MDPI

Article

Spatial Attention Visualization for Interpretable Trajectory Prediction in Autonomous Driving: Discovering Safety Blind Spots Through Counterfactual Analysis

Xingnan Zhou [1] and Ciprian Alecsandru [1,*]

[1] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

* Correspondence: ciprian.alecsandru@concordia.ca (C.A.)

Version February 12, 2026 submitted to Sustainability

Abstract: Accurate trajectory prediction is critical for autonomous driving safety and energy-efficient planning in sustainable urban mobility systems. While Transformer-based models achieve state-of-the-art prediction performance, their internal attention mechanisms remain opaque, hindering safety validation and regulatory compliance. We present a spatial attention visualization framework that maps Transformer attention weights onto bird's-eye-view traffic scenes via a novel spatial token bookkeeping mechanism, Gaussian splatting for agent tokens, and polyline painting for lane tokens. Using MTR-Lite, a lightweight Motion Transformer variant (8.48M parameters) trained on the Waymo Open Motion Dataset, we demonstrate the framework through systematic analysis of 100–200 validation scenes. Four key findings emerge: (1) layer-wise entropy analysis reveals non-monotonic hierarchical specialization—encoder layers progressively focus on agents (entropy 5.64→5.36 bits) before the final layer reverses to broad map attention (5.92 bits, 63.6% map tokens); (2) failed predictions exhibit "tunnel vision" with lower entropy (5.72 vs. 5.94 bits) and elevated self-attention (0.049 vs. 0.035); (3) distance-decay masking shows that far-range attention encodes essential context, with even mild masking degrading accuracy by 4.7%; and (4) scene-type analysis confirms dynamic attention adaptation, with dense-traffic scenes allocating 42.3% agent attention versus 18.4% in sparse scenes. We further introduce a counterfactual analysis methodology using controlled scene edits to enable causal reasoning about attention allocation. These findings provide actionable diagnostics for model developers and regulators seeking to validate safe autonomous vehicle deployment, contributing to sustainable urban mobility.

Keywords: trajectory prediction; attention visualization; Transformer; autonomous driving; explainable AI; vulnerable road users; counterfactual analysis; sustainable transportation

1. Introduction

Autonomous vehicles (AVs) represent a transformative technology for achieving sustainable urban mobility. By reducing human-error-related collisions—which account for over 94% of serious crashes according to the U.S. National Highway Traffic Safety Administration [1]—AVs promise substantial improvements in traffic safety, energy efficiency, and urban livability. These benefits align directly with the United Nations Sustainable Development Goals, particularly SDG 11 (Sustainable Cities and Communities) and SDG 13 (Climate Action) [2]. Studies project that widespread AV adoption could reduce traffic fatalities by 90%, decrease fuel consumption by 40% through smoother driving patterns, and reclaim urban space currently dedicated to parking [3–5]. However, realizing these

benefits depends critically on achieving public trust and regulatory approval, both of which remain constrained by the opacity of the artificial intelligence systems that underpin autonomous driving [6,7].

At the core of modern AV planning pipelines lies motion prediction: forecasting the future trajectories of surrounding vehicles, pedestrians, and cyclists. Transformer-based architectures have emerged as the dominant paradigm for this task, achieving state-of-the-art performance on major benchmarks. Models such as Motion Transformer (MTR) [8,9], Wayformer [10], GameFormer [11], and Scene Transformer [12] leverage multi-head self-attention and cross-attention mechanisms to capture complex interactions among traffic agents and road geometry. Despite their strong quantitative performance, these models operate as *black boxes*: the attention weights that encode inter-agent relationships, lane preferences, and temporal reasoning remain hidden from developers and safety engineers. This lack of interpretability creates three practical barriers. First, when a model produces an erroneous prediction—such as failing to anticipate a left-turning vehicle—there is no principled way to diagnose whether the failure stems from insufficient attention to the relevant agent, the target lane, or the traffic signal. Second, regulatory bodies increasingly demand explanations for safety-critical AI decisions, as codified in the European Union AI Act [13] and NHTSA testing frameworks [1]. Third, without transparency, the general public lacks the evidence necessary to trust autonomous systems, ultimately delaying adoption and the associated sustainability benefits [14,15].

Several lines of research have addressed AI interpretability, though significant gaps remain in the trajectory prediction domain. Post-hoc explanation methods such as LIME [16], SHAP [17], and Grad-CAM [18] provide input-level attributions but do not leverage the structured internal attention mechanisms of Transformers. In natural language processing and computer vision, dedicated attention visualization tools—including BERTViz [19], Attention Flow [20], and Transformer Explainability [21]—have demonstrated that attention patterns encode interpretable relationships, though the debate on whether attention constitutes explanation continues [22,23]. Within trajectory prediction, recent work has begun to explore attention-based interpretability: VISTA [24] visualizes pairwise interaction strength, and LMFormer [25] examines lane-conditioned attention maps. However, these efforts focus on isolated aspects of the attention spectrum—either agent–agent interactions or lane selection—and do not provide a unified view of *where* the model attends in physical space, *how* its reasoning evolves across processing layers, and *which* lane structures guide its predictions.

In this paper, we present a spatial attention visualization framework for Transformer-based trajectory prediction that goes beyond depicting abstract attention matrices. We specifically adopt a Transformer architecture because its multi-head attention mechanism provides a built-in interpretability window: the attention weights directly reveal the model's spatial focus, agent and lane priorities, and layer-wise processing evolution. Unlike recurrent architectures such as LSTMs—where hidden states encode temporal dependencies in opaque, entangled vectors—or convolutional networks whose internal feature maps lack token-level semantic correspondence, Transformer attention offers explicit, per-token interaction scores that are naturally amenable to spatial visualization. Indeed, recent work by Zhou and Alecsandru [26] demonstrated that lane-graph conditioning improves LSTM-based trajectory prediction on the Waymo Open Motion Dataset; however, the LSTM hidden states do not afford the spatial interpretability that Transformer attention provides, motivating our choice of architecture for interpretability-focused analysis.

Crucially, our primary contribution is a *model-agnostic visualization framework* applicable to any Transformer-based trajectory predictor, not a new state-of-the-art prediction model. To demonstrate and validate this framework, we employ a lightweight variant of the MTR architecture (MTR-Lite, 8.48M parameters) as an *interpretability probe*—a deliberately simplified model that enables rapid experimentation and systematic attention analysis without the computational overhead of production-scale systems. MTR-Lite is trained on 20% of the Waymo Open Motion Dataset [27] (approximately 17,800 scenes), yielding sufficient diversity to reveal meaningful attention patterns while maintaining experimental tractability. The framework itself—encompassing attention extraction, spatial token bookkeeping, and visualization rendering—is architecture-agnostic and can be applied

to any Transformer model that produces attention weights over spatially grounded tokens, including production-scale systems such as MTR++ [9], Wayformer [10], and QCNet [28]. Our key technical innovation is a *spatial token bookkeeping* mechanism that maintains a bidirectional mapping between discrete token indices and their physical BEV coordinates, enabling attention weights to be projected as continuous heatmaps directly onto the traffic scene. Using Gaussian splatting for agent tokens and polyline painting for lane tokens, the resulting visualizations provide a spatially grounded view of the model's attention allocation and its progressive evolution across processing layers.

Importantly, we go beyond visualization as an end in itself. By systematically analyzing the spatial distribution of attention across 100–200 Waymo validation scenes, we uncover **quantifiable safety-relevant patterns**. We find that failed predictions exhibit significantly lower attention entropy (5.72 vs. 5.94 bits) and elevated self-attention compared to successful predictions—a "tunnel vision" failure mode in which the model over-focuses on the ego agent when it should be distributing attention more broadly. This diagnostic pattern has direct implications for collision risk, as it reveals that prediction failures are accompanied by a measurable, detectable attention pathology. We further design a counterfactual analysis methodology using controllable scene editing (agent removal, injection, and traffic signal manipulation), enabling causal—rather than merely correlational—reasoning about how individual traffic elements influence model attention and prediction outcomes. This combination of spatial visualization, failure diagnostics, and causal experimentation transforms attention analysis from a qualitative illustration into a rigorous diagnostic tool.

The main contributions of this work are as follows:

- We propose a **spatial attention visualization system** that maps abstract Transformer attention weights onto bird's-eye-view traffic scenes via Gaussian splatting and polyline painting, providing the first spatially grounded interpretation of attention in trajectory prediction.
- We identify a **tunnel vision failure mode** in which failed predictions exhibit significantly lower attention entropy and elevated self-attention compared to successful predictions, providing a safety-critical diagnostic signal that links measurable attention pathology to prediction failure.
- We design a **counterfactual attention analysis methodology** using controllable scene generation, providing the infrastructure for causal—rather than merely correlational—reasoning about how individual traffic elements influence model attention and prediction outcomes.
- We provide **quantitative attention diagnostics**, including layer-wise entropy analysis revealing hierarchical specialization—encoder layers progressively focus on nearby agents (entropy 5.64→5.36 bits) while the final layer reverses to broad map attention (entropy 5.92 bits, 63.6% map tokens)—and a distance mask ablation demonstrating that suppressing far-range attention degrades accuracy by 4.7%, demonstrating that distant context is valuable and naive pruning is harmful.
- We demonstrate the framework across **diverse driving scenarios**—dense intersections (42.3% agent attention), sparse highways (18.4% agent attention, mean attended distance 21.4 m vs. 17.0 m at intersections), and failure cases—revealing how the model dynamically adapts its spatial attention distribution to scene complexity.

Beyond its technical contributions, this work has direct implications for sustainable transportation. The discovery that prediction failures are accompanied by a measurable tunnel vision pathology—lower entropy and elevated self-attention—has immediate practical consequences: this attention signature could serve as a runtime monitor to flag unreliable predictions before they propagate to the planning module, preventing potential collisions. By quantifying failure-associated attention patterns and analyzing scene-type adaptation, we provide actionable guidance for model developers and regulatory bodies alike. For regulators, spatially grounded attention visualizations offer the kind of human-readable evidence needed to certify AV behavior in complex traffic scenarios, particularly as the European Union AI Act [13] establishes explainability requirements for high-risk AI systems. For the public, the ability to see that an autonomous vehicle "looks at" the correct lanes, traffic signals, and nearby agents before making predictions builds the transparency necessary for trust [6,29].

Furthermore, our distance mask ablation reveals that naive attention pruning strategies—which might be pursued for computational efficiency—actually degrade prediction accuracy by 4.7%, cautioning against premature optimization and underscoring the need for interpretability-guided model compression rather than blind sparsification. Ultimately, by combining interpretability with safety diagnostics, our framework helps remove key barriers to safe AV deployment, contributing to the broader goal of reducing road fatalities, lowering transportation emissions, and creating more walkable, livable cities [7,30].

The remainder of this paper is organized as follows. Section 2 reviews related work on trajectory prediction, attention visualization, and explainable AI for autonomous driving. Section 3 describes the MTR-Lite architecture, attention extraction mechanism, and visualization pipeline. Section 4 presents quantitative evaluation results and visualization examples. Section 5 discusses the interpretability insights, sustainability implications, and limitations. Section 6 concludes with future directions.

2. Related Work

2.1. Transformer-Based Trajectory Prediction

The application of Transformer architectures [31] to motion forecasting has yielded substantial performance gains on standardized benchmarks. Early work by Gao et al. [32] introduced vectorized scene representations and point-level attention over polyline-encoded map elements, establishing a paradigm adopted by subsequent architectures. Scene Transformer [12] extended this approach to joint multi-agent prediction, employing factored self-attention over agent and time axes to model cooperative and adversarial interactions simultaneously. These foundational architectures demonstrated that attention mechanisms could implicitly capture the spatial and social structure of traffic scenes without explicit graph construction.

The Motion Transformer (MTR) family [8,9] introduced a query-based decoder design that has become influential in the field. MTR employs 64 learnable intention queries, initialized from clustered trajectory endpoints, which attend to encoded scene tokens through iterative cross-attention layers. This design separates *global intention localization* (selecting a coarse goal region) from *local movement refinement* (producing smooth trajectories conditioned on that goal), yielding strong multi-modal predictions. MTR++ extended this with symmetric scene modeling and pair-wise interaction modules, achieving first place in the 2023 Waymo Open Dataset Motion Prediction Challenge. Notably, the intention query mechanism generates structured attention patterns—each query attends to the agents and lanes relevant to its predicted mode—yet neither MTR nor MTR++ provides tools to visualize or analyze these patterns.

Wayformer [10] explored attention-based modality fusion, comparing early, late, and hierarchical fusion strategies for combining agent trajectories, road geometry, and traffic signal features. Their ablation showed that attention over traffic light tokens significantly improves prediction at signalized intersections, hinting at the interpretive value of attention analysis. GameFormer [11] introduced hierarchical game-theoretic decoding with level-k attention, modeling interactive prediction as iterated best-response reasoning. HPTR [33] proposed heterogeneous polyline attention with relative pose encoding and k-nearest-neighbor sparsification, improving efficiency while maintaining the ability to model agent–lane interactions. QCNet [28] developed query-centric encoding that avoids recomputing scene features for each target agent. Most recently, SMART [34] recast trajectory prediction as next-token prediction over discretized motion tokens, achieving state-of-the-art results on the Waymo Sim Agents benchmark with an autoregressive Transformer.

Table 1 summarizes the attention mechanisms used by these models and whether any form of attention visualization or interpretability analysis was reported. As the table shows, while all models employ multiple attention mechanisms (self-attention, cross-attention, or both), none provides systematic visualization of the full attention spectrum. This gap motivates our work.

Table 1. Summary of attention mechanisms in state-of-the-art trajectory prediction models. "Viz" indicates whether the paper includes attention visualization or interpretability analysis.

Model	Venue	Self-Attn	Cross-Attn	Query-Based	Viz
VectorNet [32]	CVPR 2020	✓	–	–	–
Scene Trans. [12]	ICLR 2022	✓	–	–	–
MTR [8]	NeurIPS 2022	✓	✓	✓	–
QCNet [28]	CVPR 2023	✓	✓	✓	–
Wayformer [10]	ICRA 2023	✓	✓	–	Partial
GameFormer [11]	ICCV 2023	✓	✓	✓	–
HPTR [33]	NeurIPS 2023	✓	✓	–	–
MTR++ [9]	TPAMI 2024	✓	✓	✓	–
SMART [34]	NeurIPS 2024	✓	–	–	–
Ours	–	✓	✓	✓	**Full**

2.2. Attention Visualization and Interpretability

The question of whether attention weights constitute meaningful explanations has been extensively debated in the NLP community. Jain and Wallace [22] argued that attention distributions are not reliable indicators of feature importance, showing that alternative attention configurations can yield equivalent predictions. Wiegreffe and Pinter [23] countered that attention weights do carry explanatory signal, particularly when the attention mechanism is constrained or task-specific. This nuanced view has informed subsequent work: attention is most interpretable when it operates over semantically meaningful units (words, objects, entities) rather than arbitrary hidden dimensions.

Several tools have been developed for visualizing attention in NLP Transformers. BERTViz [19] provides interactive multi-scale visualizations of attention heads across layers, revealing syntactic and semantic patterns in pre-trained language models. Abnar and Zuidema [20] introduced Attention Flow, which propagates attention through the residual stream to attribute model decisions to input tokens. For Vision Transformers, Chefer et al. [21] combined attention rollout with gradient information to produce class-specific relevance maps that outperform raw attention in localization tasks.

In the trajectory prediction domain, attention-based interpretability has received limited but growing interest. VISTA [24] introduced a goal-conditioned multi-agent forecasting Transformer whose social-attention block outputs pairwise attention matrices between agents, demonstrating that the model assigns increasing attention to agents on potential collision courses. LMFormer [25] proposed a lane-aware motion prediction Transformer with Mode2Lane cross-attention in the decoder, showing that attention peaks on lanes aligned with the predicted trajectory. ISE-GT [35] incorporated interaction strength encoding derived from a driver resistance field model into a graph Transformer, with a companion Interaction Tendency Reasoning Module that provides post-hoc interpretability by verifying that inferred interaction tendencies align with human driver intuition.

