

Article

A Unified Explainable Autonomous Driving Framework via Cross-Attention Scene Selection and Semantic–Object Fusion

Habib Dhahri , Fahad Alotaibi , Awais Mahmood  and Mousa Jari 

College of Applied Computer Science, King Saud University, Riyadh 11545, Saudi Arabia;
faalotaibi@ksu.edu.sa (F.A.); mawais@ksu.edu.sa (A.M.); maljari@ksu.edu.sa (M.J.)

* Correspondence: hdhahril@ksu.edu.sa

Abstract

Intelligent autonomous driving systems must not only predict the appropriate driving manoeuvre but also provide human-interpretable evidence that justifies the decision. However, existing methods typically address these objectives separately, leading to three practical limitations: multi-stage perception-to-language pipelines can propagate upstream perception errors into downstream explanations; post hoc saliency methods often produce pixel-level highlights that are difficult to interpret semantically; and decoupled decision and explanation modules cannot guarantee that the explanation reflects the same scene evidence used for behaviour prediction. In this paper, we propose a unified framework that jointly performs vehicle behaviour prediction and human-centric interpretation from a shared visual backbone. Specifically, a hierarchical Swin Transformer encodes the driving scene into a sequence of spatial tokens, which are processed by two complementary branches. The first branch, termed the Object Selection Module (OSM), learns a compact scene-level semantic representation through query-guided cross-attention, while the second branch extracts a small set of class-agnostic object-centric tokens without requiring bounding-box or segmentation supervision. These two representations are subsequently integrated by a Semantic–Object Fusion (SOF) module based on scaled dot-product attention, residual connections, and a feed-forward network. The behaviour prediction head operates on the fused representation, whereas the interpretation head leverages the semantic representation through a skip connection to preserve decision-relevant context. For surround-view perception, learnable per-camera embeddings are introduced to maintain viewpoint identity with negligible additional parameter cost. Furthermore, a compact language model fine-tuned via Low-Rank Adaptation (LoRA) generates fluent, label-conditioned natural-language justifications. Extensive experiments on two public benchmarks, BDD-OIA and nu-AD, demonstrate that the proposed framework consistently delivers superior performance and provides effective, human-readable interpretations of driving decisions.



Academic Editors: Xianjian Jin and Xin Xia

Received: 26 April 2026

Revised: 30 May 2026

Accepted: 4 June 2026

Published: 10 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: autonomous driving; explainable AI; vehicle-behaviour decision; cross-attention; semantic–object fusion; multi-camera perception; LoRA; BDD-OIA; nu-AD

1. Introduction

Human factors such as inattention and fatigue remain the dominant contributors to road traffic incidents, and autonomous driving systems are being developed precisely to mitigate these risks. A deployable autonomous driving system must, however, do more than select a safe manoeuvre: it must also communicate, in terms that a non-expert can evaluate, the scene evidence on which that selection rests. Predicting the action alone

is insufficient, because without a traceable justification, engineers cannot diagnose the conditions under which the system fails, regulators cannot assess whether it fails for acceptable reasons, and users have no basis on which to extend calibrated trust [1–3]. The problem we address is therefore a joint one: learning a single network that produces a vehicle-behaviour decision and a grounded natural-language interpretation from the same visual encoder and under the same objective.

Existing work has largely treated the two sub-problems in isolation. Post hoc saliency methods, including gradient-based visualisations and attention rollouts [4–6], can mark regions that a classifier finds influential, but they deliver pixel-level heatmaps rather than semantically grounded justifications, and user studies have shown that such outputs are difficult for non-experts to parse. A complementary line of work couples a pretrained detector or segmentation network with a downstream language component, so that decisions are conditioned on the symbolic perception output and then verbalised [7–9]. These pipelines recover interpretability but inherit a multi-stage training procedure and propagate perception errors downstream. More recently, the BDD-OIA benchmark [10] formalised explainable driving as joint classification over actions and a fixed set of reason labels, and the nu-AD dataset introduced in NLE-DM [11] extended this formulation to surround-view perception. VB-CASeg [12] further showed that a foundation-model backbone together with a fusion module can improve interpretability prediction under this joint formulation. From a learning-paradigm perspective, deep-learning methods for perception and decision-making can be broadly categorized into supervised, weakly supervised, and unsupervised or self-supervised approaches. Supervised methods use explicit task labels and remain dominant in safety-critical prediction settings because their outputs can be directly evaluated against annotated behaviours or reasons. Beyond road-traffic scenarios, supervised deep models have been used for multimodal single-image super-resolution with reversible guidance and cyclical knowledge distillation [13], whereas unsupervised deep transfer learning has been explored for rotating-machinery fault diagnosis through spatial-channel graph interaction [14]. Unsupervised and self-supervised methods are valuable when labels are scarce and have been successfully explored in domains such as image restoration, transfer learning, and fault diagnosis. In explainable autonomous driving, however, the need to evaluate both the predicted manoeuvre and the stated reason makes supervised or weakly supervised formulations particularly important. CrossAD follows this direction by jointly learning behaviour and interpretation outputs from annotated driving benchmarks, while avoiding additional bounding-box or segmentation supervision.

Three design challenges become visible once joint prediction is pursued. First, the action distribution in driving benchmarks is strongly imbalanced, with turning classes (and in particular left turns) severely under-represented relative to forward and stop, so that naive class reweighting only partially mitigates this bias [15,16]. Second, fusing tokens from multiple cameras without preserving viewpoint identity leads to unstable training, because the network cannot distinguish spatially symmetric evidence such as front-left and back-left cues, an observation consistent with the ablation we report in Section 4.5. Third, open-vocabulary language generation must remain faithful to the evidence used by the prediction heads; fluent but ungrounded text can, in the context of driving, be actively misleading.

To address these challenges, we propose a unified framework, dubbed CrossAD, that couples a single Swin-T encoder with two complementary branches, a semantic–object fusion module, and a label-conditioned language generator. The main contributions of this work are summarised as follows.

