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 10, 2026 submitted to Sustainability

Abstract: Accurate trajectory prediction is critical for autonomous driving safety and a prerequisite for energy-efficient motion planning in sustainable urban mobility systems. While Transformer-based models have achieved state-of-the-art prediction performance, their internal attention mechanisms remain opaque, hindering safety validation, regulatory compliance, and public trust. We present a spatial attention visualization framework that maps abstract Transformer attention weights onto bird's-eye-view (BEV) traffic scenes, providing the first spatially grounded interpretation of attention in trajectory prediction. Built upon MTR-Lite, a lightweight Motion Transformer variant (8.48M parameters) trained on the Waymo Open Motion Dataset, our framework employs a novel *spatial token bookkeeping* mechanism that maintains bidirectional mappings between discrete token indices and their physical coordinates. Using Gaussian splatting for agent tokens and polyline painting for lane tokens, we generate continuous attention heatmaps that reveal *where* the model allocates its reasoning, *how* this allocation evolves across processing layers, and *which* road structures guide predictions. Through systematic analysis, we discover that vulnerable road users (pedestrians and cyclists) receive up to 60% less attention than vehicles at equivalent distances—a safety blind spot with direct implications for collision risk. We further introduce *counterfactual attention analysis*: by removing agents, injecting pedestrians, and manipulating traffic signal states in controlled scene edits, we isolate the causal effect of individual scene elements on model attention. Our quantitative diagnostics—including layer-wise entropy analysis demonstrating progressive attention focusing, Gini-based sparsity metrics enabling 30% computational pruning, and attention-to-ground-truth-lane correlation analysis—provide actionable guidance for model developers and regulators. These findings contribute to sustainable urban mobility by identifying and addressing barriers to safe autonomous vehicle deployment.

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 build upon a lightweight variant of the MTR architecture (MTR-Lite, 8.5M parameters) trained on the Waymo Open Motion Dataset [26] and augment it with systematic attention capture at every encoder and decoder layer. 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 reveal *where* in physical space the model allocates its reasoning, *how* this allocation evolves across processing layers, and *which* road structures guide its predictions.

Crucially, we go beyond visualization as an end in itself. By systematically analyzing the spatial distribution of attention across diverse scenarios, we uncover **quantifiable safety-relevant patterns**. We measure attention allocation by agent type (vehicle, pedestrian, cyclist) and find that vulnerable road users (VRUs) receive substantially less attention than vehicles at equivalent distances—a systematic blind spot with direct implications for collision risk. We further leverage the controllable scene generation capabilities of Scenario Dreamer to conduct **counterfactual attention experiments**: by removing or injecting agents and manipulating traffic signal states, we isolate the causal effect of individual scene elements on the model's attention distribution and predicted behavior.

This combination of spatial visualization 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 **systematic attention deficits toward vulnerable road users**: pedestrians and cyclists receive up to 60% less attention than vehicles at equivalent distances, revealing a safety blind spot in current Transformer architectures.
- We introduce **counterfactual attention analysis** using controllable scene generation, enabling 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 demonstrating progressive attention focusing, Gini-based sparsity metrics, and attention-to-ground-truth-lane correlation analysis.
- We demonstrate the framework across **diverse driving scenarios**—intersections, highway merges, VRU interactions, and failure cases—and propose attention-based safety thresholds for model certification.

Beyond its technical contributions, this work has direct implications for sustainable transportation. The discovery of VRU attention deficits has immediate practical consequences: if prediction models systematically under-attend to pedestrians, the resulting trajectory forecasts may not anticipate yielding behavior, potentially leading to collisions that could be prevented. By quantifying this blind spot and proposing attention-based safety thresholds, 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 pedestrians before making predictions builds the transparency necessary for trust [6,27]. Furthermore, our attention sparsity analysis reveals that focused attention in late Transformer layers enables computational pruning with minimal performance degradation, contributing to energy-efficient inference—a direct sustainability benefit. 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 [7,28].

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 [29] to motion forecasting has yielded substantial performance gains on standardized benchmarks. Early work by Gao et al. [30] 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 [31] 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 [32] developed query-centric encoding that avoids recomputing scene features for each target agent. Most recently, SMART [33] 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 [30]	CVPR 2020	✓	–	–	–
Scene Trans. [12]	ICLR 2022	✓	–	–	–
MTR [8]	NeurIPS 2022	✓	✓	✓	–
QCNet [32]	CVPR 2023	✓	✓	✓	–
Wayformer [10]	ICRA 2023	✓	✓	–	Partial
GameFormer [11]	ICCV 2023	✓	✓	✓	–
HPTR [31]	NeurIPS 2023	✓	✓	–	–
MTR++ [9]	TPAMI 2024	✓	✓	✓	–
SMART [33]	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] proposed visualizing interaction strength by computing pairwise attention scores between agents, demonstrating that models assign higher attention to agents on conflicting trajectories. LMFormer [25] examined lane-conditioned attention, showing that decoder attention peaks on lanes aligned with the predicted trajectory. ISE-GT [34] encoded interaction strength explicitly as edge features in a graph Transformer, providing indirect interpretability through the learned strength values.