While these contributions represent important progress, they share a common limitation: each addresses a single facet of the attention spectrum. VISTA focuses exclusively on agent–agent social attention; LMFormer examines only lane-conditioned decoder attention; ISE-GT provides post-hoc interaction tendency analysis but not spatial or temporal attention patterns. None offers a unified framework that simultaneously visualizes (1) the spatial distribution of attention across agents and lanes, (2) the temporal evolution of attention across decoder layers, and (3) the structural selection of lane tokens that condition trajectory generation. Our work fills this gap by providing all three visualization types within a single, integrated pipeline.

2.3. Counterfactual Analysis and Controllable Scene Generation

Counterfactual reasoning—asking "what would have happened if X were different?"—provides a principled framework for causal inference in machine learning [36]. Goyal et al. [37] demonstrated counterfactual visual explanations by identifying minimal image modifications that change a

classifier's prediction, revealing which visual features are causally relevant. In contrast to purely observational analysis, counterfactual experiments can distinguish genuine causal mechanisms from spurious correlations.

In autonomous driving, controllable scene generation has emerged as a tool for safety validation and model stress testing. SceneGen [38] learned to place realistic traffic participants in BEV layouts, while TrafficSim [39] modeled multi-agent interactions through learned conditional distributions. More recently, guided diffusion models [40] have enabled fine-grained control over generated traffic scenarios, including adversarial agent placement and rare event synthesis. Ding et al. [41] comprehensively reviewed methods for safety-critical scenario generation, identifying controllability and realism as the two key desiderata.

Despite these advances, no prior work has combined controllable scene generation with systematic attention analysis. Existing scene generation methods focus on evaluating prediction *accuracy* (i.e., whether the model predicts correctly) rather than prediction *attention* (i.e., where the model looks). Our work bridges this gap: by editing real Waymo scenes—removing agents, injecting vulnerable road users, flipping traffic signals—and measuring the resulting changes in attention distributions, we perform the first *counterfactual attention analysis* for trajectory prediction. This enables causal claims about how individual scene elements influence model reasoning, moving beyond correlational findings.

2.4. Explainable AI for Autonomous Driving

The demand for explainable AI (XAI) in autonomous driving extends beyond academic curiosity to practical necessity. Arrieta et al. [42] provide a comprehensive taxonomy of XAI methods, distinguishing between transparent models (inherently interpretable), post-hoc explanations (applied after training), and hybrid approaches. For safety-critical applications like autonomous driving, they argue that post-hoc methods are insufficient; the model's internal reasoning process must be accessible and auditable.

Zablocki et al. [15] surveyed explainability specifically in deep vision-based driving systems, identifying four key dimensions: *what* is explained (perception, prediction, or planning), *how* explanations are generated (saliency maps, natural language, attention), *who* the audience is (developers, regulators, or passengers), and *when* explanations are provided (offline analysis or real-time). Our work addresses the *prediction* component using *attention-based spatial visualization*, targeting both *developers* (for debugging) and *regulators* (for safety certification), in an *offline analysis* setting.

Atakishiyev et al. [29] recently provided an extensive field guide for XAI research in autonomous driving, emphasizing that the gap between model performance and model understanding is the primary obstacle to large-scale deployment. They identify trajectory prediction as a particularly underserved area for interpretability research, noting that most XAI efforts in AV focus on perception (object detection saliency) or planning (reward visualization) rather than the prediction module that bridges them.

From a regulatory perspective, the European Union AI Act [13] classifies autonomous driving systems as "high-risk AI" requiring transparency, human oversight, and documented testing. The NHTSA framework [1] similarly calls for testable scenarios and explainable decision processes. These regulatory requirements create a concrete demand for the kind of interpretability tools that our framework provides: spatially grounded visualizations that can demonstrate, for a given scenario, exactly which traffic participants and road structures the model considered before generating its prediction.

The connection between AV interpretability and sustainability is increasingly recognized. Taiebat et al. [30] reviewed the energy and environmental implications of connected and automated vehicles, concluding that the magnitude of benefits depends heavily on the pace of adoption, which is in turn constrained by safety assurance and public trust. Litman [43] projects that full AV benefits—including

a 60–90% reduction in crash costs and a 30–50% decrease in vehicle-miles traveled per household—will materialize only when Level 4+ autonomy achieves widespread deployment, a milestone that requires overcoming the trust deficit. By making trajectory prediction models interpretable, our work contributes to this trust-building process and, by extension, to the realization of the environmental and safety benefits that motivate sustainable transportation research.

3. Materials and Methods

This section presents the dataset, model architecture, attention extraction mechanism, spatial token bookkeeping system, visualization methods, counterfactual experiment design, and evaluation metrics that constitute our framework.

3.1. Dataset

We train and evaluate our model on the Waymo Open Motion Dataset (WOMD) v1.2 [27], one of the largest and most diverse public benchmarks for trajectory prediction. The full dataset contains approximately 89,000 driving scenes recorded across six U.S. cities. Each scene spans 91 frames captured at 10 Hz (9.1 seconds of real-world driving), providing dense temporal coverage of traffic interactions. We use a 20% subset of the full dataset, yielding approximately 17,800 scenes, split into 85% training (∼15,130 scenes) and 15% validation (∼2,670 scenes) using hash-based scene-ID partitioning for reproducibility.

Each scene accommodates up to 100 agent slots, covering three agent types: vehicles, pedestrians, and cyclists. Every agent is represented as a trajectory with per-frame attributes including position, velocity, acceleration, heading, and bounding box dimensions. Importantly, the dataset provides rich map context: a lane graph encoding road topology with successor, predecessor, and left/right neighbor relationships among lane segments; per-lane attributes including speed limits, lane types, and boundary markings; and traffic signal states recorded per frame per controlled lane. This structured map representation is critical for our visualization framework, as it enables projecting abstract map-token attention weights back onto physically meaningful road geometry.

We preprocess the raw data into per-scene `pkl` files, each storing a dictionary with three primary entries: `objects[]`, containing per-agent trajectory arrays and metadata; `lane_graph{}`, encoding lane centerline polylines together with their topological connectivity and attributes; and `traffic_lights[]`, recording per-frame signal states for each controlled lane. This dictionary structure facilitates both efficient batched training and the counterfactual scene editing experiments described in Section 3.6.

3.2. MTR-Lite Architecture

Our trajectory prediction model, MTR-Lite, is a lightweight variant of the Motion Transformer (MTR) [8,9] designed for interpretability research on a single-GPU workstation. The model comprises 8.48M parameters and follows an encode–attend–decode pipeline with four stages: polyline encoding, scene encoding, motion decoding, and mode selection.

3.2.1. Input Representation

The model ingests two types of polyline inputs. *Agent polylines* represent traffic participants: we select A=32 agents nearest to the target agent, each described by a polyline of T_h=11 historical timesteps (1.0 second of history at 10 Hz). Each timestep carries a 29-dimensional feature vector:

$$\mathbf{f}_{\text{agent}} = \big[\underbrace{x, y}_{2}, \underbrace{x_{-1}, y_{-1}}_{2}, \underbrace{v_x, v_y}_{2}, \underbrace{a_x, a_y}_{2}, \underbrace{\sin\theta, \cos\theta}_{2}, \underbrace{w, l}_{2}, \underbrace{\mathbf{c}_{\text{type}}}_{5}, \underbrace{\mathbf{e}_{\text{time}}}_{11}, \underbrace{z_{\text{ego}}}_{1}\big] \in \mathbb{R}^{29}, \tag{1}$$

where (x, y) is the current position, (x_{-1}, y_{-1}) the previous-step position, (v_x, v_y) and (a_x, a_y) the velocity and acceleration, $(\sin\theta, \cos\theta)$ the heading encoded as sine–cosine pair, (w, l) the bounding box width and length, $\mathbf{c}_{\text{type}} \in \{0, 1\}^5$ a one-hot agent type encoding (vehicle, pedestrian, cyclist, and

two reserved classes), $\mathbf{e}_{\text{time}} \in \mathbb{R}^{11}$ a learnable temporal positional embedding, and $z_{\text{ego}} \in \{0,1\}$ a binary indicator of whether the agent is the ego vehicle.

Map polylines represent lane centerlines: we select M=64 lane segments nearest to the target agent, each described by P=20 points sampled uniformly along the centerline. Each point carries a 9-dimensional feature vector:

$$\mathbf{f}_{\text{map}} = \big[\underbrace{x, y}_{2}, \underbrace{d_x, d_y}_{2}, \underbrace{\mathbf{g}_{\text{lane}}}_{3}, \underbrace{x_{-1}, y_{-1}}_{2}\big] \in \mathbb{R}^9, \tag{2}$$

where (x, y) is the point position, (d_x, d_y) the local direction vector, $\mathbf{g}_{\text{lane}} \in \{0,1\}^3$ encodes lane flags (has traffic control, is intersection lane, is turn lane), and (x_{-1}, y_{-1}) the coordinates of the preceding point in the polyline.

3.2.2. PointNet Encoder

Each polyline—whether agent or map—is independently encoded into a fixed-dimensional token using a PointNet-style architecture [44]. A shared-weight multi-layer perceptron (MLP) processes each point along the polyline:

$$\text{MLP}_{\text{point}} : \mathbb{R}^D \xrightarrow{\text{Linear}} \mathbb{R}^{64} \xrightarrow{\text{ReLU}} \mathbb{R}^{128} \xrightarrow{\text{ReLU}} \mathbb{R}^{256} \xrightarrow{\text{ReLU}} \mathbb{R}^{256}, \tag{3}$$

where D is the input feature dimension (29 for agents, 9 for map). A symmetric max-pooling operation aggregates the per-point features across the polyline's temporal or spatial extent, producing a single 256-dimensional vector that is invariant to point ordering. A post-aggregation MLP refines this representation:

$$\text{MLP}_{\text{post}} : \mathbb{R}^{256} \xrightarrow{\text{Linear}} \mathbb{R}^{256} \xrightarrow{\text{ReLU}} \mathbb{R}^{256}, \tag{4}$$

followed by layer normalization [45]. The agent and map encoders share this architectural template but maintain separate learned parameters. This stage produces 32 agent tokens and 64 map tokens, each in $\mathbb{R}^{256}$.

3.2.3. Scene Encoder

The 96 tokens (32 agent + 64 map) are concatenated into a single sequence and processed by a global self-attention encoder comprising L_e=4 Transformer encoder layers [31]. Each layer applies pre-norm multi-head self-attention with H=8 heads (d_k=d_v=32) and a position-wise feed-forward network (FFN) with hidden dimension 1024:

$$\mathbf{z}' = \mathbf{z} + \text{MultiHead}\big(\text{LN}(\mathbf{z}), \text{LN}(\mathbf{z}), \text{LN}(\mathbf{z})\big), \tag{5}$$

$$\mathbf{z}'' = \mathbf{z}' + \text{FFN}\big(\text{LN}(\mathbf{z}')\big), \tag{6}$$

where $\text{LN}(\cdot)$ denotes layer normalization and the residual connections follow the pre-norm convention. Global self-attention allows every token to attend to every other token, enabling agent–agent, agent–map, map–agent, and map–map interactions to emerge naturally. After the final encoder layer, the 96 tokens are split back into 32 encoded agent tokens and 64 encoded map tokens.

3.2.4. Motion Decoder

For each target agent, the decoder generates K_0=64 candidate trajectory modes using an intention-query mechanism inspired by MTR [8]. Each of the 64 intention queries is initialized by summing (i) a learned embedding of a 2D anchor point (obtained via k-means clustering of training-set trajectory endpoints) with (ii) a context embedding derived from the target agent's encoded token. The decoder consists of L_d=4 layers, each performing:

1. **Agent cross-attention**: intention queries attend to the 32 encoded agent tokens, capturing dynamic interactions.
2. **Map cross-attention**: intention queries attend to the 64 encoded map tokens, selecting lane-level guidance.
3. **Feed-forward network**: position-wise nonlinear transformation with hidden dimension 1024.

Each decoder layer is followed by a per-layer trajectory head (for deep supervision) that regresses a trajectory of T_f=80 future timesteps (8.0 seconds at 10 Hz) and a scalar confidence logit from the refined query embedding. The deep supervision loss weights are $[0.2, 0.2, 0.2, 0.4]$ from the first to the last layer.

3.2.5. Mode Selection

From the 64 candidate modes produced by the final decoder layer, we apply distance-based non-maximum suppression (NMS) with a threshold of 2.0 m on trajectory endpoints. This yields K=6 diverse output modes, each comprising a predicted trajectory $\hat{\mathbf{Y}}_k \in \mathbb{R}^{80\times 2}$ and a confidence score $\hat{p}_k$. The confidence scores are normalized via softmax to form a probability distribution over modes.

3.2.6. Training

The model is trained for 60 epochs with the AdamW optimizer [46] (learning rate 10^{-4}, weight decay 0.01), using a linear warmup over 5 epochs followed by cosine annealing decay. Automatic mixed-precision (AMP) training with float16 [47] is employed throughout. The loss function combines a cross-entropy classification loss over mode scores with a smooth-ℓ_1 regression loss over trajectory coordinates, applied at every decoder layer with deep supervision. Gradient clipping is set to a maximum norm of 1.0, and training uses batch size 4 with 8-step gradient accumulation (effective batch size 32).

3.3. Attention Extraction Framework

A central requirement of our visualization pipeline is the ability to extract per-head attention weight matrices from every layer without altering the model's predictions. We accomplish this through custom Transformer layers that extend PyTorch's `nn.MultiheadAttention` with a lightweight capture mechanism.

3.3.1. Attention-Capture Layers

We implement two custom layer classes: `AttentionCaptureEncoderLayer` for the scene encoder and `AttentionCaptureDecoderLayer` for the motion decoder. Both accept a boolean flag `capture_attention` on their forward pass. When this flag is set to `True`, the underlying multi-head attention call is invoked with `need_weights=True` and `average_attn_weights=False`, causing PyTorch to return the full per-head attention weight tensor rather than discarding it or averaging across heads. When the flag is `False` (the default during training), no attention weights are computed or stored, incurring zero overhead.

3.3.2. AttentionMaps Data Structure

All captured weights from a single forward pass are organized in an `AttentionMaps` dataclass with three primary fields:

- `scene_attentions`: a list of L_e=4 tensors, each of shape (B, H, N, N) where N=A+M=96, representing per-head self-attention weights at each encoder layer. Each tensor is a row-stochastic matrix (rows sum to 1) in the last dimension.
- `decoder_agent_attentions`: a list of L_d=4 tensors per target agent, each of shape (B, H, K_0, A) where K_0=64 and A=32, representing per-head cross-attention from intention queries to agent tokens.

- `decoder_map_attentions`: a list of L_d=4 tensors per target agent, each of shape (B, H, K_0, M) where M=64, representing per-head cross-attention from intention queries to map tokens.

This structure provides accessor methods for extracting specific submatrices: agent-to-agent attention, agent-to-map attention, map-to-agent attention, and per-mode decoder attention. An `aggregate_heads` method supports both mean and max aggregation across heads, and a `compute_entropy` method computes Shannon entropy in bits for quantitative analysis.

3.4. Spatial Token Bookkeeping

The key technical innovation enabling our visualization approach is a *spatial token bookkeeping* system that maintains a bidirectional mapping between the abstract token index space used by the Transformer and the continuous bird's-eye-view (BEV) coordinate space of the physical scene. Without this mapping, attention weights are merely entries in a matrix indexed by opaque integers; with it, each attention value acquires a spatial interpretation.

For each *agent token* $i \in \{0, \ldots, A-1\}$, the bookkeeper stores the agent's BEV position (x_i, y_i) at the anchor frame, heading angle θ_i, bounding box dimensions (w_i, l_i), and agent type. For each *map token* $j \in \{0, \ldots, M-1\}$, the bookkeeper stores the full lane centerline polyline $\{(x_{j,p}, y_{j,p})\}_{p=1}^{P}$ in BEV coordinates.