1. We present CrossAD, a novel end-to-end framework that integrates a cross-attention scene selector with a class-agnostic object-token extractor over a shared visual encoder,

jointly predicting vehicle behaviour and human-centric reason labels without any detection or segmentation supervision.

2. We introduce the Object Selection Module (OSM), which adds a horizontal positional bias, scores the spatial tokens, retains the most informative subset, and aggregates them through cross-attention with a pooled query to produce a compact scene-level semantic vector. The module is driven entirely by the joint prediction objective and adds only a small number of learnable parameters to the backbone.
3. We design the semantic–object fusion module (SOF), which uses the semantic vector as a query and the object tokens as keys and values in scaled dot-product attention, followed by a residual connection, LayerNorm, and a feed-forward network. The behaviour head reads the fused representation, while the interpretation head reads the semantic vector through a skip connection, so that each head operates on the level of abstraction best matched to its output space.
4. We further introduce two lightweight mechanisms that stabilise training without additional supervision: per-camera embeddings that preserve viewpoint identity on surround-view inputs and a Frobenius-norm diversity regulariser on the Gram matrix of the refined object tokens that encourages complementary rather than redundant evidence.
5. We provide extensive experimental results on two public autonomous-driving benchmarks. On BDD-OIA and nu-AD, the framework achieves the best reported description-prediction F1 on both datasets and the best behaviour mean F1 on nu-AD, while remaining competitive on BDD-OIA behaviour prediction and maintaining practical parameter and runtime cost.

The remainder of the article is organised as follows. Section 2 reviews related work on explainable driving, multi-camera reasoning, vision–language systems, and class-imbalance strategies. Section 3 describes the architecture of the proposed CrossAD framework in detail. Section 4 presents the experimental setup, comparative results, ablation studies, robustness analysis, and natural-language explanation results. Section 5 summarizes the main limitations of the current study. Section 6 concludes the paper and discusses directions for future work.

2. Related Work

In this section, we review the four literature strands that intersect at the framework proposed here: (i) explainable driving formulated as joint behaviour and interpretation prediction; (ii) multi-camera reasoning with and without bird’s-eye-view projection; (iii) vision–language and vision–language–action systems for driving; and (iv) strategies for mitigating action-class imbalance in driving benchmarks.

2.1. Explainable Driving and Joint Behaviour–Interpretation Prediction

The earliest explainable-driving models produced attribution maps, either by gradient back-propagation onto the input image or by highlighting the tokens with the largest attention weights [4–6]. Such heatmaps are computationally cheap, but they provide only a localisation signal, and do not articulate why a region matters. The BDD-X dataset [7] introduced free-form natural-language descriptions of driving clips, and the BDD-OIA benchmark [10] reformulated the problem as joint classification over four driving actions and twenty-one binary reason labels, turning the task into a supervised multi-label learning problem and opening the door to systematic comparison. Inaction [17] and NLE-DM [11] continued this direction, the latter introducing the surround-view nu-AD dataset and evaluating it with semantic-segmentation-based front-ends. Interrelation [18] coupled the outputs of a pretrained object detector with an attention-based relational module, while

concept-bottleneck variants [19–21] applied a general interpretability technique to the same benchmark. More recently, VB-CASeg [12] integrated a vision foundation model through a lightweight adapter and a class-agnostic segmentation objective, achieving state-of-the-art interpretability scores at the time of writing. Recent work by Hu et al. [22] investigates multi-task generalization in autonomous navigation through a human-in-the-loop adversarial reinforcement-learning framework with diffusion policy. This direction is complementary to our work: it focuses primarily on generalizable navigation policy learning under human guidance, whereas CrossAD focuses on explainable visual decision-making by jointly predicting driving behaviour and human-interpretable reason labels from a shared visual encoder. Unlike policy-centered methods, our framework explicitly models semantic-object fusion and produces label-conditioned natural-language justifications, making it suitable for auditing the visual evidence associated with each predicted behaviour.

The framework proposed here differs from the above in two respects. First, the scene branch dispenses with segmentation supervision entirely: the Object Selection Module is a pure top- K token selector driven by the joint prediction objective rather than a segmentation head with a class-agnostic mask. Second, the architecture routes the two prediction heads asymmetrically, so that the behaviour head consumes the fused scene/object representation while the interpretation head consumes the semantic vector directly. To the best of our knowledge, this asymmetric routing has not previously been evaluated for the joint behaviour/interpretation formulation.

2.2. Multi-Camera Reasoning and Surround-View Fusion

Surround-view perception pipelines typically lift multi-camera features into a shared bird’s-eye-view (BEV) representation [23–26] before applying downstream heads for detection, segmentation, or planning. BEV lifting is effective when the target output is geometric (trajectories, occupancy), but it introduces calibration dependencies and discards camera-specific context that is important when explanations are tied to a named viewpoint. UniAD and VAD [27,28] are representative of the BEV-plus-planning line and target trajectory-level outputs; they therefore optimise for a different output space from the structured behaviour/description prediction studied here. In contrast, the surround-view variant of our framework operates entirely in token space: learnable per-camera embeddings label the origin of each spatial token, and the OSM selects tokens across all six views without geometric lifting. This design preserves the viewpoint provenance of every selected token, which the ablation in Section 4.5 shows is essential for area-specific reason labels such as back-left area is clear.

2.3. Vision–Language and Vision–Language–Action Models

Large language models have been coupled with driving perception in several recent systems. DriveGPT4 [29], LMDrive [30], and DriveLM [31] use language as an input–output interface for instruction following or interactive reasoning; the last adopts a graph-VQA formulation that is adjacent to, but not directly comparable with, the multi-label BDD-OIA/nuAD protocol. More recent vision–language–action models such as OpenDriveVLA [32] and OccVLA [33] extend the idea to unified perception and action generation, and X-Driver [34] targets explainability directly through a large vision–language backbone. The framework proposed here is closest to X-Driver in motivation but differs in architecture and deployment profile: we keep the language model compact (~ 1.1 B parameters) and label-conditioned, so that generation is grounded in discrete prediction outputs. This design controls hallucination risk while retaining open-vocabulary phrasing. The language generator itself is adapted via Low-Rank Adaptation [35,36].