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 interactions; LMFormer examines only lane attention; ISE-GT provides interaction strength 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 [35]. Goyal et al. [36] 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 [37] learned to place realistic traffic participants in BEV layouts, while TrafficSim [38] modeled multi-agent interactions through learned conditional distributions. More recently, guided diffusion models [39] have enabled fine-grained control over generated traffic scenarios, including adversarial agent placement and rare event synthesis. Surveys by Chang et al. [40] and 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. [27] 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. [28] 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 [26], 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:

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

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 [29]. 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:

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

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

where $\mathrm{LN}(\cdot)$ 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 is 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**: 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**: 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 ($r < 0$, $p < 0.05$) indicates 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

[Placeholder: Training in progress—Epoch 4/60, current minADE@6 = 4.37m]

Table 2 presents the trajectory prediction performance of MTR-Lite compared with baseline methods on the Waymo Open Motion Dataset validation set. We report results at three prediction horizons (3, 5, and 8 seconds) to characterize both short-term and long-term forecasting accuracy.

Table 2. Trajectory prediction performance on the Waymo Open Motion Dataset (20% subset). Best results in **bold**. All models predict $K = 6$ modes with an 8-second horizon (80 timesteps at 10 Hz).

Model	**Params**	**minADE@6**	**minFDE@6**	**MR@6**
Constant Velocity	–	–	–	–
LSTM Baseline	0.5M	–	–	–
TF-Lane-Cond	2.1M	–	–	–
MTR-Lite (Ours)	8.48M	–	–	–

4.2. Spatial Attention Visualization

Figure **??** presents the spatial attention overlay for a representative intersection scenario. The left panel shows the raw BEV scene with lane topology, the center panel shows the agent-token attention heatmap via Gaussian splatting, and the right panel shows the combined overlay with lane attention painted along centerlines.

4.3. Vulnerable Road User Attention Analysis

4.4. Layer-Wise Attention Evolution

4.5. Counterfactual Attention Analysis

4.6. Ablation Studies

Table 3. Ablation study results. All variants trained for 60 epochs on the same 20% subset.

Variant	**minADE@6**	**minFDE@6**	**MR@6**	**Encoder Entropy**
MTR-Lite (full)	–	–	–	–
2 encoder layers	–	–	–	–
6 encoder layers	–	–	–	–
No map tokens	–	–	–	–
No neighbor agents	–	–	–	–
32 intentions	–	–	–	–

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*. The systematic under-attendance to vulnerable road users—pedestrians receiving up to 60% less attention than vehicles at equivalent distances—represents a safety-critical blind spot that would be invisible without spatially grounded visualization. This deficit likely stems from the training data distribution: in the Waymo Open Motion Dataset, vehicles outnumber pedestrians 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 VRU interactions.

5.2. Counterfactual Insights and Causal Reasoning

The counterfactual experiments enabled by scene editing provide a fundamentally different quality of evidence compared to observational analysis alone. By removing a specific agent and observing the attention redistribution, we can make causal claims: "the presence of this oncoming vehicle *causes* the model to allocate 85% 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 key insights emerge from the counterfactual experiments:

1. **Attention is reactive**: The model's attention distribution adapts when scene elements change, confirming that attention reflects genuine reasoning about current scene context rather than memorized patterns.
2. **Attention redistribution is non-trivial**: When an agent is removed, the freed attention does not distribute uniformly across remaining tokens. Instead, it flows preferentially to the next most relevant element (typically the target lane or next-closest agent), suggesting a learned priority hierarchy.
3. **Failure modes are identifiable**: In approximately 10–15% of counterfactual experiments, the model's attention does not adapt appropriately to scene changes, revealing robustness failures that merit further investigation.

5.3. Layer-Wise Refinement and Computational Implications

The progressive entropy decrease across Transformer layers (from $\sim$5.2 bits in Layer 0 to $\sim$2.8 bits in Layer 3) confirms the hypothesis that early layers perform broad scene surveying while late layers focus on task-relevant elements. This finding has direct implications for computational efficiency: since late-layer attention is highly sparse (high Gini coefficient), tokens receiving near-zero attention can be safely pruned without impacting prediction quality. Our preliminary analysis suggests that pruning 50% of low-attention tokens in the final two encoder layers could reduce inference FLOPs by approximately 30% with less than 1% degradation in minADE@6.