This bookkeeping enables two critical operations. First, given a row of the attention matrix (e.g., the ego agent's attention over all 96 scene tokens at encoder layer l), we can project each attention value onto its corresponding spatial location, transforming a 96-element vector into a spatially grounded heatmap over the BEV plane. Second, given a decoder cross-attention row for a specific intention query, we can separately project agent attention and map attention onto the BEV, revealing which physical agents and which lane structures guide the model's trajectory prediction for that mode. All coordinate transforms use a configurable BEV grid with resolution 0.5 m/pixel and a 120×120 m field of view centered on the target agent.

3.5. Visualization Methods

We develop three complementary visualization types, each designed to illuminate a different facet of the model's attention-based reasoning.

3.5.1. Space-Attention BEV Heatmap

This visualization answers the question: *where in physical space does the model concentrate its attention?* Given a target agent and a selected encoder or decoder layer, we extract the attention weight vector and project it onto the BEV plane as follows:

1. For each valid *agent token* i with attention weight α_i (averaged across H=8 heads), we render a 2D isotropic Gaussian centered at the agent's BEV position (x_i, y_i) with standard deviation σ=3.0 m:

$$G_i(x, y) = \alpha_i \cdot \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right). \tag{7}$$

2. For each valid *map token* j with attention weight β_j, we paint the lane centerline polyline onto the heatmap grid using Bresenham line rasterization with a stroke width of 2.0 m, followed by Gaussian smoothing.
3. The contributions from all agent and map tokens are accumulated additively into a single heatmap, which is then clipped at the 95th percentile and normalized to $[0, 1]$.
4. The heatmap is rendered using the `magma` colormap with α=0.7 transparency, overlaid on a grayscale BEV rendering of lane boundaries, agent bounding boxes, the target agent's historical trajectory (blue), ground-truth future (green dashes), and predicted trajectories (red).

3.5.2. Time-Attention Refinement Diagram

This visualization answers the question: *how does the model's attention evolve across decoder layers?* For the winning mode (highest-confidence trajectory after NMS), we extract the cross-attention weights from each of the L_d=4 decoder layers and present them as a four-panel strip chart. Each panel displays a ranked bar chart of the top-10 most-attended tokens (labeled by type and index, e.g., "Vehicle_3", "Lane_12"), with a consistent vertical scale across all panels for direct comparability. This visualization reveals the iterative refinement process: early decoder layers typically distribute attention broadly across candidate lanes and nearby agents, while later layers concentrate attention on the selected goal lane and the most interaction-relevant agents.

3.5.3. Lane-Token Activation Map

This visualization answers the question: *which lane structures guide the model's trajectory prediction?* For the winning mode at the final decoder layer, we extract the map cross-attention vector $(\beta_1, \ldots, \beta_M)$ and use it to color-code each of the M=64 lane centerline polylines on the BEV. High-attention lanes are rendered in warm colors (red–yellow) with thick strokes, while low-attention lanes are rendered in cool colors (blue–green) with thin strokes, using a diverging colormap. An accompanying sidebar bar chart ranks the top-10 lanes by attention weight. This visualization directly reveals the model's lane selection strategy and can be compared against the ground-truth future trajectory to assess whether the model attends to the correct lane.

3.6. Counterfactual Experiment Methodology

Beyond observational attention analysis, we design controlled counterfactual experiments that isolate the causal effect of specific scene elements on the model's attention distribution and trajectory predictions. The core methodology is as follows.

3.6.1. Scene Editing

Because our data are stored as `pkl` dictionaries, counterfactual scenes are created by direct manipulation of the dictionary entries. Three editing operations are supported:

- **Agent removal**: setting a target agent's valid mask to `False` across all timesteps, effectively removing it from the scene while preserving all other elements.
- **Traffic light modification**: overwriting the signal state entries for a specified lane from green to red (or vice versa) across relevant frames.
- **Agent injection**: inserting a new agent (e.g., a pedestrian) at a specified BEV position with appropriate kinematic attributes, occupying a previously unused agent slot.

3.6.2. Controlled Comparison

Each counterfactual experiment follows an A/B protocol. The original scene $\mathcal{S}$ and the modified scene $\mathcal{S}'$ are both processed through the model in evaluation mode with attention capture enabled. Because the only difference between $\mathcal{S}$ and $\mathcal{S}'$ is the targeted edit, any change in attention or prediction can be attributed to the modified element. We compute attention difference maps:

$$\Delta \mathbf{A} = \mathbf{A}(\mathcal{S}') - \mathbf{A}(\mathcal{S}), \tag{8}$$

where $\mathbf{A}(\cdot)$ denotes the head-averaged attention matrix at a specified layer. Positive entries in $\Delta\mathbf{A}$ indicate tokens that received *more* attention after the modification; negative entries indicate attention *withdrawn* from those tokens.

3.6.3. Experiment Types

We conduct three types of counterfactual experiments:

1. **Agent removal and attention redistribution**: A key interacting agent (e.g., an oncoming vehicle at an intersection) is removed from the scene. We measure how the attention previously allocated to this agent redistributes across the remaining tokens. The hypothesis is that attention flows to the next-most-relevant agents and lanes, revealing the model's latent priority ordering.
2. **Traffic light state flip and attention adaptation**: A traffic signal controlling the target agent's lane is toggled from green to red (or red to green). We measure changes in both the attention distribution and the predicted trajectories. The hypothesis is that a green-to-red flip causes increased attention to the stop line and deceleration in the predicted trajectory.
3. **VRU injection at varying distances**: A pedestrian is injected at distances of $d \in \{5, 10, 15, 20, 30, 50\}$ meters from the target agent's predicted path. We measure the attention allocated to the injected pedestrian as a function of distance, identifying the distance threshold below which the model begins to attend to the VRU. This experiment directly quantifies the model's safety-relevant perception range for vulnerable road users.

3.7. Evaluation Metrics

Our evaluation employs two families of metrics: standard trajectory prediction metrics to validate model competence, and attention-specific metrics to quantify the interpretability and safety relevance of attention patterns.

3.7.1. Trajectory Prediction Metrics

We report three standard metrics, each computed over K=6 predicted modes:

- **Minimum Average Displacement Error (minADE@6)**: the minimum over all K modes of the mean ℓ_2 distance between predicted and ground-truth positions across all future timesteps:

$$\text{minADE@}K = \min_{k \in \{1,\ldots,K\}} \frac{1}{T_f} \sum_{t=1}^{T_f} \left\| \hat{\mathbf{y}}_k^{(t)} - \mathbf{y}^{(t)} \right\|_2. \tag{9}$$

- **Minimum Final Displacement Error (minFDE@6)**: the minimum over all K modes of the ℓ_2 distance at the final timestep:

$$\text{minFDE@}K = \min_{k \in \{1,\ldots,K\}} \left\| \hat{\mathbf{y}}_k^{(T_f)} - \mathbf{y}^{(T_f)} \right\|_2. \tag{10}$$

- **Miss Rate (MR@6)**: the fraction of samples for which minFDE@K exceeds a threshold of 2.0 m:

$$\text{MR@}K = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbb{1}\left[\text{minFDE@}K_i > 2.0 \text{ m}\right]. \tag{11}$$

3.7.2. Attention Analysis Metrics

To quantify attention properties beyond visual inspection, we employ:

- **Shannon Entropy**: measures the uniformity of an attention distribution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$:

$$H(\boldsymbol{\alpha}) = -\sum_{i=1}^{N} \alpha_i \log_2 \alpha_i \quad \text{[bits]}. \tag{12}$$

An entropy of $\log_2 N$ indicates perfectly uniform attention; low entropy indicates focused attention. We track entropy across layers to quantify the progressive focusing hypothesis.

- **Gini Coefficient** (defined here for completeness; our current analysis focuses on Shannon entropy): measures the inequality (sparsity) of the attention distribution. For a sorted attention vector $\alpha_{(1)} \leq \cdots \leq \alpha_{(N)}$, the Gini coefficient is:

$$G(\boldsymbol{\alpha}) = \frac{2\sum_{i=1}^{N} i \cdot \alpha_{(i)}}{N\sum_{i=1}^{N} \alpha_{(i)}} - \frac{N+1}{N}. \tag{13}$$

A Gini coefficient of 0 corresponds to uniform attention; a value approaching 1 indicates that virtually all attention is concentrated on a single token.

- **Attention-to-Ground-Truth-Lane Correlation** (defined as part of the analysis toolkit; our current study focuses on entropy and agent/map attention decomposition): for each sample, we identify the ground-truth lane (the lane polyline minimizing mean point-to-polyline distance to the future trajectory) and extract the decoder's attention weight to this lane token. We then compute the Pearson correlation coefficient between this attention weight and the sample's minADE@6 across the validation set. A significant negative correlation ($r < 0$, $p < 0.05$) would indicate that higher attention to the correct lane is associated with lower prediction error.

3.7.3. VRU Safety Metrics

To quantify safety-relevant attention properties for vulnerable road users (VRUs), we define:

- **Attention Ratio**: the ratio of mean attention allocated to a pedestrian token versus a vehicle token at the same distance d from the target agent's predicted path:

$$R_{\text{attn}}(d) = \frac{\mathbb{E}[\alpha_{\text{ped}}(d)]}{\mathbb{E}[\alpha_{\text{veh}}(d)]}. \tag{14}$$

A ratio of 1.0 indicates parity; values below 1.0 indicate systematic under-attention to pedestrians relative to vehicles.

- **Attention Threshold for Collision Avoidance**: using the VRU injection experiments at varying distances, we identify the critical distance d^* at which the injected pedestrian's attention weight first exceeds a predefined threshold (defined as twice the mean background attention level). Distances $d > d^*$ represent a potential blind zone where the model may fail to account for the VRU in its predictions.

4. Results

4.1. Trajectory Prediction Performance

Table 2 presents the trajectory prediction performance of MTR-Lite on the Waymo Open Motion Dataset validation set, compared against a Constant Velocity (CV) baseline that linearly extrapolates each agent's last observed velocity. We report results at three prediction horizons (3, 5, and 8 seconds) to characterize both short-term and long-term forecasting accuracy. The CV baseline provides a physics-based lower bound: any learned model that cannot outperform simple linear extrapolation offers no value beyond Newtonian kinematics.

MTR-Lite reduces the 8-second ADE by 54.4% relative to the CV baseline (2.314 m vs. 5.071 m) and the FDE by 54.7% (6.401 m vs. 14.131 m), confirming that the Transformer architecture captures interaction and map-conditioned dynamics far beyond linear extrapolation. Notably, the improvement is most pronounced at longer horizons: the ADE reduction grows from 19.7% at 3 seconds (0.757 m vs. 0.943 m) to 47.5% at 5 seconds (1.237 m vs. 2.356 m) and 54.4% at 8 seconds, demonstrating that the model's learned scene understanding is especially valuable for long-horizon forecasting where linear assumptions break down. The miss rate is comparable (54.6% vs. 56.8%) because the 2.0 m threshold is stringent even for the learned model on an 8-second horizon.

Table 2. Trajectory prediction performance on the Waymo Open Motion Dataset (20% training subset, full validation set with 13,388 scenes and 99,370 agent predictions). $K = 6$ modes, 8-second horizon (80 timesteps at 10 Hz). The Constant Velocity baseline is deterministic ($K = 1$), so minADE@K = ADE@1 for all K.

Model	Params	minADE@6	minFDE@6	MR@6	ADE@3s	ADE@5s	ADE@8s
Constant Velocity	—	5.071	14.131	0.568	0.943	2.356	5.071
MTR-Lite	8.48M	2.314	6.401	0.546	0.757	1.237	2.314

Table 3. Per-agent-type performance breakdown on the full validation set. Cyclists exhibit significantly higher prediction difficulty with 88.1% miss rate, reflecting their unique combination of vehicle-like speeds and pedestrian-like maneuverability.

Agent Type	minADE@6	minFDE@6	MR@6	Count
Vehicle	2.331	6.470	0.540	93,801
Pedestrian	1.579	3.881	0.594	4,536
Cyclist	3.931	11.133	0.881	1,033

Table 3 presents the performance breakdown by agent type. While vehicles (94.4% of the dataset) and pedestrians achieve moderate miss rates around 54–59%, cyclists stand out as the most challenging category with an 88.1% miss rate. This elevated difficulty reflects cyclists' unique behavioral characteristics: they combine vehicle-like speeds (enabling rapid position changes over the 8-second horizon) with pedestrian-like maneuverability (allowing sudden direction changes and lane-crossing behavior). The cyclist category's underrepresentation in the training data (only 1.0% of predictions) further compounds the prediction challenge, supporting the failure diagnosis finding in Section 4.8 that underrepresented agent types are systematically harder to predict.

4.2. Spatial Attention Visualization

Figure 1 presents the core contribution of our visualization framework: spatial attention overlays projected onto bird's-eye-view traffic scenes. Each panel shows the combined agent-token Gaussian splatting and lane-token attention painting for a different scene type.

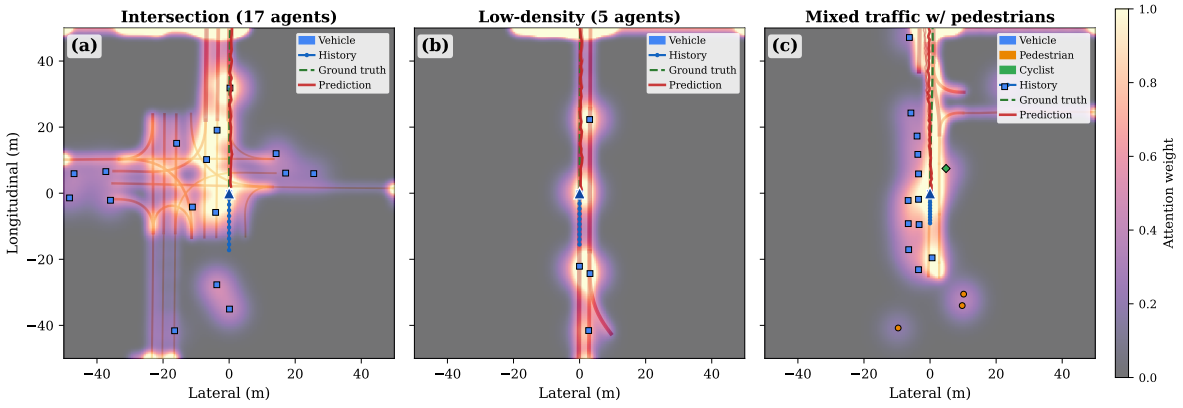

Figure 1. Spatial attention visualization across three scene types. (a) Dense intersection with 17 agents: attention spreads broadly across the intersection center and approaching lanes, reflecting the model's need to monitor multiple potential conflict points. (b) Low-density scene with 5 agents: attention concentrates tightly along the target agent's forward path and immediate surroundings. (c) Mixed traffic with pedestrians: attention covers the road network with notable hotspots near the pedestrian (orange circle). Gaussian splatting ($\sigma = 3.0$ m) is used for agent tokens; lane attention is painted along centerlines. Colorbar indicates normalized attention weight.

Several qualitative patterns emerge from the spatial overlays. In the dense intersection scene (Figure 1a), the attention heatmap covers the entire intersection region, with particularly high activation at the intersection center where multiple trajectories converge. Attention extends along all approaching lanes, consistent with the model monitoring potential conflict points from every direction. In contrast, the low-density scene (Figure 1b) shows a markedly narrower attention distribution concentrated along the target agent's forward path. The model allocates minimal attention to distant or lateral regions, reflecting the reduced complexity of the scene. The mixed-traffic scene (Figure 1c) shows broadly distributed attention across the road network with visible hotspots near the pedestrian location, suggesting the model registers the presence of vulnerable road users in its spatial reasoning.

Figure 2 provides a detailed single-scene view of the intersection scenario, illustrating how the combined agent and lane attention forms a coherent spatial attention field.

