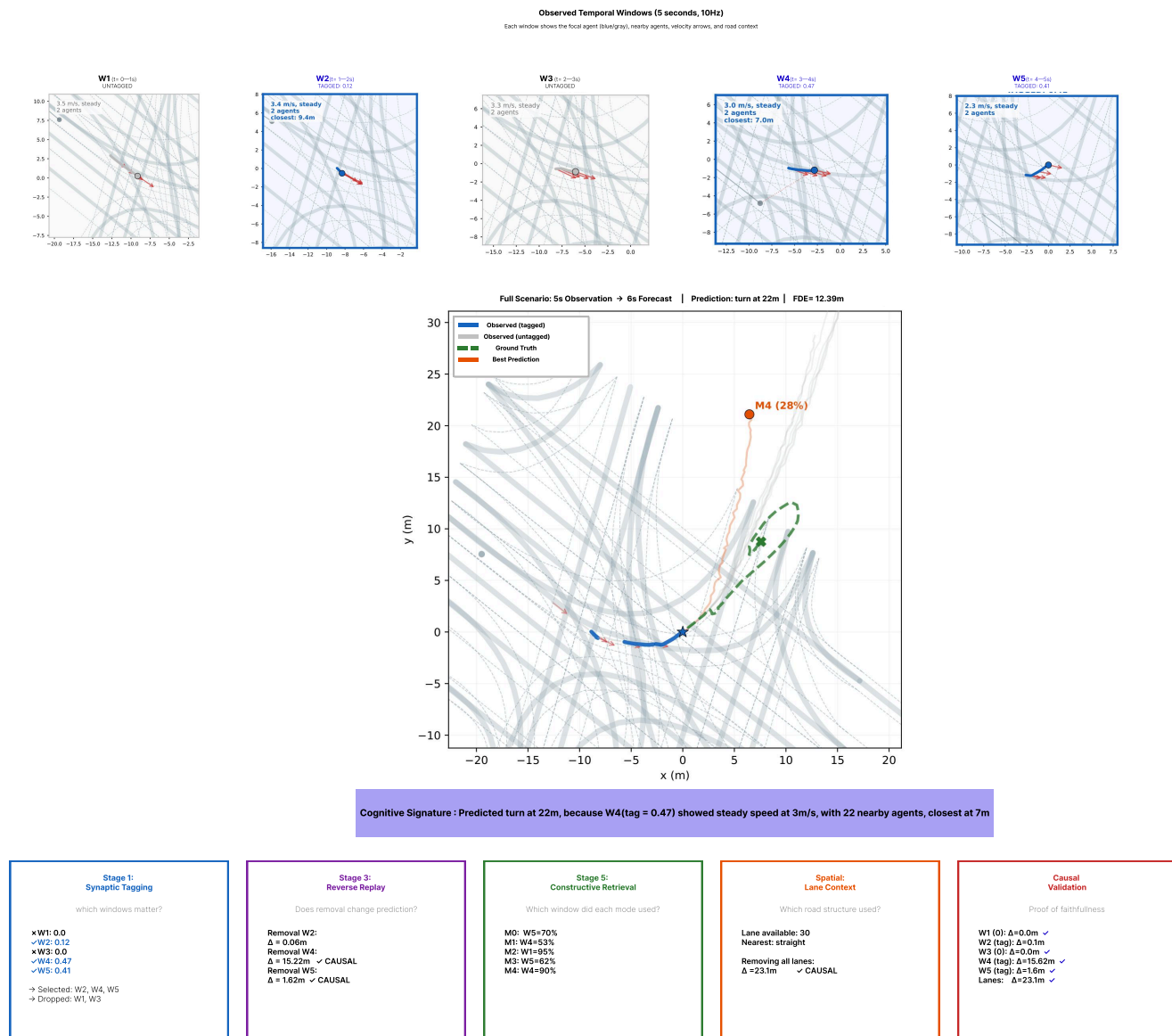


Figure 8. Qualitative result: **turn scenario** (FDE = 12.39 m). The model tags W4 (tag = 0.47) and W5 (tag = 0.41), identifying the deceleration and lateral movement preceding the turn. Despite higher FDE (a known limitation for complex maneuvers), the explanation correctly attributes the prediction to the relevant windows. Removing W4 shifts the endpoint by 15.2 m, confirming causal reliance.