2.4. Class Imbalance and Rare Actions in Driving

The left-turn class is rare in driving benchmarks, reflecting the distribution of manoeuvres on public road data more generally. Focal loss [37] and empirical surveys of imbalanced learning [15,16] document the reweighting techniques that have been applied to mitigate this effect, and the long-tail recognition literature [38,39] argues that in some regimes only architectural changes can lift minority-class performance. In the framework proposed here, we combine per-class positive weighting in the behaviour loss with architectural mechanisms that preserve rare local cues, specifically the top- K object selector and the Frobenius-norm diversity regulariser; the empirical contribution of this combination is assessed in Section 4.5.

3. The Proposed Framework

In this section, we introduce the architecture of the proposed CrossAD framework. First, we provide an overview of the full pipeline. We then describe each constituent module in detail.

3.1. Overall Architecture

The overall pipeline is depicted in Figure 1. As shown, the task is to jointly generate a vehicle-behaviour decision and a corresponding multi-label interpretation prediction from a single visual input, either a monocular driving image on BDD-OIA or a six-camera surround-view tuple on nu-AD. In detail, a Swin-T backbone encodes the input into a sequence of spatial tokens, which are consumed in parallel by two complementary branches that share tokens but use independent scoring heads. The scene branch, implemented as the Object Selection Module (OSM), adds a horizontal positional bias, scores the tokens with an MLP selector, retains the most informative subset, and aggregates them through cross-attention with a pooled query to produce a compact semantic vector. The object branch, implemented as the object-token extractor (OTE), scores the same token sequence with a separate MLP and refines the top-scoring tokens into a small class-agnostic object-token set.

The two representations are subsequently combined by the semantic–object fusion module, dubbed SOF, which uses the semantic vector as the query and the object tokens as keys and values in a scaled dot-product attention block, followed by a residual connection, LayerNorm, and a feed-forward network. The resulting fused representation is passed to the behaviour head, while the semantic vector is routed directly to the interpretation head through a skip connection. This asymmetric coupling is intentional: behaviour prediction benefits from object-aware fused evidence, whereas interpretation prediction is better matched to a scene-level semantic summary. Finally, the predicted behaviour and interpretation vectors are consumed by a natural-language explanation module that combines a deterministic rule-based engine with a LoRA-adapted TinyLlama-1.1B language model conditioned on the discrete prediction outputs.

Formally, let X_i be the visual input, which may be a single image or a six-camera tuple. The network predicts a multi-label behaviour vector $B_i \in \{0, 1\}^N$ and a multi-label interpretation vector $E_i \in \{0, 1\}^C$, where $N = 4$ for both benchmarks and $C = 21$ for BDD-OIA and $C = 8$ for nu-AD. The joint mapping is given by

$$f_{\theta}(X_i) \rightarrow (B_i, E_i), \quad 1 \leq i \leq T, \quad (1)$$

where f_{θ} denotes the model parameterized by θ , X_i is the input sample, B_i is the behaviour output, E_i is the interpretation output, i is the sample index, and T is the total number of samples. Sigmoid activations are used for both heads because multiple behaviours and multiple interpretations may be simultaneously active for a given scene.

To make the computational flow in Figure 1 more explicit, Table 1 summarizes the principal tensor dimensions at each stage for both BDD-OIA and nu-AD.