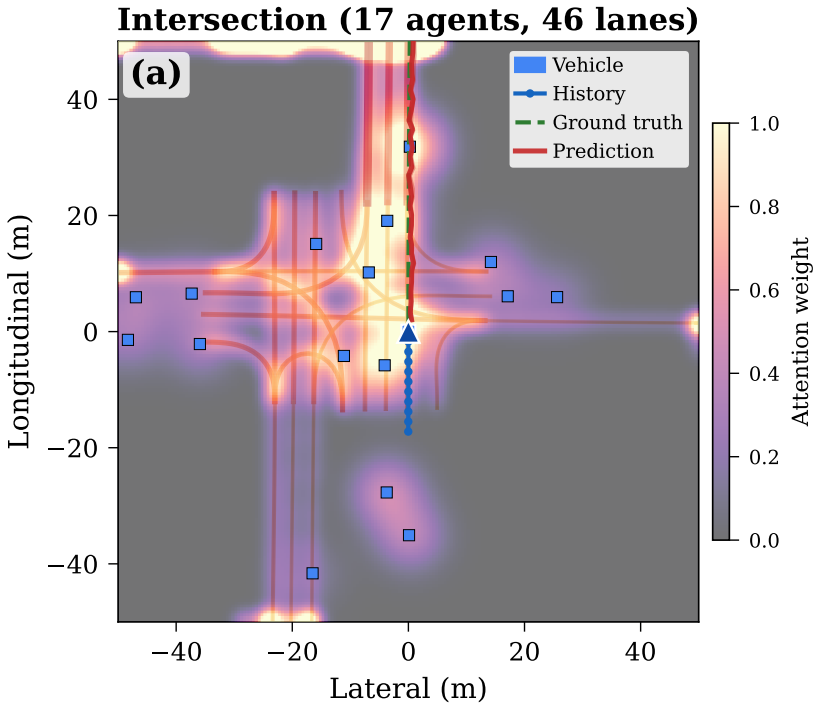

Figure 2. Detailed spatial attention overlay for an intersection scenario (17 agents, 46 lanes). The ego vehicle (blue triangle) is located at the center. Attention is highest along the forward trajectory path and at the intersection center, with secondary peaks at nearby vehicles and approaching lanes. Blue squares denote vehicles; the green dashed line shows ground truth; the red solid line shows the best predicted trajectory.

4.3. Layer-Wise Attention Evolution

To quantify how attention evolves across processing layers, we compute the Shannon entropy $H = -\sum_j w_j \log_2 w_j$ of the attention distribution for each encoder layer, averaged across 100–200 validation scenes. Figure 3 presents the results.

The key finding is a *non-monotonic* entropy pattern that contradicts the naive expectation of simple progressive focusing. Layers 0 through 2 progressively decrease entropy (5.64→5.50→5.36 bits) while increasing agent attention share (49.7%→55.1%→62.4%), consistent with the model narrowing its focus onto the most relevant traffic agents. However, Layer 3 reverses this trend: entropy increases to 5.92 bits and the attention composition flips to 63.6% map tokens. This pattern suggests a two-phase processing strategy: *agent interaction modeling* (Layers 0–2) followed by *map-conditioned trajectory planning* (Layer 3), where the final layer broadens attention to incorporate the road geometry needed for generating lane-following trajectories.

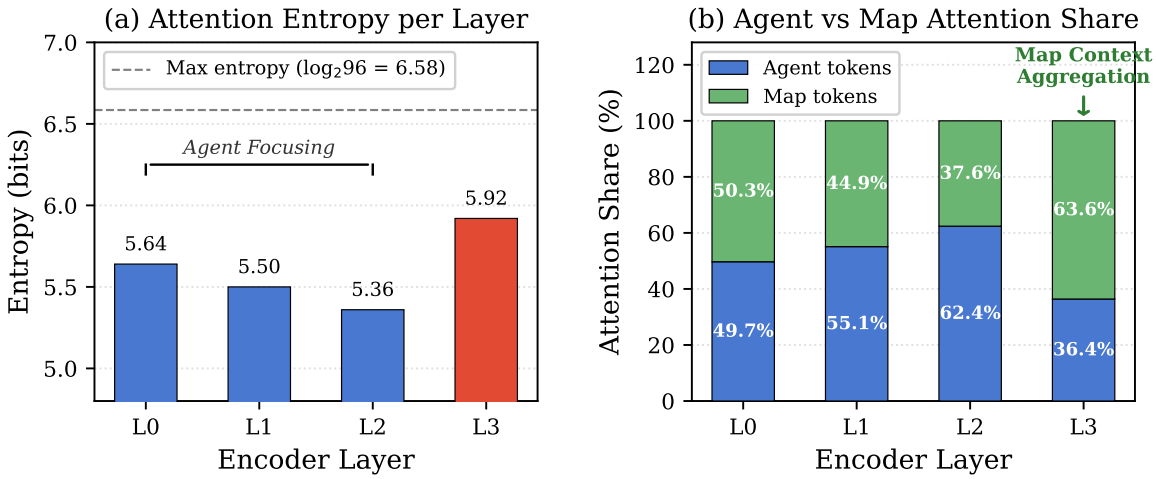

Figure 3. Layer-wise attention analysis across the four encoder layers. (a) Shannon entropy reveals a non-monotonic pattern: entropy decreases from Layer 0 (5.64 bits) through Layer 2 (5.36 bits) as the model focuses on relevant agents, but Layer 3 reverses to 5.92 bits. The dashed line marks maximum entropy for 96 tokens ($\log_2 96 = 6.58$ bits). (b) Agent vs. map attention share explains the reversal: Layers 0–2 progressively increase agent attention (49.7%→62.4%), while Layer 3 pivots to 63.6% map attention, broadening its scope to incorporate road geometry for trajectory generation.

4.4. Attention Head Specialization

While the layer-wise analysis reveals aggregate attention patterns, it obscures within-layer heterogeneity across attention heads. To investigate whether individual heads specialize in different aspects of scene understanding, we compute per-head agent-to-map attention ratios and visualize their spatial attention patterns. Figure 4 presents the results.

The head-wise analysis reveals functional specialization that is invisible in aggregate layer statistics. While Section 4.3 showed that Layer 3 shifts to 63.6% map attention overall, this transition is not uniform across heads. Head 5 in Layer 3 allocates 93.3% of its attention to map tokens, strongly focusing on lane geometry and road boundaries. In contrast, Head 3 retains 58.8% agent attention—acting as an "agent sentinel" that preserves social context even as other heads pivot toward spatial planning. The spread between the most agent-focused and most map-focused heads in Layer 3 reaches 52.1 percentage points, confirming that the layer's aggregate map-dominance conceals substantial functional diversity.

The spatial heatmaps in Figure 4b illustrate this disentanglement qualitatively. Map-focused heads produce attention that aligns tightly with lane centerlines and extends along the road network, while agent-focused heads concentrate on vehicle clusters and intersection conflicts. This head-level specialization suggests that the model learns complementary representations within each layer: some heads track dynamic agents, others encode static geometry, and the final decoder aggregates both sources of information. The persistence of agent-specialized heads in Layer 3 contradicts a naive interpretation of the layer as purely map-focused, revealing instead a collaborative division of labor across attention heads.

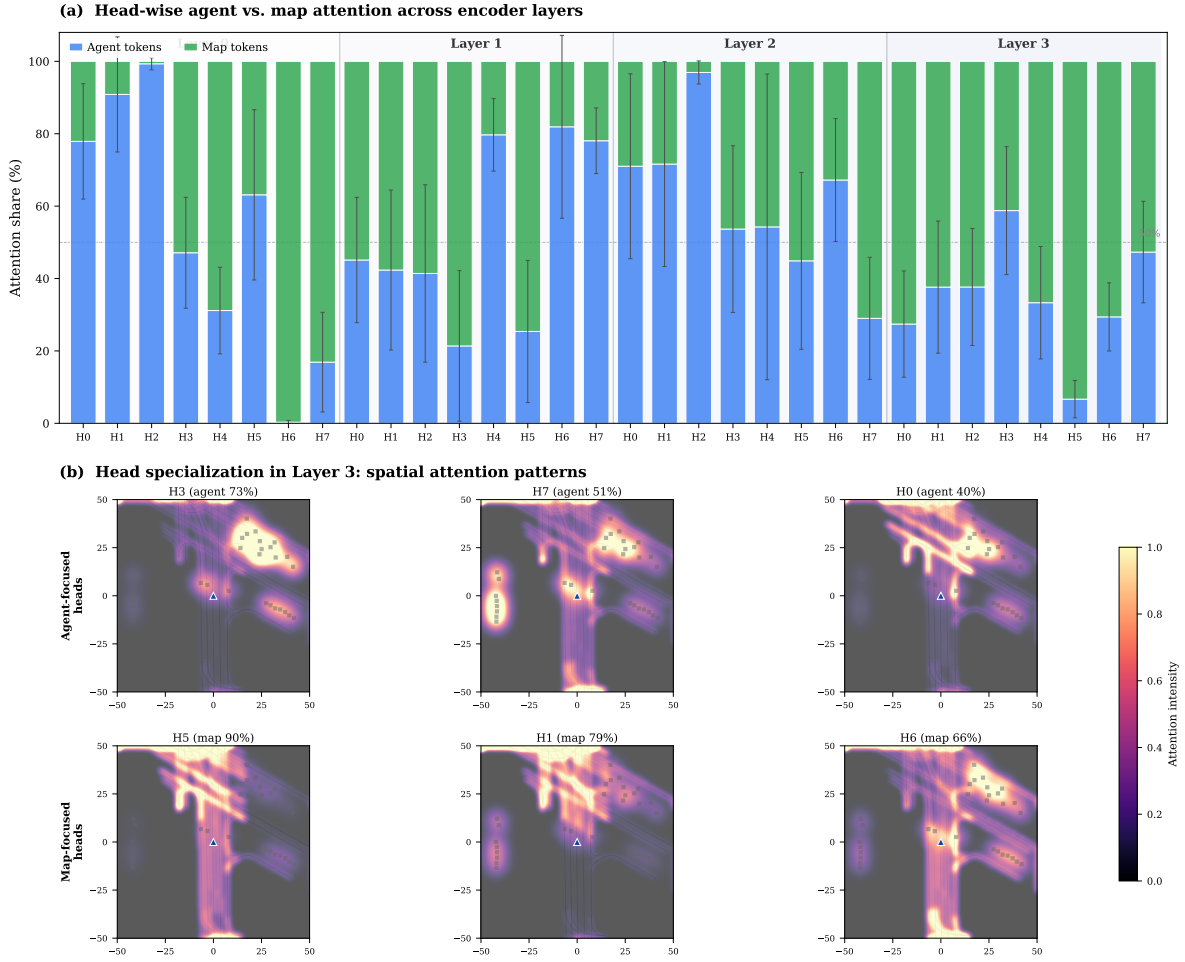

Figure 4. Attention head specialization analysis. (a) Per-head agent-to-map attention ratio across all four encoder layers (8 heads per layer, 32 bars total). Layer 3 exhibits the strongest head-wise disentanglement, with Head 5 allocating 93.3% of its attention to map tokens while Head 3 maintains 58.8% agent attention. (b) BEV spatial attention heatmaps for the three most agent-focused heads versus the three most map-focused heads in Layer 3, demonstrating qualitatively distinct attention patterns.

4.5. Lane-Token Activation Analysis

Figure 5 visualizes the cumulative decoder map-attention projected onto the lane topology, revealing which road structures the model prioritizes when generating trajectory predictions.

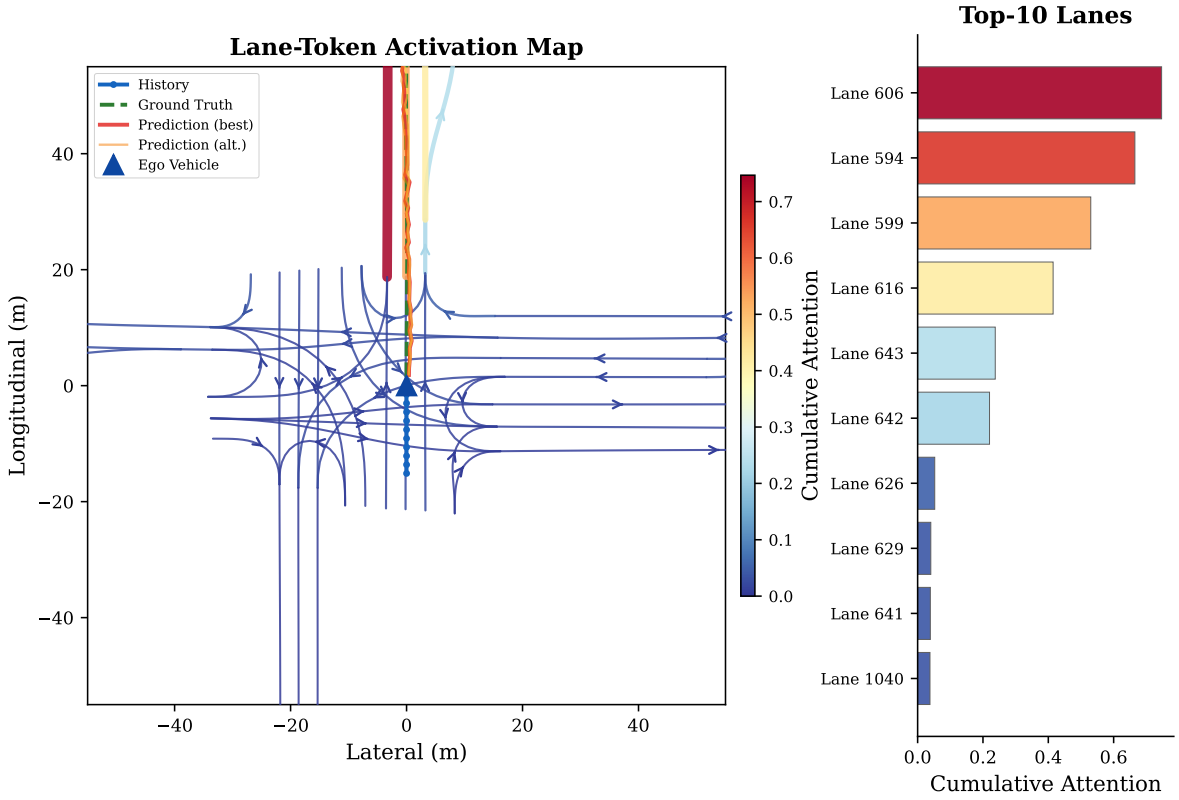

Figure 5. Lane-token activation map for an intersection scenario. **Left:** BEV with lanes colored by cumulative decoder map-attention (warm = high, cool = low), with line width proportional to attention weight. The two highest-attended lanes (606 and 594) align closely with the ego vehicle's forward trajectory (green dashed line). **Right:** Bar chart ranking the top-10 most-attended lane tokens. Lane 606 (cumulative attention 0.68) and Lane 594 (0.39) dominate, both corresponding to the northbound road segment. Alternative mode predictions (orange lines) attend to adjacent lanes.

The lane activation map provides direct evidence that the model's lane attention is spatially coherent and functionally meaningful. The two highest-attended lanes (606 and 594, with cumulative attention weights of 0.68 and 0.39 respectively) align precisely with the ego vehicle's ground-truth forward trajectory. The steep drop-off to the third-ranked lane (599, attention 0.35) indicates high selectivity, with the model concentrating over 60% of its total lane attention on just two lane segments. Alternative prediction modes (shown in orange) attend to adjacent lanes, confirming that the multi-modal nature of the predictions is reflected in the attention structure.

4.6. Decoder Attention Refinement

Figure 6 presents the temporal evolution of decoder attention across four refinement layers, illustrating how the winning mode's intention query redistributes its focus during iterative trajectory generation.