4.12. Model Behavior Analysis

Figure 9 analyzes two global properties of the learned model across all 24,988 validation scenarios.

Panel (a) reveals a strong recency bias: W4 (3–4 s) receives the highest mean tag value (0.49), followed by W5 (4–5 s, mean = 0.31) and W3 (2–3 s, mean = 0.17). Windows W1 and W2 are almost always untagged (mean < 0.02). This is consistent with the intuition that recent observations are most predictive of immediate future behavior.

Panel (b) shows that the six learned mode queries specialize in distinct regions of the prediction space. Mode M4 (8799 wins) dominates the central straight-ahead region, while other modes cover longer-range predictions (M2: 5381 wins), lateral movements (M1, M5, and M6), and rare maneuvers (M3: 253 wins).

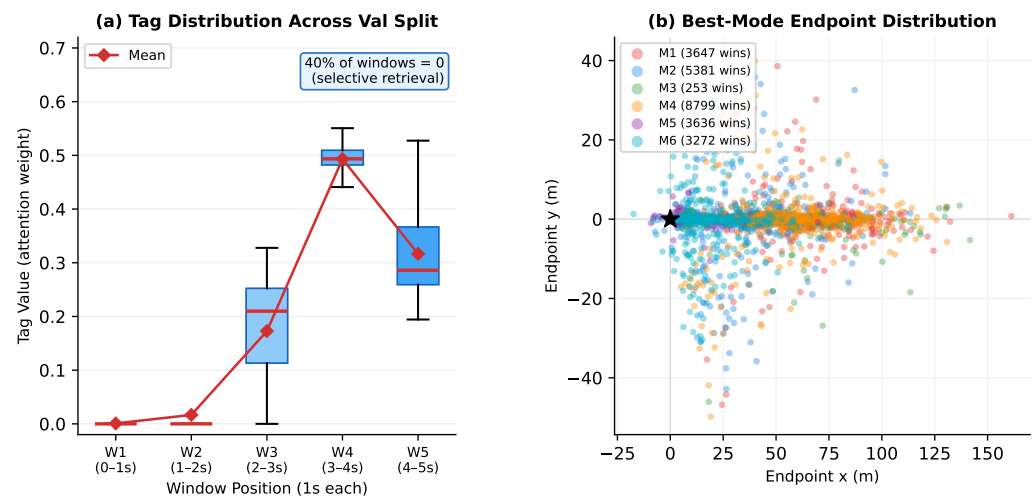


Figure 9. Model behavior analysis across 24,988 validation scenarios. (a) Tag value distribution per temporal window: later windows (W4 and W5) receive the highest attention (recency bias), and 40% of all windows receive exactly zero weight (selective memory retrieval). Stars (*) indicate the mode with the highest win count. (b) Best-mode endpoint distribution colored by mode ID: each mode specializes in a distinct region of the prediction space, confirming learned behavioral diversity.

5. Discussion

5.1. The Explainability–Accuracy Trade-Off

CogSig-Mamba’s minADE (0.908 m) is higher than pure prediction models like QCNet (0.718 m). This gap is a direct consequence of the information bottleneck: predictions must route through a five-window evidence bank with only three active windows rather than accessing raw 50-timestep features directly.

The miss rate ($MR_6 = 0.312$) represents the primary limitation. The per-maneuver analysis (Table 3) shows that this is driven by turning scenarios (59% miss rate), while safety-relevant maneuvers achieve substantially lower miss rates: stops (8%), straight driving (19%), and lane changes (27%). The failure taxonomy (Table 4) further shows that 96% of misses involve dense social interactions (>10 neighbors), indicating that the limitation is in multi-agent uncertainty coverage rather than in the explanation mechanism itself.

CogSig-Mamba’s minFDE (1.949 m) improves over SC-Mamba (2.089 m) by 6.7%, suggesting that routing predictions through a sparse evidence bottleneck acts as regularization, filtering out uninformative early observations. The model prioritizes recent causally relevant evidence—a principle that benefits endpoint prediction even at the cost of path-averaged accuracy.