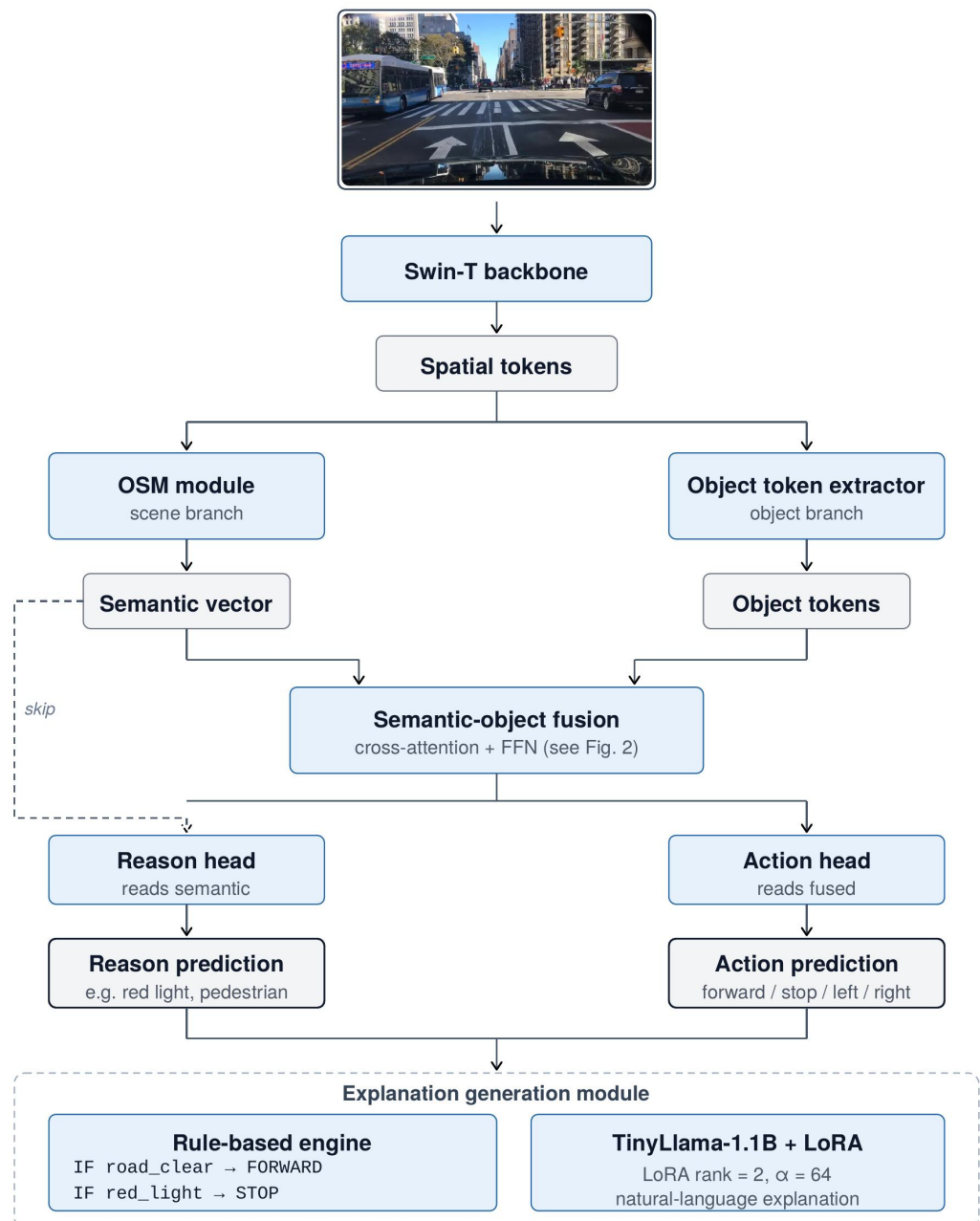


Figure 1. Overall architecture of the proposed CrossAD framework. A driving image (or a six-camera tuple on nu-AD) is encoded by a Swin-T backbone into a sequence of spatial tokens, which are consumed in parallel by the Object Selection Module (scene branch) and the object-token extractor (object branch). The semantic vector and the object tokens are combined by the semantic–object fusion module (see Figure 2). The behaviour head reads the fused representation, while the interpretation head reads the semantic vector through a skip connection. Predicted behaviour and interpretation labels are passed to the explanation generator, which combines a deterministic rule-based engine with a LoRA-adapted TinyLlama-1.1B language model to produce a natural-language justification.

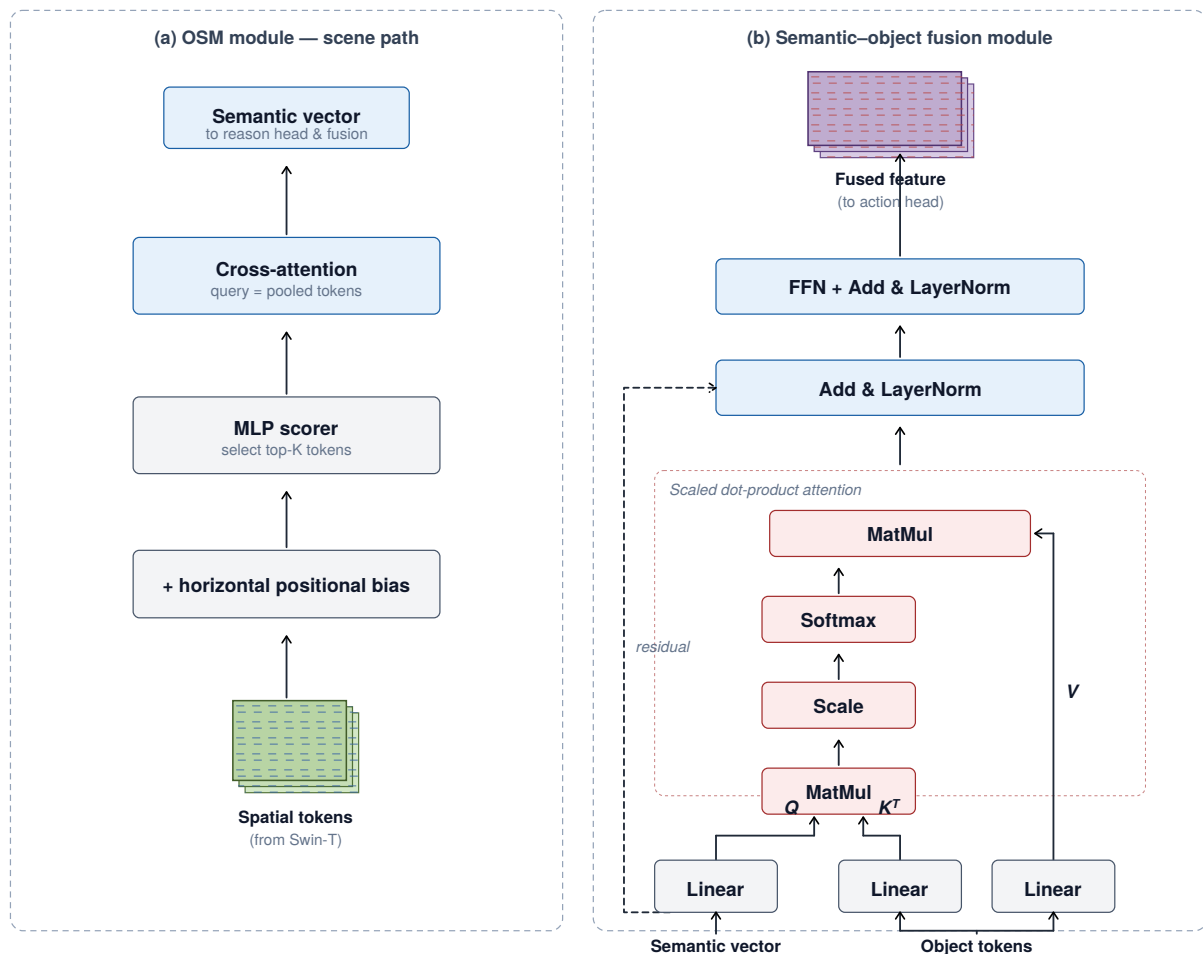


Figure 2. Structure of the two core modules. (a) The Object Selection Module adds a horizontal positional bias, scores the spatial tokens, retains the top- K tokens, and aggregates them by cross-attention with a pooled query to produce the semantic vector. (b) The semantic-object fusion module uses the semantic vector as the query and the object tokens as keys and values; the query is then passed through scaled dot-product attention, a residual with LayerNorm, and a feed-forward network to produce the fused feature consumed by the behaviour head.