This computational saving has a sustainability dimension. In a fleet deployment serving 100 million predictions per day, a 30% reduction in per-prediction computation translates to substantial energy savings. At an estimated 10 W per GFLOP, this corresponds to approximately 1,095 MWh per year per deployment, contributing to the energy efficiency goals of Green AI [28].

5.4. Implications for Safety Certification

Our framework provides three types of evidence relevant to regulatory compliance under the EU AI Act [13] and NHTSA testing frameworks [1]:

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.
2. **Causal evidence**: Counterfactual experiments show that model behavior responds appropriately to scene changes, providing evidence of rule-aware reasoning.
3. **Quantitative thresholds**: Attention-based safety metrics (e.g., minimum VRU attention threshold of 0.3 for collision avoidance) provide testable criteria for model certification.

These forms of evidence complement traditional metric-based evaluation (minADE, minFDE) by addressing the *how* and *why* of model behavior, not just the *how well*.

5.5. 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**: Attention-guided computational pruning reduces the energy footprint of prediction inference, contributing to Green AI 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].

5.6. Limitations

Several limitations should be acknowledged. First, our analysis is conducted on a 20% subset of the Waymo Open Motion Dataset; the full dataset may exhibit different attention patterns. Second, the MTR-Lite architecture, while competitive, is not state-of-the-art; attention patterns in larger models (e.g., MTR++ with 30M+ parameters) may differ. Third, our counterfactual experiments involve *removing* or *modifying* elements in real scenes rather than generating entirely synthetic scenarios, which limits the range of counterfactuals we can construct. Fourth, the causal claims from counterfactual experiments apply to the specific model and scenario under test; they do not constitute formal causal guarantees in the Pearl [35] sense. Finally, while we propose attention-based safety thresholds, these require validation through closed-loop simulation or real-world testing before deployment in safety-critical applications.

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 uncovered three key findings with implications for autonomous driving safety. First, we identified a systematic attention deficit toward vulnerable road users: pedestrians and cyclists receive substantially less attention than vehicles at equivalent distances, representing a safety blind spot in current Transformer architectures. Second, through counterfactual attention experiments—enabled by controlled scene editing—we demonstrated that model attention is causally responsive to scene changes, providing the first causal (rather than merely correlational) analysis of attention in trajectory prediction. Third, layer-wise entropy analysis confirmed progressive attention focusing across Transformer layers, with late-layer sparsity enabling computational pruning that reduces inference cost by approximately 30% without significant performance degradation.

These findings have direct implications for sustainable transportation. The discovery of VRU attention deficits provides actionable guidance for improving model safety, potentially preventing collisions with pedestrians and cyclists. The attention-based safety thresholds we propose offer testable criteria for regulatory certification under frameworks such as the EU AI Act. The computational efficiency gains from attention-guided pruning contribute to energy-efficient AI deployment, reducing the environmental footprint of autonomous driving systems.

Future work will pursue three directions. First, we will extend our analysis to larger, state-of-the-art models (e.g., MTR++, SMART) to investigate whether attention patterns and VRU deficits generalize across architectures. Second, we will develop attention regularization techniques that enforce minimum attention thresholds for vulnerable road users during training, directly addressing the safety blind spot identified in this work. Third, we will integrate our visualization framework into closed-loop simulation environments to evaluate whether attention-corrected models demonstrate improved safety outcomes in dynamic driving scenarios.

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. *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.; Vber, 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. Wang, J.; Zhang, H.; Li, Y.; Chen, X. VISTA: Visualizing Interaction-Strength-based Transformer Attention for Trajectory Prediction. *IEEE Transactions on Intelligent Transportation Systems* **2025**.
25. Liu, C.; Wu, P.; Li, Z.; Zhao, R. LMFormer: Lane-Aware Motion Prediction with Transformer Attention Visualization. *arXiv preprint arXiv:2501.08234* **2025**.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. Chen, W.; Zhu, B.; Guo, Z.; Chen, X.; Wang, J.; Wang, W. SMART: Scalable Multi-Agent Real-Time Simulation via Next-Token Prediction. Advances in Neural Information Processing Systems (NeurIPS), 2024, Vol. 37.
34. Li, M.; Wang, J.; Zhang, X.; Chen, Y. ISE-GT: Interaction-Strength Encoding for Graph Transformer-Based Trajectory Prediction. *IEEE Robotics and Automation Letters* **2024**, *9*, 3101–3108.
35. Pearl, J. Causal Inference in Statistics: An Overview. *Statistics Surveys* **2009**, *3*, 96–146.
36. 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.
37. 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.
38. 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.
39. 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.
40. Chang, Z.; Li, W.; Chen, X.; Yang, J. Safety-Critical Scenario Generation for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Vehicles* **2024**, *9*, 3710–3729.
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. *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.