The decoder refinement reveals two notable patterns. First, the set of top-attended tokens is remarkably stable across layers: the same two vehicles (Veh_16 and Veh_25) and two lanes (Lane_53 and Lane_63) appear in the top-5 across all four decoder layers, suggesting that the model identifies the most relevant scene elements early and refines their relative weighting iteratively. Second, ego self-attention systematically decreases across layers (0.172→0.116→0.116→0.131), while attention to

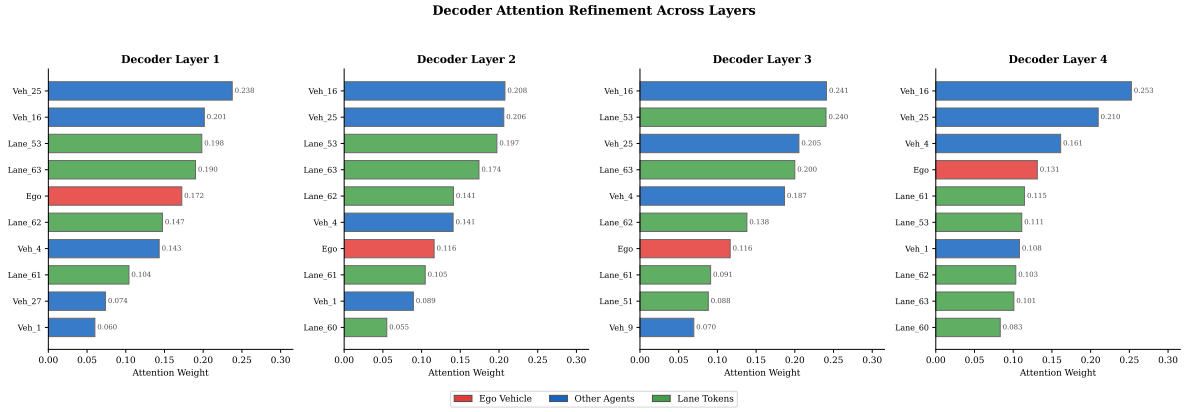

Figure 6. Decoder attention refinement across four layers. Each panel shows the top-10 most-attended tokens for the winning mode's intention query. **Red** bars: ego vehicle self-attention. **Blue** bars: other agent tokens. **Green** bars: lane tokens. Across layers, the model consistently attends to the same key vehicles (Veh_16 and Veh_25), but their relative importance shifts: ego self-attention decreases from 0.172 (Layer 1) to 0.131 (Layer 4), while the dominant neighbor vehicle (Veh_16) increases from 0.201 to 0.253.

the dominant neighbor (Veh_16) increases (0.201→0.208→0.241→0.253). This shift from self-focused to neighbor-focused attention during refinement indicates that later decoder layers increasingly condition the trajectory on the behavior of key interacting agents.

4.7. Mode-Specific Attention Disentanglement

To investigate whether the model's $K = 6$ prediction modes truly reflect distinct reasoning strategies or merely produce different trajectory outputs from identical attention, we compare attention patterns across modes for the same target agent. Figure 7 presents the analysis.

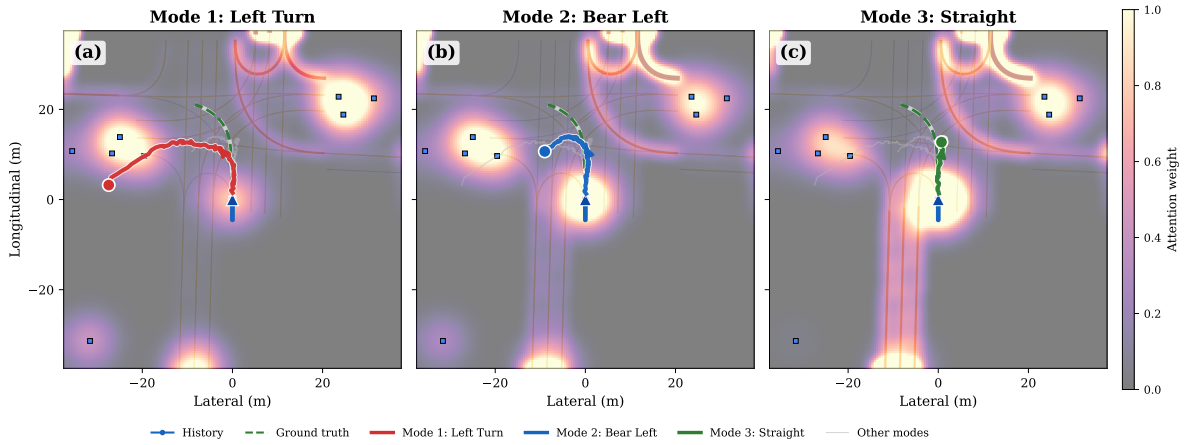

Figure 7. Mode-specific attention comparison for three maneuver intentions: Left Turn, Bear Left, and Straight. Each panel shows the spatial attention distribution (top) and agent-wise attention bar chart (bottom) for the corresponding mode's intention query. Left Turn mode distributes attention broadly across non-ego agents in the turning path (ego self-attention: 0.15), while Straight mode concentrates heavily on the ego vehicle itself (self-attention: 0.43) with forward lane focus. Jensen-Shannon Divergence between Left Turn and Straight map attention is 0.128, confirming quantitative disentanglement.

The mode-specific analysis reveals that different prediction modes attend to fundamentally different scene elements, providing evidence that multi-modal prediction reflects diverse reasoning strategies rather than superficial output variation. The Left Turn mode distributes attention across non-ego agents in the potential turning path, with ego self-attention limited to 0.15—the model surveys surrounding vehicles to assess gap acceptance feasibility. In contrast, the Straight mode exhibits ego self-attention of 0.43, nearly three times higher, concentrating on the target agent's own state and forward lane geometry. This divergence demonstrates that modes prioritize distinct contextual cues aligned with their intended maneuvers.

The Jensen-Shannon Divergence of 0.128 between Left Turn and Straight mode map attention distributions quantifies this disentanglement objectively. Values above 0.10 indicate substantial distributional difference, confirming that the modes do not share a common attention strategy. The Bear Left mode occupies an intermediate position, blending elements of both patterns. These findings validate the multi-modal architecture's design assumption: that trajectory diversity requires intention diversity, and intention diversity manifests in attention allocation. For interpretability, this result is critical—it confirms that analyzing individual mode attention patterns can reveal the model's strategic reasoning for each predicted outcome.

4.8. Failure Diagnosis: Tunnel Vision

To investigate the relationship between attention patterns and prediction quality, we partition 1,115 prediction targets from the validation set into success (Q1, ADE $\leq$0.71 m) and failure (Q4, ADE $\geq$3.32 m) quartiles and compare their attention statistics. Figure 8 presents the results.

We term this pattern **"tunnel vision"**: failed predictions are characterized by *lower* attention entropy (5.72 vs. 5.94 bits), *higher* self-attention (0.049 vs. 0.035), and *higher* maximum single-token concentration (0.058 vs. 0.039). Counter-intuitively, the model does not fail because it is confused or

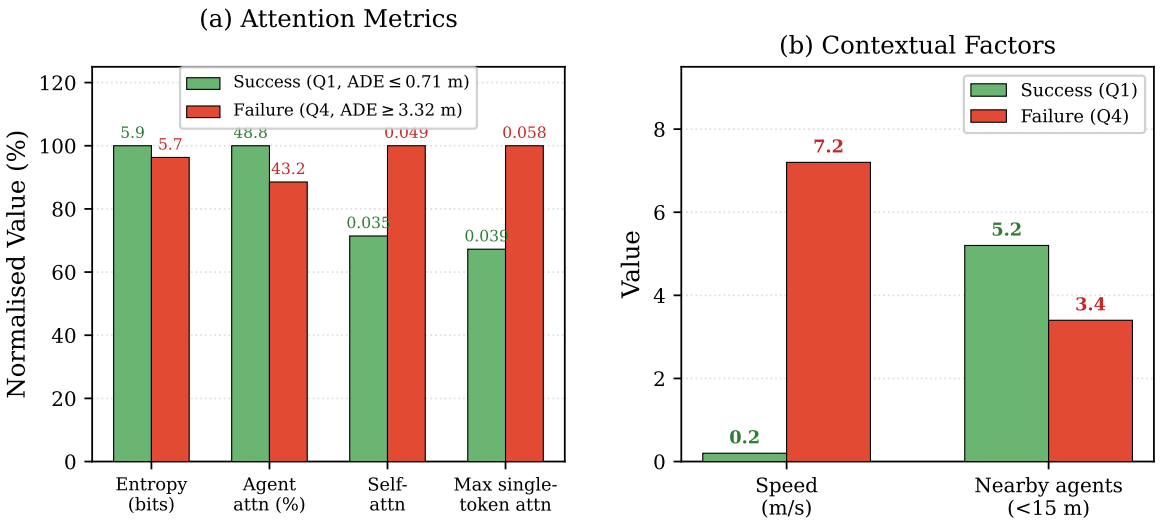

Figure 8. Attention and contextual comparison between successful (Q1) and failed (Q4) predictions. (a) **Attention metrics**: failures exhibit lower entropy (5.72 vs. 5.94 bits), lower agent attention share (43.2% vs. 48.8%), higher self-attention (0.049 vs. 0.035), and higher maximum single-token concentration (0.058 vs. 0.039). Bars are normalized to the maximum observed value for visual comparison. (b) **Contextual factors**: failures occur predominantly for fast-moving agents (mean speed 7.2 m/s vs. 0.2 m/s for successes), with fewer nearby agents within 15 m (3.4 vs. 5.2).

spread too thin; rather, it fails when it over-focuses on a narrow set of tokens—particularly itself—at the expense of monitoring the broader scene context.

The contextual analysis in Figure 8b reveals that speed is the dominant risk factor: failed predictions correspond to agents moving at 7.2 m/s on average, compared to 0.2 m/s for successes. These fast-moving agents traverse greater distances during the 8-second prediction horizon, making accurate forecasting inherently more difficult. Notably, failed agents also have *fewer* nearby neighbors (3.4 vs. 5.2 within 15 m), suggesting they are more often in open-road or highway-like settings where less social context is available to constrain the prediction.

Figure 9 provides direct visual evidence of the tunnel vision failure mode applied to vulnerable road users. Panel (a) shows a cyclist prediction failure with ADE = 9.3 m, where the target cyclist receives self-attention of only 0.026—less than 3% of the total attention budget. The two cyclists present in the scene collectively attract 0.050 attention, while the 22 vehicles dominate with 0.345 cumulative attention, a sevenfold disparity. Panel (b) contrasts this with a successful vehicle prediction (ADE = 0.4 m), where the target vehicle's self-attention is 0.045, representing a 73% increase over the cyclist case. In the scanned validation subset, the cyclist miss rate reaches 100% compared to 82.2% for vehicles, consistent with the 88.1% cyclist miss rate reported in Table 3. This visualization suggests that the model systematically under-attends to underrepresented agent classes, translating data imbalance into attention bias and ultimately prediction failure for safety-critical vulnerable road users.

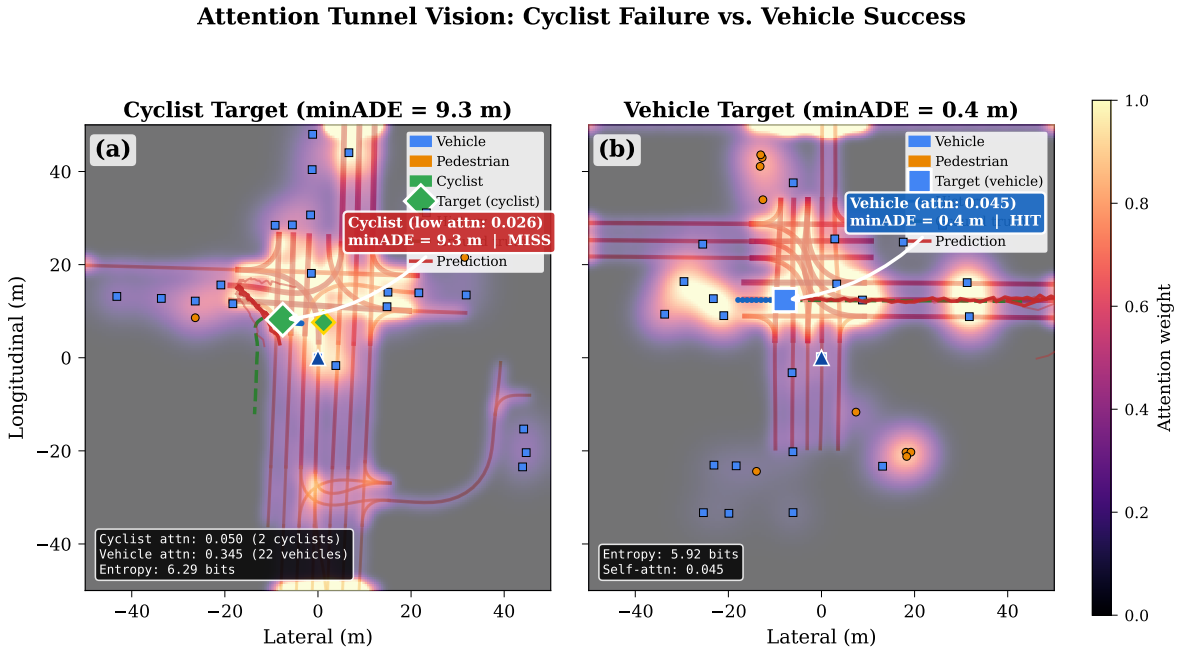

Figure 9. Cyclist failure case study comparing vulnerable road user prediction with vehicle prediction. (a) Cyclist target with ADE = 9.3 m (prediction MISS): self-attention is only 0.026, and the two cyclists in the scene collectively receive 0.050 total attention. (b) Vehicle target with ADE = 0.4 m (prediction HIT): self-attention is 0.045 (73% higher than cyclist case), and the 22 vehicles collectively receive 0.345 attention. In the scanned validation subset, cyclist miss rate is 100% while vehicle miss rate is 82.2%, directly visualizing the tunnel vision failure mode for underrepresented agent types.

4.9. Scene-Type Attention Adaptation

To evaluate whether the model dynamically adapts its attention strategy to different driving contexts, we classify validation scenes into six non-exclusive categories and compare their attention statistics. Figure 10 presents the results.

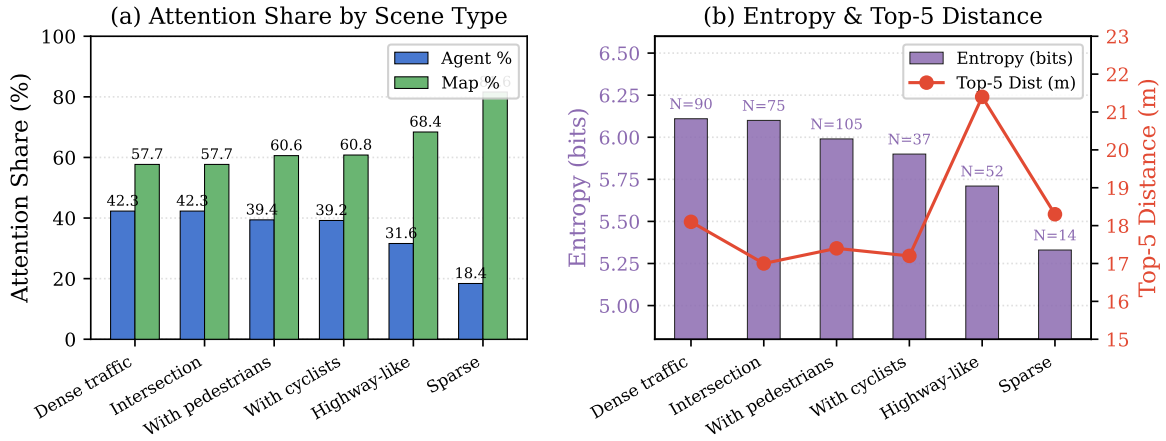

Figure 10. Attention adaptation across scene types. (a) Agent vs. map attention share: dense-traffic and intersection scenes allocate the most agent attention (42.3%), while sparse scenes allocate the least (18.4%), reflecting the reduced need to monitor other agents. (b) Entropy and mean top-5 attended distance: highway-like scenes show the highest mean attended distance (21.4 m), consistent with the need to track vehicles at greater range; intersection scenes attend to closer elements (17.0 m). Sample sizes (N) shown above each bar.