Table 1. Tensor-shape summary for the main CrossAD modules.

Component	BDD-OIA Shape	nu-AD Shape
Input image(s)	$3 \times 224 \times 224$	$6 \times 3 \times 224 \times 224$
Swin-T token sequence	$Z \in \mathbb{R}^{49 \times 768}$	$T \in \mathbb{R}^{294 \times 768}$
OSM retained tokens	$Z_{\text{top}} \in \mathbb{R}^{16 \times 768}$	$Z_{\text{top}} \in \mathbb{R}^{64 \times 768}$
Semantic vector	$h_{\text{sem}} \in \mathbb{R}^{768}$	$h_{\text{sem}} \in \mathbb{R}^{768}$
Object tokens	$O \in \mathbb{R}^{8 \times 768}$	$O \in \mathbb{R}^{8 \times 768}$
Fused representation	$h_{\text{fuse}} \in \mathbb{R}^{768}$	$h_{\text{fuse}} \in \mathbb{R}^{768}$
Behaviour output	$\hat{B}_i \in \mathbb{R}^4$	$\hat{B}_i \in \mathbb{R}^4$
Interpretation output	$\hat{E}_i \in \mathbb{R}^{21}$	$\hat{E}_i \in \mathbb{R}^8$

3.2. Visual Encoder

The visual encoder must simultaneously preserve fine-grained local cues, such as lane markings, traffic lights, and cyclists, and model the wider road layout. A hierarchical Swin Transformer [40] provides a favourable trade-off between representational quality and cost: attention is computed within local windows, cross-window information is exchanged through the shifted-window mechanism, and patch merging produces progressively coarser

feature maps. We adopt Swin-T initialised from ImageNet pretraining. For a 224×224 input image, the final-stage feature map is passed through a LayerNorm and reshaped into $N = 49$ spatial tokens,

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N], \quad \mathbf{z}_j \in \mathbb{R}^d, \quad (2)$$

with $d = 768$. The sequence $\mathbf{Z}$ is shared by the scene and object branches and serves as the common substrate from which both representations are derived.

3.3. Object Selection Module (Scene Branch)

Tokens produced by the backbone are not equally informative: sky, vegetation, and uniform road surface constitute most of the sequence but carry limited discriminative value for either behaviour or interpretation prediction. To filter the sequence and aggregate the retained tokens into a single semantic vector, we introduce the Object Selection Module, dubbed OSM.

In detail, a horizontal positional bias is first injected into the token sequence. Let $x_j \in [-1, 1]$ be the horizontal coordinate of token j in the 7×7 spatial grid. A learnable linear map projects this scalar into the token space,

$$\mathbf{p}_j = W_p x_j + \mathbf{b}_p, \quad \tilde{\mathbf{z}}_j = \mathbf{z}_j + \mathbf{p}_j, \quad W_p \in \mathbb{R}^{d \times 1}, \mathbf{b}_p \in \mathbb{R}^d, \quad (3)$$

preserving the left/right information on which turning manoeuvres depend. To be specific, each position-aware token is then scored by a learned linear scorer,

$$s_j = W_s \tilde{\mathbf{z}}_j + b_s, \quad (4)$$

and the top K_{osm} tokens are retained as the reduced set $\mathbf{Z}_{\text{top}}$. Empirically, we set $K_{\text{osm}} = 16$ on BDD-OIA and $K_{\text{osm}} = 64$ on nu-AD, the values being chosen on the validation split. Rather than using a bank of learned semantic queries, we form a single pooled scene query by averaging the full position-aware sequence,

$$\mathbf{q}_{\text{pool}} = \frac{1}{N} \sum_{j=1}^N \tilde{\mathbf{z}}_j, \quad (5)$$

and apply multi-head cross-attention,

$$\mathbf{h}_{\text{sem}} = \text{MHA}(\mathbf{q}_{\text{pool}}, \mathbf{Z}_{\text{top}}, \mathbf{Z}_{\text{top}}). \quad (6)$$

The output $\mathbf{h}_{\text{sem}} \in \mathbb{R}^d$ is the semantic vector: a saliency-weighted scene summary used both as a direct input to the interpretation head and as the query in the semantic-object fusion block. The total number of trainable parameters of the OSM is approximately 2.4 M. Mean pooling is adopted here for simplicity; we verified on the validation split that a single-layer attention-pooled query did not improve the metric reported in Section 4.4.

3.4. Object-Token Extractor (Object Branch)

The semantic vector is well matched to interpretation prediction, for which the outputs describe the scene as a whole (red light, pedestrian in crosswalk). Behaviour prediction, however, frequently turns on a small number of localised entities, such as the lead vehicle, the color of the nearest signal, or a cyclist entering the corridor. A single pooled scene vector can wash out these focal cues; to preserve them, we propose the object-token extractor (OTE), a second lightweight branch that operates in parallel with the OSM over the same token sequence $\mathbf{Z}$.

Each token is scored by a small MLP,

$$r_j = g(\mathbf{z}_j), \quad (7)$$

and the top K_{obj} tokens are gathered into a raw set $\mathbf{O}_{\text{raw}}$. A refinement block with a linear layer, GELU activation, dropout, and LayerNorm produces the final object-token set,

$$\mathbf{O} = \text{LayerNorm}\left(\text{Dropout}\left(\text{GELU}(W_o \mathbf{O}_{\text{raw}} + \mathbf{b}_o)\right)\right) \in \mathbb{R}^{K_{\text{obj}} \times d}. \quad (8)$$

The extractor is class-agnostic: no detection, bounding-box, or segmentation supervision is used, and the scorer and refinement layers are shaped end-to-end by the downstream fusion module and the behaviour head. We set $K_{\text{obj}} = 8$ throughout, the value chosen on the validation split over $K_{\text{obj}} \in \{4, 8, 16\}$. The total number of trainable parameters of the OTE is approximately 1.2 M. A diversity penalty, defined in Section 3.7, encourages the resulting tokens to cover complementary evidence.