5.2. Comparison with Post Hoc Explanation Methods

SHAP [11] and LIME [12] can be applied to any model but have significant limitations for real-time deployment: (1) they require hundreds to thousands of forward passes per explanation (vs. a single pass for tags + five passes for validation); (2) they provide no faithfulness guarantee: the explanation may not reflect actual model reasoning [13]; (3) they operate on individual features rather than semantically meaningful temporal windows. CogSig-Mamba’s explanations are produced as a natural byproduct of the forward pass (zero additional cost for tags) and verified with only five additional passes (10 ms total).

5.3. Safety Certification Implications

ISO 21448 (SOTIF) [7] requires that autonomous systems demonstrate behavioral traceability—the ability to explain why a specific action was taken given specific inputs. CogSig-Mamba’s Cognitive Signature provides a pathway toward this requirement: for

each prediction, it states which temporal observations drove the decision, which road structure was used, and verifies via causal intervention that these attributions are faithful. This audit trail can be logged, reviewed, and used for offline failure analysis—capabilities that are absent from black-box approaches.

For modern vehicles with SAE automation of levels 2–4, where over-the-air (OTA) software updates continuously modify ADAS behavior, such audit trails are increasingly relevant to regulatory compliance. Recent work on SOTIF validation through perception performance insufficiency injection reinforces this traceability requirement for automated driving systems [48]. The lightweight nature of CogSig-Mamba (1.9M parameters; 35 ms with full explanation on an A100 GPU) suggests potential suitability for onboard hardware, although deployment on production vehicle platforms would require additional profiling on embedded processors (e.g., NVIDIA Orin or similar) and validation under real-time constraints, which remains future work. Causally validated temporal attribution may also benefit downstream collision avoidance and path planning controllers [1]: if the forecasting module can articulate exactly which temporal window triggered a predicted maneuver, the planning module could cross-validate that evidence against its own safety constraints before executing a response.

It should be noted that the current prediction accuracy ($MR_6 = 0.312$) does not meet the requirements for safety-critical online deployment. The contribution of this work is not a deployment-ready prediction system but rather the *explanation mechanism* and demonstration that causal temporal attribution is achievable with minimal overhead. The model is best suited for offline safety auditing, post-incident analysis, and compliance documentation in its current form.

5.4. Limitations and Future Work

The model's primary limitation is the elevated miss rate in complex turning scenarios (59% MR for turns vs. 8% for stops). The failure taxonomy (Table 4) reveals that this is driven by two factors: (1) the evidence bank compresses 10 timesteps into a single vector per window, losing fine-grained curvature progression during turn initiation; and (2) dense social interactions create trajectory uncertainty that exceeds the coverage of six prediction modes. Future work could address the first factor through sub-window temporal attention or overlapping windows and the second through adaptive mode count or hierarchical prediction.

The sparsity parameter analysis (Table 1) shows that $k = 3$ is optimal for the current architecture but cannot be adjusted at inference time without retraining. Future work could explore training models at multiple k values or implementing a learned scenario-dependent sparsity mechanism that adjusts k based on prediction confidence.

The recency bias (W4/W5 dominating, Table 8) suggests that the model may underweight early observations that could provide context for gradual maneuvers. Curriculum learning or explicit long-range dependency objectives could address this.

The comparison with published baselines (Table 2) uses reported values from their respective papers, following the standard practice in the literature [3,4]. Fully matched comparisons (identical preprocessing, training budget, and hardware) were not feasible for all the methods but would strengthen future evaluations. The reported seed variance (Table 1 footnote: ± 0.027 for minADE₆ across three seeds) indicates stable training, although additional seeds could narrow confidence intervals.

Finally, the latency measurements are reported on an NVIDIA A100 GPU. Deployment on production vehicle hardware would require profiling on automotive-grade processors, which remains future work. The model's $O(T)$ complexity and small parameter count

(1.9M) are favorable for embedded deployment, but this has not been empirically validated on target platforms.

The Argoverse 2 explainability framework introduced in this work also supports integration with adaptive optimization techniques for improved classification reliability [38], although such extensions remain future work.

6. Conclusions

This paper presented CogSig-Mamba, a motion forecasting model that provides causally validated temporal explanations alongside trajectory predictions. By mapping hippocampal memory mechanisms—synaptic tagging, reverse replay, and constructive retrieval—to neural network components, both reasonable prediction accuracy and verifiably faithful explanations are achieved within a single architecture. The model produces a “Cognitive Signature” for each prediction: an audit trail stating which past observations drove the decision, verified by causal intervention across 24,988 scenarios.