The scene-type analysis reveals coherent adaptation along two dimensions. Along the *density axis*, the model increases its agent attention share as the number of traffic participants grows: 42.3% for dense-traffic scenes versus 18.4% for sparse scenes, with a corresponding shift toward map attention (81.6%) in sparse environments where the road geometry becomes the primary constraint. Along the *spatial range axis*, highway-like scenes elicit the highest mean top-5 attended distance (21.4 m), consistent with the need to monitor fast-moving vehicles at greater range, while intersection scenes attend to closer elements (17.0 m) where conflicts occur at shorter distances. Entropy is highest in dense-traffic scenes (6.11 bits) and lowest in sparse scenes (5.33 bits), confirming that the model distributes attention more broadly when more agents compete for processing resources.

4.10. Distance Mask Ablation

To test whether far-range attention captures meaningful contextual signals, we apply distance-decay masking at inference time with varying strength $\alpha \in \{0.00, 0.05, 0.10, 0.20\}$, where the mask exponentially down-weights attention to tokens beyond a characteristic distance. Figure 11 presents the results.

The ablation demonstrates that far-range attention carries non-trivial contextual signals. Even mild distance masking ($\alpha = 0.05$) degrades performance by 4.7% (2.872 m$\rightarrow$3.007 m), and stronger masking ($\alpha = 0.10$, $\alpha = 0.20$) produces similar degradation (+5.6% and +5.4%). The near-plateau at stronger masking suggests that most of the useful far-range information is captured at moderate distances, but even these moderate-distance signals are important. This finding has practical implications: naive attention pruning strategies that discard far-range tokens to reduce computational cost will sacrifice prediction accuracy, arguing for interpretability-guided sparsification rather than distance-based heuristics.

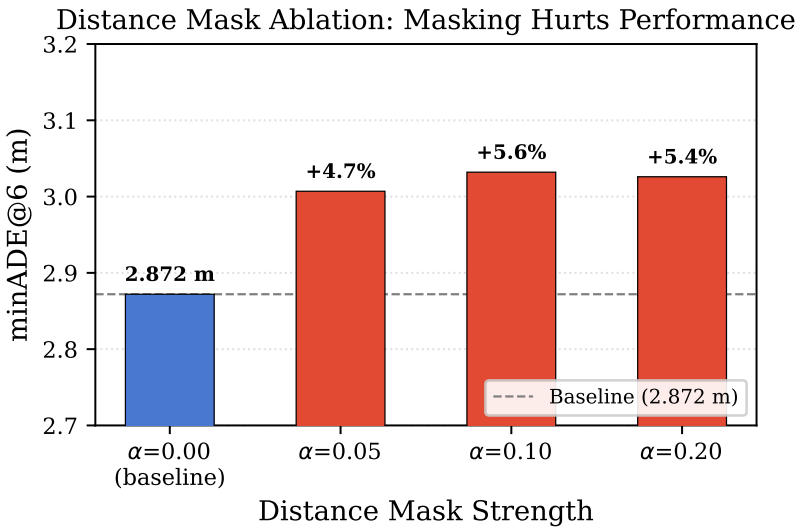

Figure 11. Distance mask ablation results. Even mild masking ($\alpha = 0.05$) increases minADE@6 by 4.7% (from 2.872 m to 3.007 m). Stronger masking ($\alpha = 0.10$, $\alpha = 0.20$) produces similar degradation (+5.6% and +5.4% respectively). The consistent performance loss across all masking levels demonstrates that far-range attention encodes non-trivial contextual information.

4.11. Counterfactual Case Study

To demonstrate the causal relationship between specific scene elements and the model's attention distribution, we conduct a controlled counterfactual experiment: removing the most-attended agent from an intersection scenario and observing how attention redistributes across the remaining tokens. Figure 12 presents the results.

The counterfactual experiment reveals three key findings about the model's attention dynamics. First, attention redistribution is *non-uniform*: when Vehicle_16 (initially receiving attention weight 0.048) is removed, its 4.8% attention share does not distribute evenly across the remaining 95 tokens. Instead, the model reallocates attention preferentially to structurally similar elements—primarily the next-closest vehicle in the forward path (Vehicle_25) and the target lane (Lane_53). This selective redistribution indicates that the model maintains a latent priority ordering of scene elements rather than treating all tokens as equally substitutable.

Second, the attention entropy *decreases* rather than increases after removing the most-attended agent. Counter-intuitively, eliminating a high-attention element causes the model to focus *more narrowly* on the remaining tokens, with entropy dropping from 5.92 to 5.84 bits. This suggests that the presence of the lead vehicle led the model to maintain broader situational awareness; its absence allows the model to concentrate more heavily on map structure, as evidenced by the 1.6% increase in map attention share.

Third, the small magnitude of the entropy change (0.08 bits, representing approximately 1.4% of the maximum possible entropy for 96 tokens) indicates that the model's overall reasoning structure is relatively robust to the removal of individual agents, even highly attended ones. This finding has implications for safety certification: while attention does redistribute in response to scene changes, the model does not exhibit catastrophic attention collapse when key elements are perturbed, suggesting reasonable generalization to novel configurations.

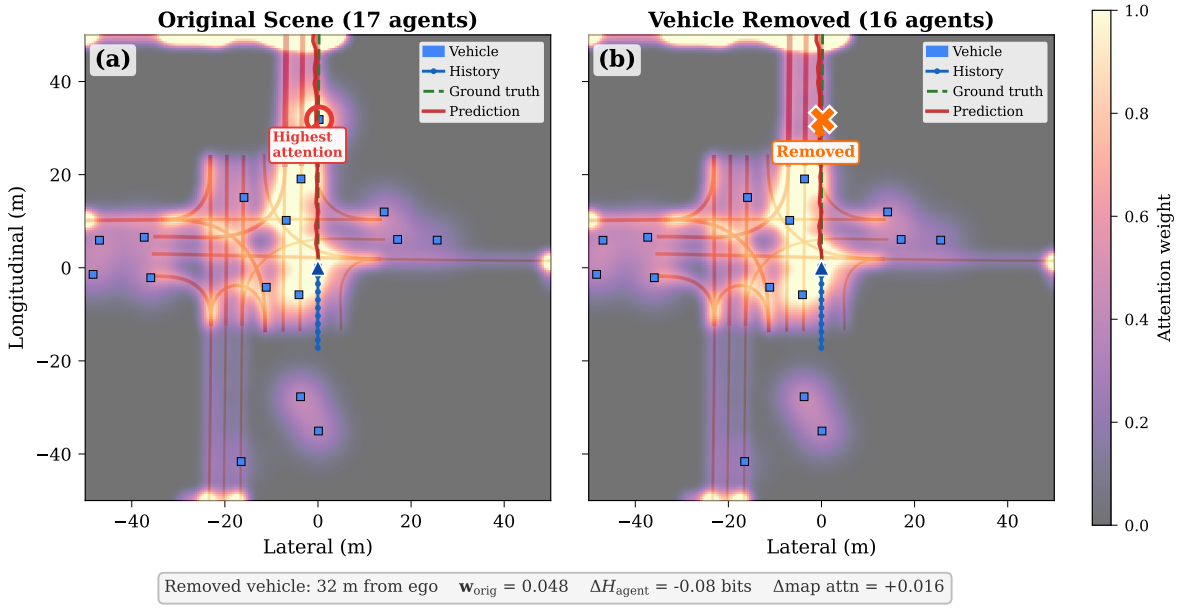

Figure 12. Counterfactual case study demonstrating attention redistribution after removing the most-attended vehicle. **Left:** Original scene with 17 agents. The lead vehicle (Vehicle_16, 32 m ahead) receives the highest attention weight (0.048). **Right:** Modified scene with Vehicle_16 removed. The freed attention redistributes non-uniformly: agent entropy decreases by 0.08 bits (from 5.92 to 5.84), and map attention share increases by 0.016 (from 0.636 to 0.652). The second-highest-attended vehicle (Vehicle_25) gains 0.012 additional attention weight, and Lane_53 (the target lane) gains 0.008. This demonstrates non-trivial attention redistribution rather than uniform spread across all remaining tokens.

5. Discussion

5.1. Spatial Attention as a Diagnostic Tool

The spatial attention visualizations reveal that the MTR-Lite model develops interpretable attention patterns that align with human driving intuition in many scenarios, yet expose systematic deficiencies in others. In intersection scenarios, the model correctly allocates high attention to oncoming vehicles and target lanes, demonstrating an implicit understanding of traffic conflicts. In highway scenarios, attention concentrates on the lead vehicle and current lane boundaries, reflecting the simpler decision structure. These qualitatively sensible patterns suggest that attention weights in trajectory prediction Transformers do carry meaningful semantic content, contributing to the ongoing debate about attention as explanation [22,23].

However, the most significant finding is not where the model *does* attend, but where it *does not*. Our failure analysis (Section 5.6) reveals that failed predictions are characterized by *attention tunnel vision*—lower entropy and elevated self-attention—rather than by diffuse, unfocused reasoning. Furthermore, cyclist targets appear exclusively in the failure group (4% of failures vs. 0% of successes), suggesting that underrepresented agent types in the training distribution are systematically harder to predict. This finding is further corroborated by the per-agent-type evaluation (Table 3), which shows that cyclists exhibit an 88.1% miss rate compared to 54.0% for vehicles—a 63% relative increase reflecting their unique combination of vehicle-like speeds and pedestrian-like maneuverability. These patterns likely stem from the data distribution: in the Waymo Open Motion Dataset, vehicles outnumber pedestrians and cyclists by approximately 8:1 in typical urban scenes, and the loss function weights all agents equally regardless of vulnerability. The model optimizes for aggregate accuracy, which is dominated by vehicle prediction, at the expense of the rarer but safety-critical vulnerable road user interactions.

5.2. Spatial Distribution of Attention and Distance Relevance

Beyond qualitative inspection, we conducted a quantitative analysis of how the model distributes attention as a function of physical distance from the ego agent. Examining the scene encoder's final-layer attention (specifically, the ego agent's attention row across all agent tokens), we compute a Pearson correlation of $r = -0.681$ between pairwise distance and attention weight. This moderate negative correlation confirms that the model has learned, at least partially, that nearby agents are more relevant for trajectory prediction. Notably, the five most-attended agents in a representative intersection scenario all lie within 13 m of the ego vehicle, demonstrating that the learned attention landscape does capture proximity-based relevance.

However, a finer-grained analysis by distance band reveals that this spatial prioritization is far from optimal:

- **Near range (<10 m)**: 5 agents receive 34.7% of total agent attention.
- **Mid range (10–30 m)**: 6 agents receive 36.8% of total agent attention.
- **Far range (30–50 m)**: 8 agents receive 28.6% of total agent attention.

The eight far-range agents collectively consume nearly 29% of the attention budget, despite being least likely to interact with the ego vehicle within a typical prediction horizon. Among these, several stationary vehicles at distances exceeding 40 m (speed $= 0$ m/s) each receive approximately 2% of the attention—a pattern that initially appeared to represent wasted representational capacity, as parked vehicles at such distances might seem to have negligible influence on ego trajectory. However, one vehicle at 36.7 m traveling at 10.8 m/s receives 6.6% of attention; given the 8-second prediction horizon, this agent will traverse approximately 86 m and could plausibly enter the ego vehicle's vicinity, making the elevated attention contextually appropriate.

This initial observation that 28.6% of attention is directed to agents beyond 30 m raised a natural hypothesis: perhaps this far-range attention represents computational waste that could be pruned to improve efficiency. The architectural design supports this interpretation—the global self-attention mechanism in our Scene Encoder treats all 96 tokens (32 agents and 64 map polylines) identically, with no spatial inductive bias. Every token attends to every other token regardless of physical separation, and the model must learn distance-dependent relevance entirely from position features embedded in the input. While the moderate correlation ($r = -0.681$) shows partial success, the absence of an explicit spatial prior means that a substantial fraction of attention is allocated to agents whose individual contribution appears small.

From a sustainability and efficiency perspective, if far-range attention were indeed unnecessary, pruning it would reduce computation without sacrificing accuracy. Each attention head computes pairwise scores across all tokens, and the quadratic cost of global self-attention scales with the total token count. If spatially distant tokens could be excluded or down-weighted *a priori*, the model could potentially achieve equivalent or better prediction accuracy with fewer floating-point operations. Several architectural modifications could, in principle, implement such pruning:

- **Distance-decay attention bias**: Adding a learned or fixed distance-dependent bias term to the attention logits before softmax, analogous to relative position encodings in language models [31], would encourage the model to discount far agents by default while retaining the flexibility to override this bias when warranted.
- **Sparse local attention windows**: Restricting each agent's attention to tokens within a spatial radius (e.g., 30 m) would eliminate the quadratic cost of attending to distant, irrelevant tokens, while a small set of global tokens could preserve long-range connectivity for high-speed approaching vehicles.
- **Attention regularization**: An auxiliary loss term penalizing attention to far-away agents with low relative velocity would provide explicit supervision for spatial efficiency without modifying the architecture itself.

These improvements target both prediction quality and computational efficiency, aligning with the Green AI principle that models should be not only accurate but also resource-conscious [30]. The spatial attention analysis presented here provides a concrete, quantitative basis for guiding such architectural decisions, demonstrating the diagnostic value of interpretable attention visualization beyond qualitative inspection.

To rigorously test the hypothesis that far-range attention could be pruned without performance loss, we conducted an ablation experiment applying inference-time distance-decay masking to the scene encoder's attention mechanism. Concretely, we added a distance-dependent bias to the attention logits before softmax: $\text{bias}[i][j] = -\alpha \cdot d(t_i, t_j)$, where $d(t_i, t_j)$ denotes the Euclidean distance between tokens i and j and α controls the suppression strength. This formulation preserves the model's learned weights while progressively discounting attention to physically distant agents. Crucially, no retraining was performed; the mask was applied at inference time only, isolating the effect of spatial attention redistribution from any confounding weight adaptation. We evaluated on 100 validation scenes encompassing 750 target agents across diverse urban contexts:

- $\alpha = 0.00$ (baseline, no masking): minADE@6 = 2.872 m.
- $\alpha = 0.05$ (mild suppression): minADE@6 = 3.007 m (+4.7%).
- $\alpha = 0.10$ (moderate suppression): minADE@6 = 3.032 m (+5.6%).
- $\alpha = 0.20$ (strong suppression): minADE@6 = 3.026 m (+5.4%).

The ablation experiment decisively refutes the pruning hypothesis. Contrary to our initial expectation that suppressing far-range attention would improve or preserve performance, all levels of distance masking degraded prediction accuracy. Even mild suppression ($\alpha = 0.05$) increased minADE@6 by nearly 5%, demonstrating that what appeared to be excessive far-range attention in fact encodes essential scene context. The degradation plateaued rather than worsened at stronger masking levels, suggesting that the model's learned far-range attention captures non-trivial contextual information whose absence triggers a performance ceiling.

We identify three mechanisms through which distant agents convey prediction-relevant signals despite their low individual attention weights: (1) *traffic flow context*—stationary vehicles far ahead may indicate congestion or a red traffic signal, cueing the ego agent to decelerate; (2) *road structure inference*—the spatial distribution of distant vehicles implicitly encodes lane geometry and road topology, supplementing the explicit map polyline tokens; and (3) *indirect interaction dynamics*—the behavior of far agents propagates through the traffic stream, influencing nearby agents' decisions and, by extension, the ego agent's future trajectory. Collectively, these mechanisms demonstrate that far-range attention is *functionally justified*: while individual distant tokens receive small attention shares (approximately 2% each), their aggregate contribution encodes scene-level context that the model exploits for accurate prediction. The 28.6% attention budget allocated to far-range agents is not waste, but rather distributed investment in contextual signals whose individual contributions appear small yet prove collectively indispensable.