3.5. Semantic–Object Fusion Module

The semantic vector and the object tokens describe the scene at different levels of abstraction and must be combined before the behaviour head can use them jointly. This combination is performed by the semantic–object fusion module (SOF), whose structure is shown in Figure 2. In detail, the semantic vector, viewed as a $1 \times d$ query, is paired with the object tokens as keys and values in multi-head cross-attention,

$$\mathbf{a} = \text{MHA}(\mathbf{h}_{\text{sem}}, \mathbf{O}, \mathbf{O}), \quad \mathbf{a} \in \mathbb{R}^d. \quad (9)$$

A residual connection with LayerNorm gives

$$\mathbf{u} = \text{LayerNorm}(\mathbf{h}_{\text{sem}} + \mathbf{a}), \quad (10)$$

and a two-layer feed-forward network with a second residual and LayerNorm produces the fused representation,

$$\mathbf{h}_{\text{fuse}} = \text{LayerNorm}(\mathbf{u} + \text{FFN}(\mathbf{u})) \in \mathbb{R}^d. \quad (11)$$

Intuitively, the semantic vector determines which kinds of evidence matter for the current scene, and the attention weights in Equation (9) select the object tokens that carry that evidence. The fused output therefore combines holistic scene character with the localised cues that support a given decision. The total number of trainable parameters of the SOF module is approximately 4.1 M.

3.6. Prediction Heads and Asymmetric Routing

The two heads operate on different representations,

$$\hat{B}_i = \sigma(\text{MLP}_{\text{beh}}(\mathbf{h}_{\text{fuse}})), \quad (12)$$

$$\hat{E}_i = \sigma(\text{MLP}_{\text{int}}(\mathbf{h}_{\text{sem}})), \quad (13)$$

where σ denotes the element-wise sigmoid and each MLP consists of two fully connected layers. The skip connection from the semantic vector to the interpretation head, visible in Figure 1, is deliberate. Interpretation labels describe the scene at the level of abstraction that $\mathbf{h}_{\text{sem}}$ already captures, and routing them through the fusion module would inject object-level noise irrelevant to their prediction. Conversely, behaviour labels depend on the specific localised entities summarised in $\mathbf{h}_{\text{fuse}}$. The design is consistent with the ablation in

Section 4.5, where interpretation F1 follows a cleaner monotonic pattern across surround-view configurations than behaviour F1, as predicted by the abstraction-level argument.

3.7. Loss Design

We propose a multi-task loss function composed of three complementary terms: a vehicle-behaviour loss $\mathcal{L}_{VB}$, an interpretation-prediction loss $\mathcal{L}_{des}$, and an object-diversity regulariser $\mathcal{L}_{div}$.

For behaviour prediction, we adopt a class-weighted binary cross-entropy loss with fixed class weights $\mathbf{w}_B = [1.0, 1.0, 2.0, 2.0]$ over {Forward, Stop, Left, Right}, emphasising the two rarer turning classes,

$$\mathcal{L}_{VB} = \text{BCE}_{\mathbf{w}_B}(\hat{B}_i, B_i). \quad (14)$$

For interpretation prediction, we use per-label positive weights computed from the training-set label frequencies and clipped to a maximum value, which prevents very rare labels from dominating the gradient,

$$\mathcal{L}_{des} = \text{BCE}_{\mathbf{w}^+}(\hat{E}_i, E_i). \quad (15)$$

To encourage the object tokens to cover complementary rather than redundant evidence, we apply a Frobenius-norm diversity penalty to the Gram matrix of the L_2 -normalised refined tokens. Let $\mathbf{O} \in \mathbb{R}^{K_{\text{obj}} \times d}$ denote the refined object-token matrix from Equation (8), and let $\tilde{\mathbf{O}}$ denote its row-wise L_2 -normalised version, $\tilde{\mathbf{O}}_{k,:} = \mathbf{O}_{k,:} / \|\mathbf{O}_{k,:}\|_2$ for $k = 1, \dots, K_{\text{obj}}$. Then

$$\mathcal{L}_{div} = \|\tilde{\mathbf{O}}\tilde{\mathbf{O}}^\top - I_{K_{\text{obj}}}\|_F^2, \quad I_{K_{\text{obj}}} \in \mathbb{R}^{K_{\text{obj}} \times K_{\text{obj}}}. \quad (16)$$

The total training objective is then calculated as follows:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{VB} + \beta \cdot \mathcal{L}_{des} + \lambda \cdot \mathcal{L}_{div}, \quad (17)$$

where α , β , and λ are the hyper-parameters that weight the vehicle-behaviour, interpretation-prediction, and object-diversity terms, respectively. We fix $(\alpha, \beta, \lambda) = (0.5, 1.0, 0.005)$ throughout all experiments; this setting was chosen on a small validation sweep over $\alpha \in \{0.3, 0.5, 1.0\}$, $\beta \in \{1.0, 1.5\}$, and $\lambda \in \{0.001, 0.005, 0.01\}$. The higher weight on $\mathcal{L}_{des}$ reflects the dense supervision that the interpretation labels provide, whereas the small diversity coefficient is sufficient to regularise the object branch without dominating the main prediction terms.

3.8. Multi-Camera Surround-View Fusion

For the nu-AD benchmark, behaviour decisions and interpretation labels depend on evidence distributed across a six-camera rig. Concatenating the per-camera token sequences directly is insufficient: without viewpoint information, the OSM cannot distinguish spatially symmetric cues from, for example, the front-left and back-left cameras, and this ambiguity empirically prevents convergence.

To resolve the issue, we introduce a lightweight viewpoint-identity mechanism. A learnable camera embedding $\mathbf{e}_k \in \mathbb{R}^d$ is defined for each of the six cameras ($k = 1, \dots, 6$), shared across all scenes, and added element-wise to every token originating from camera k . Addition is preferred over concatenation because it preserves the token dimension d and therefore does not require an additional projection; the same choice is standard for

positional embeddings in transformer encoders. The identity-augmented sequences are concatenated into a unified surround-view token matrix,

$$\mathbf{T} = \text{Concat}(\mathbf{Z}^{(1)} + \mathbf{e}_1, \dots, \mathbf{Z}^{(6)} + \mathbf{e}_6) \in \mathbb{R}^{6N \times d}. \quad (18)$$

The OSM and the object-token extractor then attend over all $6N$ tokens simultaneously without any structural modification, selecting evidence from whichever camera is most informative for a given prediction. This design is particularly consequential for area-specific interpretation labels such as back-left area is clear, which cannot be observed from the frontal camera alone. The camera embeddings add only $6 \times 768 = 4608$ parameters, a negligible overhead relative to the backbone.