The key empirical findings are: (1) learned synaptic tags produce $277\times$ higher causal discrimination than random window selection ($p < 0.001$), confirming genuine temporal credit assignment; (2) causal faithfulness holds across all the maneuver categories, including failure cases; (3) the model achieves a 92% success rate in stopping scenarios and 81% in straight driving, with the elevated overall miss rate (31%) concentrated in complex turning maneuvers.

With 1.9M parameters and $O(T)$ temporal complexity, CogSig-Mamba demonstrates that causal temporal attribution can be achieved with minimal computational overhead (10 ms for full explanation). The current work establishes the feasibility of intrinsically explainable motion forecasting with causal verification. Future work should address the miss rate limitation through adaptive sparsity mechanisms and richer temporal encoding and validate deployment characteristics on automotive-grade hardware. The explanation framework introduced here may support the behavioral traceability requirements of standards such as ISO 21448, particularly for offline safety auditing and post-incident analysis.

Author Contributions: Conceptualization, E.B.; methodology, E.B.; software, E.B.; validation, E.B.; formal analysis, E.B.; investigation, E.B.; data curation, E.B.; writing—original draft preparation, E.B.; writing—review and editing, E.B. and Z.I.; visualization, E.B.; supervision, Z.I.; project administration, Z.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Argoverse 2 Motion Forecasting dataset used in this study is publicly available at <https://www.argoverse.org/av2.html> (accessed on 21 May 2026). The trained model weights and inference code will be made available upon publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ADAS	Advanced Driver Assistance System
ADE	Average Displacement Error
CV	Constant Velocity
FDE	Final Displacement Error
HD	High Definition
MHA	Multi-Head Attention
MLP	Multi-Layer Perceptron
MR	Miss Rate
SOTIF	Safety of the Intended Functionality

SSM	State Space Model
XAI	Explainable Artificial Intelligence

References

- Chen, C.; Lin, Y.; Zhan, L.; He, Y.; Zhang, Y.; Chi, X.; Chen, M. Dynamic Obstacle Avoidance Approach Based on Integration of A-Star and APF Algorithms for Vehicles in Complex Mountainous Environments. *Vehicles* **2025**, *7*, 65. <https://doi.org/10.3390/vehicles7030065>.
- Chen, K.; Wu, G. The Vehicle Intention Recognition with Vehicle-Following Scene Based on Probabilistic Neural Networks. *Vehicles* **2023**, *5*, 332–343. <https://doi.org/10.3390/vehicles5010019>.
- Zhou, Z.; Wang, J.; Li, Y.H.; Huang, Y.K. QCNet: Query-Centric Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 17863–17873. <https://doi.org/10.1109/CVPR52729.2023.01713>.
- Zhou, Z.; Ye, L.; Wang, J.; Wu, K.; Lu, K. HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8813–8823. <https://doi.org/10.1109/CVPR52688.2022.00862>.
- Cheng, J.; Mei, X.; Liu, M. Forecast-MAE: Self-supervised Pre-training for Motion Forecasting with Masked Autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 8679–8689. <https://doi.org/10.1109/ICCV51070.2023.00797>.
- Shi, S.; Jiang, L.; Dai, D.; Schiele, B. MTR: Motion Transformer with Global Intention Localization and Local Movement Refinement. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 6531–6543. <https://doi.org/10.52202/068431-0473>.
- ISO 21448:2022; Road Vehicles—Safety of the Intended Functionality (SOTIF). Technical Report; International Organization for Standardization (ISO): Geneva, Switzerland, 2022.
- UL 4600; Standard for Safety for the Evaluation of Autonomous Products. Technical Report; Underwriters Laboratories (UL): Northbrook, IL, USA, 2020.
- Koopman, P.; Wagner, M. Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intell. Transp. Syst. Mag.* **2017**, *9*, 90–96. <https://doi.org/10.1109/MITS.2016.2583491>.
- Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf (accessed on 23 July 2026).
- Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>.
- Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3543–3556. <https://doi.org/10.18653/v1/N19-1357>.
- Ellouze, A.; Karray, M.; Ksantini, M. A Hybrid Decision-Making Framework for Autonomous Vehicles in Urban Environments Based on Multi-Agent Reinforcement Learning with Explainable AI. *Vehicles* **2026**, *8*, 8. <https://doi.org/10.3390/vehicles8010008>.
- Zhao, H.; Gao, J.; Lan, T.; Sun, C.; Sapp, B.; Varadarajan, B.; Shen, Y.; Shen, Y.; Chai, Y.; Schmid, C.; et al. TNT: Target-driven Trajectory Prediction. In Proceedings of the Conference on Robot Learning (CoRL), Cambridge, MA, USA (Virtual), 16–18 November 2020; PMLR: New York, NY, USA, 2020; Volume 155, pp. 895–904. Available online: <https://proceedings.mlr.press/v155/zhao21b.html> (accessed on 23 July 2026).
- Liang, M.; Yang, B.; Hu, R.; Chen, Y.; Liao, R.; Feng, S.; Urtasun, R. LaneGCN: Learning Lane Graph Representations for Motion Forecasting. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 541–556. https://doi.org/10.1007/978-3-030-58536-5_32.
- Minsky, M. Steps toward artificial intelligence. *Proc. IRE* **1961**, *49*, 8–30. <https://doi.org/10.1109/JRPROC.1961.287775>.
- Sutton, R.S. Temporal Credit Assignment in Reinforcement Learning. Ph.D. Thesis, University of Massachusetts Amherst, Amherst, MA, USA, 1984.
- Frey, U.; Morris, R.G.M. Synaptic tagging and long-term potentiation. *Nature* **1997**, *385*, 533–536. <https://doi.org/10.1038/385533a0>.
- Foster, D.J.; Wilson, M.A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **2006**, *440*, 680–683. <https://doi.org/10.1038/nature04587>.