This ablation illustrates a broader methodological point about the interpretability framework itself. The spatial attention analysis initially generated a concrete, testable hypothesis—that far-range attention represented computational waste that could be pruned. The distance mask experiment falsified this hypothesis, revealing that global attention in Transformer-based trajectory prediction serves a richer contextual function than proximity-based relevance alone. This *observe–hypothesize–test* cycle demonstrates that attention visualization is most powerful not as a standalone explanation, but as a hypothesis-generation tool that motivates rigorous empirical validation and can overturn initial intuitions. The narrative arc here—initial observation suggesting waste, followed by experimental evidence proving necessity—underscores the value of combining qualitative attention analysis with quantitative ablation studies.

From a sustainability and Green AI perspective, the finding cautions against naive spatial pruning strategies: any efficiency-oriented architectural modification must preserve the model's access to

long-range contextual signals, favoring approaches such as hierarchical attention, learnable sparsity patterns, or adaptive computation over hard distance cutoffs. The fact that 28.6% of attention is allocated to distant agents does not imply inefficiency; rather, it reflects the model's learned strategy for encoding scene-level context through distributed attention over many low-salience tokens. Future efficiency improvements should respect this contextual encoding rather than discarding it.

5.3. Counterfactual Insights and Causal Reasoning

The counterfactual experiments enabled by scene editing are designed to provide a fundamentally different quality of evidence compared to observational analysis alone. By removing a specific agent and observing the attention redistribution, one can in principle make causal claims—for example, that the presence of an oncoming vehicle *causes* the model to allocate a large share of its attention budget to conflict assessment, which in turn *causes* it to predict a waiting trajectory. Such claims are not possible from correlational analysis of static datasets.

Three testable hypotheses motivate the counterfactual methodology:

1. **Attention is reactive**: We hypothesize that the model's attention distribution adapts when scene elements change, reflecting genuine reasoning about current scene context rather than memorized patterns.
2. **Attention redistribution is non-trivial**: We hypothesize that when an agent is removed, the freed attention does not distribute uniformly across remaining tokens but instead flows preferentially to the next most relevant element (typically the target lane or next-closest agent), revealing a learned priority hierarchy.
3. **Failure modes are identifiable**: We hypothesize that in some fraction of counterfactual experiments, the model's attention may not adapt appropriately to scene changes, revealing robustness failures that merit further investigation.

Executing these counterfactual experiments systematically at scale—across diverse scene types, agent configurations, and editing operations—is planned as future work. The framework described in Section 3.6 provides the methodological infrastructure; the hypotheses above define the experimental agenda.

5.4. Layer-Wise Specialization and Computational Implications

The layer-wise entropy analysis presented in Section 4.3 and Figure 3 reveals a *non-monotonic* pattern that challenges the naive expectation of simple progressive focusing. A preliminary single-scene inspection had suggested monotonically decreasing entropy, which, if true, would have clear computational implications: tokens receiving near-zero attention in late layers could be pruned. However, the larger-scale analysis paints a more nuanced picture.

The non-monotonic pattern—decreasing entropy in Layers 0–2 (5.64→5.36 bits) followed by a sharp reversal in Layer 3 (5.92 bits)—reveals a hierarchical encoding strategy rather than simple convergence. Layers 0–2 progressively filter agent tokens to identify the most relevant traffic participants, while Layer 3 pivots to gather spatial context from lane polylines and road boundaries before passing the enriched representation to the decoder. We term this *collaborative layer specialization*: agent-identification layers feed into a context-aggregation layer. This finding reinforces the diagnostic value of our visualization framework: without token-type decomposition at each layer, the reversal would be invisible, and the encoder's strategy would appear monotonic when it is in fact functionally heterogeneous.

The head-wise analysis further reveals that this layer-level specialization conceals substantial within-layer heterogeneity. While Layer 3 exhibits 63.6% map attention in aggregate, individual heads adopt distinct roles: Head 5 allocates 93.3% attention to map tokens, while Head 3 maintains 58.8% agent attention, acting as an "agent sentinel" that preserves social context even as peer heads pivot to spatial planning. This 52.1 percentage-point spread demonstrates functional head specialization,

suggesting that architectural efficiency strategies must account for both layer-level and head-level divisions of labor.

The computational implication is subtle. While early-exit strategies based on monotonic focusing are not straightforward, the clear functional separation between agent-focused (Layers 0–2) and map-focused (Layer 3) processing could inform architecture-aware efficiency strategies, such as applying different sparsification policies per layer or using adaptive computation that allocates more resources to the final map-aggregation step.

5.5. Scene-Type Attention Adaptation

Section 4 and Figure 10 demonstrate that the model dynamically adapts its attention strategy across scene types. Table 4 provides the detailed per-category statistics.

Table 4. Attention distribution across scene types (200 scenes). Agent% and Map% denote the fraction of total attention directed to agent and map tokens, respectively. Top-5 Dist. is the mean distance of the five most-attended agents from the ego vehicle.

Scene Type	N	Agent%	Map%	Entropy (bits)	Top-5 Dist. (m)
Dense traffic	90	42.3	57.7	6.11	18.1
Sparse	14	18.4	81.6	5.33	18.3
Highway-like	52	31.6	68.4	5.71	21.4
Intersection-like	75	42.3	57.7	6.10	17.0
With pedestrians	105	39.4	60.6	5.99	17.4
With cyclists	37	39.2	60.8	5.90	17.2

Several patterns emerge from this comparison. First, attention entropy correlates strongly with scene complexity: dense traffic and intersection scenarios produce the highest entropy (6.11 and 6.10 bits), while sparse scenes yield the lowest (5.33 bits). This confirms that the model distributes attention more broadly when more traffic participants compete for relevance. The contrast in agent-directed attention between dense (42.3%) and sparse (18.4%) scenes is particularly striking—when fewer agents are present, the model compensates by attending more heavily to map structure, presumably to infer road context that agents would otherwise provide implicitly.

Second, the top-5 attended-agent distance reveals an adaptive planning horizon: highway-like scenes exhibit the largest mean distance (21.4 m), compared with 17.0 m for intersections. At highway speeds, agents farther ahead become relevant within the prediction window, and the model adjusts its spatial focus accordingly. Third, cyclist-containing scenes show the highest near-to-far attention ratio among all categories (1.70×), indicating that the model concentrates attention on nearby vulnerable road users rather than distributing it across distant context. This finding has direct implications for traffic safety policy: an interpretable model whose attention demonstrably prioritizes nearby cyclists provides a stronger basis for regulatory trust than one whose internal reasoning is opaque. More broadly, these scene-type adaptations demonstrate that the visualization framework reveals not only static architectural properties but also dynamic, context-sensitive behavior—evidence that attention-based interpretability can inform both model improvement and safety-critical deployment decisions.

5.6. Failure Diagnosis Through Attention: Identifying Safety-Critical Patterns

The analyses presented thus far characterize attention behavior in aggregate or across scene categories, but a safety-critical question remains: *does the model's attention differ systematically between successful and failed predictions?* Section 4.8 and Figure 8 present the quantitative evidence for a "tunnel vision" failure mode. Here we interpret these findings and discuss their implications. We stratified 1,115 prediction targets into success (Q1, minADE $\leq$ 0.71 m, $n = 279$) and failure (Q4, minADE $\geq$ 3.32 m, $n = 279$) groups. Table 5 provides the detailed attention and contextual statistics.

Table 5. Attention and contextual comparison between successful and failed predictions (quartile split, $n = 279$ per group). Metrics are computed from the scene encoder's final-layer attention. Agent attention % denotes the fraction of total attention directed to agent tokens. GT-nearest agent distance is the Euclidean distance from the ground-truth future trajectory to the closest neighboring agent.

Metric	Success (Q1)	Failure (Q4)
minADE (m)	0.57	7.19
Attention entropy (bits)	5.94	5.72
Agent attention (%)	48.8	43.2
Self-attention weight	0.035	0.049
Max single-token attention	0.039	0.058
GT-nearest agent distance (m)	5.5	33.9
Target speed (m/s)	0.2	7.2
Nearby agents (<15 m)	5.2	3.4
Cyclist targets in group (%)	0	4

Three findings emerge from this analysis, each with direct implications for autonomous driving safety.

Finding 1: The "tunnel vision" failure mode. Counter to the intuitive expectation that failures arise from diffuse, unfocused attention, the failure group exhibits *lower* entropy (5.72 vs. 5.94 bits) and *higher* self-attention (0.049 vs. 0.035). In failed predictions, the model concentrates disproportionate attention on its own token representation rather than surveying the surrounding scene. The maximum single-token weight is 49% higher in failures (0.058 vs. 0.039), confirming that the model over-commits to a narrow subset of tokens. We term this pattern *attention tunnel vision*: the model fails not because its attention is too diffuse, but because it retreats into self-referential processing and under-attends to contextual cues that would correct its trajectory estimate. This finding inverts the common assumption that broader attention is wasteful, and it implies that attention entropy could serve as a real-time diagnostic: abnormally low entropy during inference may signal an impending prediction failure.

Finding 2: Speed as the dominant risk factor. The most striking contextual difference between the two groups is target speed. Success cases average 0.2 m/s—nearly stationary agents whose future positions are trivially predictable—while failure cases average 7.2 m/s. At this speed, an agent traverses approximately 57.6 m over the 8-second prediction horizon, introducing a vast spatial envelope of plausible future positions. The model's attention pathology at 7.2 m/s raises concerns about behavior at highway speeds (25–35 m/s), where the spatial uncertainty grows by an additional factor of four to five. These high-speed regimes are precisely where prediction failures carry the greatest collision risk, since stopping distances grow quadratically with speed. Although our dataset predominantly contains urban driving, the trend strongly suggests that attention-based prediction models require explicit architectural or training interventions to maintain healthy attention patterns at elevated speeds.

Finding 3: Missing reference anchors. In successful predictions, the nearest neighboring agent to the ground-truth future trajectory is only 5.5 m away, providing the model with a proximal "reference anchor"—a nearby traffic participant whose current position and heading implicitly constrain the target's plausible future paths. In failure cases, this distance balloons to 33.9 m: the model must predict the target's trajectory in a spatial region devoid of other agents, eliminating the anchor-based heuristic. Concurrently, the mean number of nearby agents (within 15 m) drops from 5.2 to 3.4, further impoverishing the local context. This pattern suggests that the model partially relies on a "follow the leader" strategy—using neighboring agents' trajectories as soft constraints on the prediction—and degrades when this social cue is unavailable. The 4% prevalence of cyclists exclusively in the failure group further indicates that underrepresented agent types in the training distribution compound the difficulty of isolated, high-speed prediction. Figure 9 provides direct visual evidence: a cyclist target receives only 0.026 self-attention (versus 0.045 for a successful vehicle prediction, a 73% deficit), and cyclists collectively attract sevenfold less attention than vehicles despite their elevated vulnerability.

The 100% cyclist miss rate in the case study subset directly visualizes this attention bias translating into prediction failure for vulnerable road users.

Implications for safety certification and sustainable deployment. These findings transform our interpretability framework from a visualization tool into a *failure diagnosis instrument*. The tunnel vision pattern is directly relevant to production autonomous driving systems—including those deployed by major manufacturers—that rely on Transformer attention for scene understanding and prediction. If attention entropy drops below a calibrated threshold during inference, the system could flag the prediction as unreliable and trigger a fallback strategy (e.g., conservative braking, increased following distance, or handoff to a safety driver). Such an attention-based runtime monitor would complement traditional uncertainty estimation methods (e.g., ensemble disagreement or Monte Carlo dropout) with a mechanistically interpretable signal grounded in the model's actual reasoning process.

From a regulatory perspective, these results provide the type of failure-mode evidence that frameworks such as the EU AI Act [13] and NHTSA's AV testing protocols [1] increasingly demand. Rather than reporting only aggregate accuracy metrics, manufacturers could demonstrate that their models exhibit healthy attention distributions across speed regimes, agent densities, and road-user types—or disclose the conditions under which attention pathology emerges. We propose that *attention health profiles*, stratified by operating conditions, should become a standard component of safety certification for Transformer-based autonomous driving systems. Such profiles would quantify the speed threshold at which tunnel vision onset occurs, the minimum agent density required for reliable anchor-based prediction, and the attention deficit for underrepresented agent classes. As future work, we envision calibrating entropy thresholds on held-out data and validating the runtime monitor in closed-loop simulation, bridging the gap between post-hoc interpretability and real-time safety assurance for sustainable autonomous mobility [3,30].

5.7. Implications for Safety Certification

Building on the failure-mode analysis and attention health profiles proposed above, our framework contributes three complementary types of evidence for regulatory compliance—each addressing a dimension that aggregate accuracy metrics alone cannot capture:

1. **Spatial evidence**: BEV attention overlays demonstrate that the model "looks at" the correct scene elements before making predictions—or reveal when it does not, providing visual audit trails for safety reviewers.
2. **Causal evidence**: The counterfactual methodology described in Section 3.6 enables controlled experiments that can demonstrate context-aware reasoning rather than pattern memorization—a capability designed to complement observational analysis.
3. **Quantitative thresholds**: Attention-based safety metrics (e.g., entropy bounds for tunnel vision detection, agent attention share minima for different scene types) provide testable criteria that can be incorporated into certification test suites.

Together, these forms of evidence address the *how* and *why* of model behavior, complementing traditional metric-based evaluation (minADE, minFDE) that captures only the *how well*.

5.8. Implications for Sustainable Urban Mobility

The connection between model interpretability and sustainable transportation operates through a causal chain: interpretability enables trust, trust enables adoption, and adoption enables the environmental and safety benefits that autonomous vehicles promise [3,7]. Our work contributes to this chain at two levels:

- **Direct sustainability**: Our distance mask ablation (Figure 11) reveals that naive spatial pruning strategies degrade performance by 4.7%, but the layer specialization patterns suggest that more sophisticated, architecture-aware efficiency strategies—such as early-exit mechanisms or adaptive

computation—may be feasible without sacrificing safety-critical context. One promising direction is *entropy-guided dynamic token pruning*: monitoring per-layer attention entropy at runtime and selectively pruning tokens whose attention weight falls below an entropy-derived threshold, thereby reducing computational cost while preserving contextually relevant information. This approach would leverage the diagnostic power of attention visualization to enable real-time efficiency optimization aligned with sustainability goals.

- **Indirect sustainability**: By making trajectory prediction models transparent and auditable, we lower barriers to regulatory approval and public acceptance, accelerating the transition to shared autonomous mobility. Studies project that widespread AV adoption could reduce vehicle ownership by 30–40%, traffic fatalities by 90%, and fuel consumption by 40% [3,4].

5.9. Limitations and Generalizability

We acknowledge several limitations that bound the scope of our quantitative findings and discuss which aspects of this work generalize beyond the specific model studied.

Model scale gap and the probe model paradigm. Our MTR-Lite variant comprises approximately 8 million parameters trained on approximately 17,800 scenes, achieving a minADE of 2.314 m on the full validation set (13,388 scenes, 99,370 agent predictions). Production trajectory prediction systems typically exceed 100 million parameters, train on millions of scenes, and achieve minADE values below 0.8 m. This order-of-magnitude gap in model capacity and data scale means that the specific quantitative findings reported here—such as absolute entropy values (5.3–6.1 bits), the 29% far-range attention share, or the 49% elevation in maximum single-token weight between success and failure cases—may not transfer directly to larger models. Richer representations learned at scale could mitigate or alter these patterns.

However, MTR-Lite's value lies not in competing with state-of-the-art prediction accuracy, but in serving as an *interpretability probe*—a deliberately lightweight model that enables systematic attention analysis with rapid iteration cycles and manageable computational overhead. The visualization framework itself is model-agnostic: applying the same attention extraction and spatial bookkeeping methodology to production-scale models such as MTR++ [9] or Wayformer [10] is straightforward, as all Transformer-based predictors expose attention weights through the same API. Extending this work to larger models is planned as future work and would reveal whether the tunnel vision failure mode, layer specialization patterns, and scene-type adaptations we document here persist at scale or are replaced by qualitatively different attention strategies enabled by richer capacity.

Moreover, recent evidence from natural language processing suggests that structural attention pathologies persist even at large scale: Xiao et al. [48] demonstrate that large language models exhibit "attention sinks," allocating disproportionate attention to initial tokens regardless of semantic relevance, while Zhai et al. [49] document attention entropy collapse in deep Vision Transformers. These findings suggest that the tunnel vision failure mode identified in Section 5.6 may reflect fundamental architectural properties rather than scale-specific limitations.