3.9. Natural-Language Explanation Module

Generating faithful natural-language explanations from structured prediction outputs requires balancing two competing desiderata: coverage, in that every predicted label combination must yield a well-formed explanation, and expressiveness, in that explanations should not be limited to rigid templates. The module shown at the bottom of Figure 1 addresses both through a complementary dual-generator design.

In detail, a deterministic rule-based engine maintains a priority-ordered mapping from active reason-label combinations to grammatically complete explanatory sentences. For each input, the active labels are identified from $\hat{E}_i$ and matched against a handcrafted template hierarchy of 47 templates for BDD-OIA and 18 templates for nu-AD, and the highest-priority matching template is instantiated. Example rules include IF `road_clear` $\rightarrow$ FORWARD and IF `red_light` $\rightarrow$ STOP; the rule set covers every label combination observed in the two benchmarks with no generation failure. The engine requires no GPU resources at inference, and its determinism makes it directly auditable, which is a useful property in safety-critical deployments.

For fluent open-vocabulary explanations beyond the expressive constraints of fixed templates, we fine-tune TinyLlama-1.1B [36] with Low-Rank Adaptation [35] using rank $r = 2$ and scaling factor $\alpha_{\text{LoRA}} = 64$, restricting updates to the attention projection layers. Fine-tuning uses stratified, template-diverse prompt-completion pairs ($\sim 12,300$ pairs in total, split 90/10 train/validation), with each prompt encoding the predicted behaviour and interpretation label vectors and each completion being a reference explanation that exhibits varied syntactic structure. Stratified sampling ensures proportional representation of all label combinations, including rare configurations involving left turns or complex multi-label scenes. Importantly, the module is label-conditioned: it operates on the discrete prediction outputs rather than on continuous visual embeddings, which keeps the generated text aligned with the structured behaviour and interpretation predictions. This design reduces unsupported open-ended generation, but it does not by itself establish causal faithfulness to the raw visual evidence; that faithfulness remains bounded by the correctness and grounding of the prediction heads. TinyLlama generator on the BDD-OIA held-out test set.

The language-generation faithfulness evaluation considers the LoRA-adapted TinyLlama generator on the BDD-OIA held-out test set.

The deterministic rule-based engine is retained as an auditable lightweight deployment path, but it is not used to compute the reported decision-faithfulness or task-specific evaluation scores.

3.10. Notation Summary and Forward-Pass Pseudocode

For clarity and reproducibility, Table 2 consolidates the principal variables used throughout the proposed framework.

Algorithm 1 summarizes the computational process of a single CrossAD forward pass during inference.

Algorithm 1: Single forward pass of CrossAD during inference.

Inputs: Visual sample X_i ; a single image for BDD-OIA or a six-view set $\{X_i^1, \dots, X_i^6\}$ for nu-AD; shared Swin-T backbone SwinT; token-retention sizes K_{osm} and K_{obj} ; thresholds τ_{beh} and τ_{des} .

Outputs: Behaviour probability vector $\mathbf{b}_{\text{score}} \in \mathbb{R}^{N_{\text{beh}}}$, interpretation probability vector $\mathbf{e}_{\text{score}} \in \mathbb{R}^{N_{\text{int}}}$, and generated explanation y_i .

Notation: d is the token dimension, N is the number of single-view tokens, N_{tot} is the token count after multi-camera concatenation, and d_k, d_v are the key/query and value dimensions. Attention denotes scaled dot-product attention with learned query, key, value, and output projections; in the multi-head setting, the operation is applied independently across heads and the head outputs are concatenated and projected. Dropout is disabled during inference.