21. Buzsáki, G.; Moser, E.I. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat. Neurosci.* **2013**, *16*, 130–138. <https://doi.org/10.1038/nn.3304>.
22. Mattar, M.G.; Daw, N.D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **2018**, *21*, 1609–1617. <https://doi.org/10.1038/s41593-018-0232-z>.
23. Marchetti, F.; Becattini, F.; Seidenari, L.; Del Bimbo, A. MANTRA: Memory Augmented Networks for Multiple Trajectory Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7141–7150. <https://doi.org/10.1109/CVPR42600.2020.00717>.
24. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 961–971. <https://doi.org/10.1109/CVPR.2016.110>.
25. Salzmann, T.; Ivanovic, B.; Chakravarty, P.; Pavone, M. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 683–700. https://doi.org/10.1007/978-3-030-58523-5_40.
26. Bayramov, E.; Istenes, Z. SC-Mamba: Causal Disentanglement of Driver Intent and Style for Autonomous Driving Motion Forecasting. In Proceedings of the International Conference on Information, Intelligence, Systems and Applications (IISA), Budapest, Hungary, 2026; Accepted; to appear.
27. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the Conference on Language Modeling (COLM), Philadelphia, PA, USA, 7–9 October 2024. Available online: <https://arxiv.org/pdf/2312.00752> (accessed on 23 July 2026).
28. Wilson, B.; Qi, W.; Agarwal, T.; Lambert, J.; Singh, J.; Khandelwal, S.; Pan, B.; Kumar, R.; Hartnett, A.; Pontes, J.K.; et al. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, Online, 6–14 December 2021. Available online: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4734ba6f3de83d861c3176a6273cac6d-Paper-round2.pdf (accessed on 23 July 2026).
29. Arjona-Medina, J.A.; Gillhofer, M.; Widrich, M.; Unterthiner, T.; Brandstetter, J.; Hochreiter, S. RUDDER: Return Decomposition for Delayed Rewards. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32. Available online: https://proceedings.neurips.cc/paper_files/paper/2019/file/16105fb9cc614fc29e1bda00dab60d41-Paper.pdf (accessed on 23 July 2026).
30. Hung, C.C.; Lillicrap, T.; Abramson, J.; Wu, Y.; Mirza, M.; Carnevale, F.; Ahuja, A.; Wayne, G. Optimizing agent behavior over long time scales by transporting value. *Nat. Commun.* **2019**, *10*, 5223. <https://doi.org/10.1038/s41467-019-13073-w>.
31. Bengio, Y.; Lecun, Y.; Hinton, G. Deep learning for AI. *Commun. ACM* **2021**, *64*, 58–65. <https://doi.org/10.1145/3448250>.
32. Chen, M.; Chen, Z.; Luo, L.; Tang, Y.; Cheng, J.; Wei, H.; Wang, J. Dynamic Visual Servo Control Methods for Continuous Operation of a Fruit Harvesting Robot Working Throughout an Orchard. *Comput. Electron. Agric.* **2024**, *219*, 108774. <https://doi.org/10.1016/j.compag.2024.108774>.
33. Zablocki, E.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452. <https://doi.org/10.1007/s11263-022-01657-x>.
34. Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; Akata, Z. Textual Explanations for Self-Driving Vehicles. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 577–593. https://doi.org/10.1007/978-3-030-01216-8_35.
35. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide. *IEEE Access* **2024**, *12*, 101603–101625. <https://doi.org/10.1109/ACCESS.2024.3431437>.
36. Wiegrefe, S.; Pinter, Y. Attention is not not Explanation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019; pp. 11–20. <https://doi.org/10.18653/v1/D19-1002>.
37. Lipton, Z.C. The Mythos of Model Interpretability. *Queue* **2018**, *16*, 31–57. <https://doi.org/10.1145/3236386.3241340>.
38. Trillo, J.R.; Del Moral, M.J.; Tapia, J.M.; García-Cabello, J.; Cabrerizo, F.J. Explainable Classifier with Adaptive Optimisation for Medical Data. *Appl. Intell.* **2026**, *56*, 77. <https://doi.org/10.1007/s10489-025-07081-1>.
39. O’Keefe, J.; Dostrovsky, J. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **1971**, *34*, 171–175. [https://doi.org/10.1016/0006-8993\(71\)90358-1](https://doi.org/10.1016/0006-8993(71)90358-1).
40. Tung, W.L.; Quek, C. A Hippocampal-Inspired Self-Organising Learning Memory Model with Analogical Reasoning for Decision Support. In Proceedings of the 2006 IEEE International Conference on Fuzzy Systems; IEEE: Piscataway, NJ, USA, 2006; pp. 1154–1161. <https://doi.org/10.1109/FUZZY.2006.1681855>.
41. Lv, X.; Wang, D. A Bionic Memory Replay Model Based on Hippocampal-Prefrontal Collaboration: Design, Mechanism, and Robot Navigation Verification. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 2852–2863. <https://doi.org/10.1109/TASE.2026.3657422>.