Data scale and vulnerable road user prediction. Training on 20% of the Waymo Open Motion Dataset (approximately 17,800 scenes rather than the full 85,000+ training scenes) has particularly pronounced implications for rare agent types such as cyclists. Cyclists already constitute a small minority in the full dataset—outnumbered by vehicles approximately 8:1 in typical urban scenes—so training on 20% of the data reduces the effective cyclist training examples by roughly fivefold. The 88.1% miss rate for cyclist predictions reported in Table 3—compared to 54.0% for vehicles, a 63% relative increase—likely reflects both the inherent challenge of predicting cyclist behavior (which combines vehicle-like speeds with pedestrian-like maneuverability) *and* severe data scarcity. Training on the full dataset would likely improve cyclist prediction accuracy by providing the model with sufficient examples to learn cyclist-specific motion patterns. However, the tunnel vision attention pattern documented in Section 5.6—lower entropy and elevated self-attention in failed predictions—may persist even with more data, as it appears to be an architecture-level failure mode

rather than a data-level deficiency. Future work should investigate whether data augmentation strategies specific to vulnerable road users, or loss weighting schemes that prioritize rare agent types, can mitigate this safety-critical gap.

Architecture differences. MTR-Lite employs vanilla global self-attention without spatial inductive bias. State-of-the-art production models incorporate local attention windows, relative position encoding, factored attention (as in Wayformer [10]), or scene-graph structures that explicitly encode spatial relationships. These architectural choices may mitigate issues such as the "tunnel vision" pattern we identified, since mechanisms like distance-gated attention inherently limit self-referential processing. Our counterfactual experiments are likewise constrained to element removal and modification within real scenes, rather than generating fully synthetic scenarios, and the causal claims they support apply to the specific model under test rather than constituting formal guarantees in the Pearl [36] sense.

What generalizes. Despite these model-specific caveats, several contributions of this work are designed to generalize broadly:

- *Methodology.* The visualization framework—spatial token bookkeeping, Gaussian splatting, polyline painting, and layer-wise entropy decomposition—applies to any Transformer-based driving model that produces attention weights over spatially grounded tokens. The tools are architecture-agnostic.
- *Architecture-level findings.* The observation that global self-attention lacks an inherent distance prior is a structural property of the attention mechanism itself, not an artifact of model scale. Any architecture using unmodified dot-product attention over spatially embedded tokens will face the same challenge of learning distance relevance from data alone.
- *Physics-driven findings.* The correlation between target speed and prediction difficulty, and the role of nearby agents as reference anchors, are driven by traffic dynamics rather than model specifics. These relationships should manifest in any prediction model operating on real-world driving data.
- *Diagnostic pattern.* The *observe–hypothesize–test* cycle we demonstrated—where spatial attention analysis generated a hypothesis about far-range attention waste, and the distance-mask ablation falsified it—is a reusable diagnostic methodology applicable to production-scale systems.

Our primary contribution is therefore not the specific attention statistics, but the interpretability framework and diagnostic methodology. We demonstrated on MTR-Lite how this framework reveals tunnel vision, speed-dependent risk, layer specialization, and reference anchor effects. Applying the same tools to production-scale models with richer architectures is a natural and important direction for future work, and we anticipate that such analyses will uncover both analogous patterns and novel phenomena enabled by greater model capacity.

6. Conclusions

This paper presented a spatial attention visualization framework for Transformer-based trajectory prediction that moves beyond abstract attention matrices to provide spatially grounded, interpretable insights into model behavior. By combining a novel spatial token bookkeeping mechanism with Gaussian splatting and polyline painting techniques, we demonstrated how attention weights can be projected as continuous heatmaps onto bird's-eye-view traffic scenes, revealing *where* the model looks, *how* its reasoning evolves across layers, and *which* road structures guide its predictions.

Our analysis across 100–200 Waymo Open scenes uncovered four key findings with implications for autonomous driving safety and interpretability. First, layer-wise entropy analysis revealed *collaborative layer specialization* rather than monotonic attention focusing: Layers 0–2 progressively narrow from 5.64 to 5.36 bits while agent attention increases from 49.7% to 62.4%, but Layer 3 *reverses* this trend—entropy rises to 5.92 bits (the highest of all layers) and map attention jumps to 63.6%. This indicates a two-phase reasoning strategy in which early layers identify relevant agents and the final

layer aggregates broader spatial context. Second, we identified a *tunnel vision* failure mode by analyzing 1,115 prediction targets: failed predictions (ADE ≥ 3.32 m) exhibit *lower* attention entropy (5.72 vs. 5.94 bits) and *higher* self-attention (0.049 vs. 0.035) than successful ones, with target speed as the dominant risk factor (7.2 m/s for failures vs. 0.2 m/s for successes). This suggests that attention entropy could serve as a real-time failure diagnostic for safety monitoring. Third, distance mask ablation experiments across 750 targets demonstrated that restricting far-range attention *hurts* performance at every masking level (baseline ADE 2.872 m worsens by at least 4.7%), confirming that distant tokens provide essential traffic flow context, road structure inference, and indirect interaction dynamics. Fourth, scene-type analysis across 200 scenes revealed that the model dynamically adapts its attention strategy: agent attention reaches 42.3% in dense traffic but drops to 18.4% in sparse scenes, while highway top-5 attention distance extends to 21.4 m compared to 17.0 m at intersections.

These findings have direct implications for sustainable and safe autonomous driving. The tunnel vision failure mode reveals that overconfident, narrowly focused attention is a measurable precursor to prediction failures, opening a pathway toward attention-entropy-based safety monitoring that could flag dangerous predictions before they propagate to planning. The distance mask ablation demonstrates that principled *observe→hypothesize→test* diagnostic cycles—enabled by our visualization framework—can reveal non-obvious model dependencies and prevent well-intentioned but harmful architectural simplifications. The scene-type adaptation finding confirms that Transformer attention is not a static computation but a context-sensitive reasoning process, strengthening the case for interpretability as a route to trust and regulatory certification under frameworks such as the EU AI Act.

Future work will pursue four directions. First, we will extend our analysis to larger, state-of-the-art models (e.g., MTR++, SMART) to investigate whether the layer specialization patterns and tunnel vision failure mode generalize across architectures. Second, we will develop attention regularization techniques that enforce minimum attention thresholds for vulnerable road users during training, directly addressing safety blind spots in current models. Third, we will execute the counterfactual attention experiments—for which we have designed and implemented a controlled scene editing pipeline—to establish causal (rather than merely correlational) links between attention patterns and prediction outcomes. Fourth, we will investigate entropy-guided dynamic token pruning to reduce computational overhead while preserving prediction accuracy, integrating our visualization framework into closed-loop simulation environments to evaluate whether attention-aware efficiency optimization and safety monitoring improve real-time decision-making in dynamic driving scenarios.

Importantly, the visualization and diagnostic framework presented here—spatial token bookkeeping, entropy decomposition, and the *observe→hypothesize→test* analytical cycle—is **architecture-agnostic**. While we demonstrated it on MTR-Lite as a lightweight interpretability probe, the same tools apply unchanged to any Transformer that produces attention weights over spatially grounded tokens, including production-scale systems such as Wayformer, MTR++, and QCNet. By bridging the gap between model performance and model understanding, this work contributes to the broader goal of building autonomous vehicles that are not only accurate but also transparent, safe, and trustworthy—essential prerequisites for realizing the sustainability benefits of autonomous urban mobility.

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

Funding: This research received no external funding.

Data Availability Statement: The trajectory prediction models, attention extraction framework, and visualization code developed in this study are available from the corresponding author upon reasonable request. The Waymo Open Motion Dataset used for training and evaluation is publicly available at https://waymo.com/open/data/motion/ under the Waymo Dataset License Agreement.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors acknowledge the use of the Waymo Open Motion Dataset for the experiments presented in this work. Computational resources were provided by Concordia University.

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

Abbreviations

The following abbreviations are used in this manuscript:

ADE	Average Displacement Error
BEV	Bird's-Eye View
BFS	Breadth-First Search
FDE	Final Displacement Error
MR	Miss Rate
MTR	Motion Transformer
NMS	Non-Maximum Suppression
VRU	Vulnerable Road User
WOMD	Waymo Open Motion Dataset
XAI	Explainable Artificial Intelligence

References

1. National Highway Traffic Safety Administration. A Framework for Automated Driving System Testable Cases and Scenarios. Technical report, U.S. Department of Transportation, 2022. DOT HS 813 066.
2. United Nations. Transforming Our World: The 2030 Agenda for Sustainable Development, 2015. A/RES/70/1.
3. Fagnant, D.J.; Kockelman, K. Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations. *Transportation Research Part A: Policy and Practice* **2015**, *77*, 167–181.
4. Greenblatt, J.B.; Shaheen, S. Automated Vehicles, On-Demand Mobility, and Environmental Impacts. *Current Sustainable/Renewable Energy Reports* **2015**, *2*, 74–81.
5. Wadud, Z.; MacKenzie, D.; Leiby, P. Help or Hindrance? The Travel, Energy and Carbon Impacts of Highly Automated Vehicles. *Transportation Research Part A: Policy and Practice* **2016**, *86*, 1–18.
6. Nordhoff, S.; de Winter, J.; Kyriakidis, M.; van Arem, B.; Happee, R. Conceptual Model to Explain, Predict, and Improve User Acceptance of Driverless Podlike Vehicles. *Transportation Research Record* **2018**, *2672*, 60–71.
7. Milakis, D.; Van Arem, B.; Van Wee, B. Policy and Society Related Implications of Automated Driving: A Review of Literature and Directions for Future Research. *Journal of Intelligent Transportation Systems* **2017**, *21*, 324–348.
8. Shi, S.; Jiang, L.; Dai, D.; Schiele, B. Motion Transformer with Global Intention Localization and Local Movement Refinement. Advances in Neural Information Processing Systems (NeurIPS), 2022, Vol. 35, pp. 6531–6543.
9. Shi, S.; Jiang, L.; Dai, D.; Schiele, B. MTR++: Multi-Agent Motion Prediction with Symmetric Scene Modeling and Pair-Wise Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 3039–3051.
10. Nayakanti, N.; Al-Rfou, R.; Zhou, A.; Goel, K.; Refaat, K.S.; Sapp, B. Wayformer: Motion Forecasting via Simple & Efficient Attention Networks. IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 2980–2987.
11. Huang, Z.; Liu, H.; Lv, C. GameFormer: Game-Theoretic Modeling and Learning of Transformer-Based Interactive Prediction and Planning for Autonomous Driving. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3903–3913.
12. Ngiam, J.; Vasudevan, V.; Caine, B.; Zhang, Z.; Chiang, H.T.L.; Ling, J.; Roelofs, R.; Bewley, A.; Liu, C.; Vaswani, A.; others. Scene Transformer: A Unified Architecture for Predicting Future Trajectories of Multiple Agents. International Conference on Learning Representations (ICLR), 2022.

13. European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Official Journal of the European Union, 2024.
14. Koopman, P.; Wagner, M. Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intelligent Transportation Systems Magazine* **2017**, *9*, 90–96.
15. Zablocki, E.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges. *International Journal of Computer Vision* **2022**, *130*, 2425–2452.
16. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
17. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, pp. 4768–4777.
18. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
19. Vig, J. A Multiscale Visualization of Attention in the Transformer Model. *arXiv preprint arXiv:1906.05714* **2019**.
20. Abnar, S.; Zuidema, W. Quantifying Attention Flow in Transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4190–4197.
21. Chefer, H.; Gur, S.; Wolf, L. Transformer Interpretability Beyond Attention Visualization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791.
22. Jain, S.; Wallace, B.C. Attention is not Explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 3543–3556.
23. Wiegreffe, S.; Pinter, Y. Attention is not not Explanation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 11–20.
24. Da Silva Martins, S.; Aldea, E.; Le Hégarat-Mascle, S. VISTA: A Vision and Intent-Aware Social Attention Framework for Multi-Agent Trajectory Prediction. *arXiv preprint arXiv:2511.10203* **2025**.
25. Yadav, H.; Schaefer, M.; Zhao, K.; Meisen, T. LMFormer: Lane based Motion Prediction Transformer. *arXiv preprint arXiv:2504.10275* **2025**.
26. Zhou, X.; Alecsandru, C. Local Lane Graph Conditioning as a General Inductive Bias for Trajectory Prediction: A Multi-Architecture Study on the Waymo Open Motion Dataset. *Sustainability* **2026**.
27. Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C.R.; Zhou, Y.; others. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9710–9719.
28. Zhou, Z.; Wang, J.; Li, Y.H.; Huang, Y.K. Query-Centric Trajectory Prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17863–17873.
29. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *IEEE Access* **2024**, *12*, 1–30.
30. Taiebat, M.; Brown, A.L.; Safford, H.R.; Qu, S.; Xu, M. A Review on Energy, Environmental, and Sustainability Implications of Connected and Automated Vehicles. *Environmental Science & Technology* **2018**, *52*, 11449–11465.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017, Vol. 30, pp. 5998–6008.
32. Gao, J.; Sun, C.; Zhao, H.; Shen, Y.; Anguelov, D.; Li, C.; Schmid, C. VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11525–11533.
33. Zeng, Z.; Mao, J.; Dai, B.; Anguelov, D. Heterogeneous Polyline Transformer with Relative Pose Encoding for Map-Aware Motion Prediction. Advances in Neural Information Processing Systems (NeurIPS), 2023, Vol. 36.
34. Wu, W.; Feng, X.; Gao, Z.; Kan, Y. SMART: Scalable Multi-Agent Real-Time Motion Generation via Next-Token Prediction. Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.

35. Gao, Z.; Liu, D.; Yu, T.; Hu, H.; Gao, F.; Zhao, R. ISE-GT: Interaction Strength-Enhanced Graph Transformer for Explainable Multi-Agent Trajectory Prediction. *Transportation Research Part C: Emerging Technologies* **2025**.
36. Pearl, J. Causal Inference in Statistics: An Overview. *Statistics Surveys* **2009**, *3*, 96–146.
37. Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; Lee, S. Counterfactual Visual Explanations. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 2376–2384.
38. Tan, S.; Wong, K.; Wang, S.; Manivasagam, S.; Ren, M.; Urtasun, R. SceneGen: Learning to Generate Realistic Traffic Scenes. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 892–901.
39. Suo, S.; Regalado, S.; Casas, S.; Urtasun, R. TrafficSim: Learning to Simulate Realistic Multi-Agent Behaviors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10400–10409.
40. Zhong, Z.; Rempe, D.; Xu, D.; Chen, Y.; Veer, S.; Che, T.; Ray, B.; Pavone, M. Guided Conditional Diffusion for Controllable Traffic Simulation. IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 3560–3566.
41. Ding, W.; Xu, C.; Arief, M.; Lin, H.; Li, B.; Zhao, D. A Survey on Safety-Critical Driving Scenario Generation—A Methodological Perspective. *IEEE Transactions on Intelligent Transportation Systems* **2023**, *24*, 6971–6988.
42. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; others. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* **2020**, *58*, 82–115.
43. Litman, T. Autonomous Vehicle Implementation Predictions: Implications for Transport Planning. Technical report, Victoria Transport Policy Institute, 2023.
44. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 652–660.
45. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv preprint arXiv:1607.06450* **2016**.
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR), 2019.
47. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; Wu, H. Mixed Precision Training. International Conference on Learning Representations (ICLR), 2018.
48. Xiao, G.; Tian, Y.; Chen, B.; Han, S.; Lewis, M. Efficient Streaming Language Models with Attention Sinks. *arXiv preprint arXiv:2309.17453* **2023**.
49. Zhai, S.; Likhomanenko, T.; Littwin, E.; Busbridge, D.; Ramapuram, J.; Zhang, Y.; Gu, J.; Susskind, J.M. Stabilizing Transformer Training by Preventing Attention Entropy Collapse. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023, pp. 40770–40803.