42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 23 July 2026).
43. Redish, A.D. *Beyond the Cognitive Map: From Place Cells to Episodic Memory*; MIT Press: Cambridge, MA, USA, 1999.
44. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660. <https://doi.org/10.1109/CVPR.2017.16>.
45. Wang, M.; Ren, X.; Jin, R.; Li, M.; Zhang, X.; Yu, C.; Wang, M.; Yang, W. FutureNet-LOF: Joint Trajectory Prediction and Lane Occupancy Field Prediction with Future Context Encoding. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 8841–8848. <https://doi.org/10.1109/ICRA55743.2025.11128073>.
46. Prutsch, A.; Bischof, H.; Possegger, H. Efficient Motion Prediction: A Lightweight & Accurate Trajectory Prediction Model with Fast Training and Inference Speed. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 9411–9417. <https://doi.org/10.1109/IROS58592.2024.10802425>.
47. Deo, N.; Wolff, E.; Beijbom, O. Multimodal Trajectory Prediction Conditioned on Lane-Graph Traversals. In *Proceedings of the Conference on Robot Learning (CoRL), London, UK, 8–11 November 2021*; PMLR: New York, NY, USA, 2021; Volume 164, pp. 203–212. Available online: <https://proceedings.mlr.press/v164/deo22a.html> (accessed on 23 July 2026).
48. Expósito Jiménez, V.J.; Macher, G.; Watzenig, D.; Brenner, E. Safety of the Intended Functionality Validation for Automated Driving Systems by Using Perception Performance Insufficiencies Injection. *Vehicles* **2024**, *6*, 1164–1184. <https://doi.org/10.3390/vehicles6030055>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.