1. Encode the input into spatial visual tokens. If X_i is single-view, set $\mathbf{Z} \leftarrow \text{SwinT}(X_i)$ and $N_{\text{tot}} \leftarrow N$; otherwise, for $k = 1, \dots, 6$, compute $\mathbf{Z}_k \leftarrow \text{SwinT}(X_i^k)$ and $\mathbf{T}_k \leftarrow \mathbf{Z}_k + \mathbf{e}_k$, then set $\mathbf{Z} \leftarrow \text{Concat}(\mathbf{T}_1, \dots, \mathbf{T}_6)$ and $N_{\text{tot}} \leftarrow \sum_k N_k$.
 2. For each token $\mathbf{z}_j \in \mathbf{Z}$, $j = 1, \dots, N_{\text{tot}}$, compute $\tilde{\mathbf{z}}_j \leftarrow \mathbf{z}_j + \mathbf{p}_j$ and $s_j \leftarrow \text{MLP}_{\text{osm}}(\tilde{\mathbf{z}}_j)$, where $\mathbf{p}_j$ is the horizontal positional bias associated with the spatial position of token j ; for multi-camera input, the same positional-bias rule is applied consistently within each camera stream.
 3. Select the OSM scene tokens $\mathbf{Z}_{\text{top}} \leftarrow \text{TopK}(\{\tilde{\mathbf{z}}_j\}, \{s_j\}, K_{\text{osm}})$, with $\mathbf{Z}_{\text{top}} \in \mathbb{R}^{K_{\text{osm}} \times d}$.
 4. Compute the pooled semantic query $\mathbf{q}_{\text{pool}} \leftarrow \text{Mean}(\mathbf{Z}_{\text{top}})$, where $\mathbf{q}_{\text{pool}} \in \mathbb{R}^{1 \times d}$.
 5. Compute the scene-level semantic representation $\mathbf{h}_{\text{sem}} \leftarrow \text{MHA}(\mathbf{q}_{\text{pool}}, \mathbf{Z}_{\text{top}}, \mathbf{Z}_{\text{top}})$, where $\mathbf{h}_{\text{sem}} \in \mathbb{R}^{1 \times d}$.
 6. For each raw token $\mathbf{z}_j \in \mathbf{Z}$, compute $r_j \leftarrow \text{MLP}_{\text{obj}}(\mathbf{z}_j)$. This object-token path scores raw tokens without horizontal positional bias, so object selection remains driven by localized visual evidence rather than directional scene priors.
 7. Select the object-relevant tokens $\mathbf{O}_{\text{raw}} \leftarrow \text{TopK}(\{\mathbf{z}_j\}, \{r_j\}, K_{\text{obj}})$, with $\mathbf{O}_{\text{raw}} \in \mathbb{R}^{K_{\text{obj}} \times d}$.
 8. Refine the selected object tokens: $\mathbf{O} \leftarrow \text{LayerNorm}(\text{GELU}(\text{Linear}(\mathbf{O}_{\text{raw}})))$, where $\mathbf{O} \in \mathbb{R}^{K_{\text{obj}} \times d}$; dropout is omitted because inference mode is assumed.
 9. Project semantic and object features: $\mathbf{Q} \leftarrow \text{Linear}_Q(\mathbf{h}_{\text{sem}})$, $\mathbf{K}_{\text{att}} \leftarrow \text{Linear}_K(\mathbf{O})$, and $\mathbf{V}_{\text{att}} \leftarrow \text{Linear}_V(\mathbf{O})$.
 10. Compute the attended object representation $\mathbf{o}_{\text{att}} \leftarrow \text{Linear}_O(\text{softmax}(\mathbf{Q}\mathbf{K}_{\text{att}}^\top / \sqrt{d_k})\mathbf{V}_{\text{att}})$, where $\mathbf{o}_{\text{att}} \in \mathbb{R}^{1 \times d}$.
 11. Fuse semantic and object-aware representations: $\mathbf{u} \leftarrow \text{LayerNorm}(\mathbf{h}_{\text{sem}} + \mathbf{o}_{\text{att}})$ and $\mathbf{h}_{\text{fuse}} \leftarrow \text{LayerNorm}(\mathbf{u} + \text{FFN}(\mathbf{u}))$.
 12. Predict behaviour probabilities: $\mathbf{b}_{\text{score}} \leftarrow \sigma(\text{Head}_{\text{beh}}(\mathbf{h}_{\text{fuse}}))$, with $\mathbf{b}_{\text{score}} \in \mathbb{R}^{N_{\text{beh}}}$.
 13. Determine active behaviour labels: $\mathcal{A}_i \leftarrow \{m : \mathbf{b}_{\text{score}}[m] \geq \tau_{\text{beh}}\}$.
 14. Predict interpretation probabilities: $\mathbf{e}_{\text{score}} \leftarrow \sigma(\text{Head}_{\text{int}}(\mathbf{h}_{\text{sem}}))$, with $\mathbf{e}_{\text{score}} \in \mathbb{R}^{N_{\text{int}}}$.
 15. Determine active interpretation labels: $\mathcal{C}_i \leftarrow \{c : \mathbf{e}_{\text{score}}[c] \geq \tau_{\text{des}}\}$.
 16. Generate the natural-language explanation: $y_i \leftarrow \text{ExplanationGenerator}(\text{BuildPrompt}(\mathcal{A}_i, \mathcal{C}_i))$.
 17. Return $\mathbf{b}_{\text{score}}$, $\mathbf{e}_{\text{score}}$, and y_i .
-

Table 2. Nomenclature and variable definitions used in the CrossAD formulation.

Symbol	Meaning
X_i	Input image or six-camera input tuple
$B_i, \hat{B}_i$	Ground-truth and predicted behaviour vector
$E_i, \hat{E}_i$	Ground-truth and predicted interpretation vector
N	Number of spatial tokens per image, with $N = 49$
d	Token feature dimension, with $d = 768$
$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N]$	Spatial token sequence from Swin-T
x_j	Horizontal coordinate of token j
$\mathbf{p}_j$	Learnable horizontal positional bias
$\tilde{\mathbf{z}}_j$	Position-augmented token
s_j	OSM token-selection score
K_{osm}	Number of retained OSM tokens
$\mathbf{Z}_{\text{top}}$	Top-ranked scene tokens selected by OSM
$\mathbf{q}_{\text{pool}}$	Pooled query used in OSM cross-attention
$\mathbf{h}_{\text{sem}}$	Scene-level semantic vector
r_j	Object-token score
K_{obj}	Number of object tokens
$\mathbf{O}_{\text{raw}}, \mathbf{O}$	Raw and refined object-token sets
$\mathbf{a}$	Attention output in SOF
$\mathbf{u}$	Residual-normalized SOF intermediate vector
$\mathbf{h}_{\text{fuse}}$	Fused semantic-object representation
$\mathbf{e}_k$	Learnable camera embedding for camera k
$\mathbf{T}$	Six-camera token matrix after adding camera embeddings
$\mathcal{L}_{\text{VB}}$	Vehicle-behaviour loss (multi-label BCE)
$\mathcal{L}_{\text{des}}$	Interpretation-prediction loss (weighted multi-label BCE)
$\mathcal{L}_{\text{div}}$	Object-token diversity regularizer
$\mathcal{L}_{\text{total}}$	Total training objective
α, β, λ	Loss-weighting hyperparameters

4. Experiments and Results

The main focus of this work is a unified framework that jointly predicts vehicle behaviour and possible human-centric interpretations from a shared visual encoder, using an Object Selection Module, an object-token extractor, and a semantic-object fusion module. In this section, we present a comprehensive set of experiments that demonstrate the capabilities of the proposed method.

4.1. Datasets

We conduct our experimental trials on two public autonomous-driving benchmarks, **BDD-OIA** [10] and **nu-AD** [11], in order to demonstrate the performance and generalisation of the proposed method. Table 3 summarises their main properties. Both datasets use the same four-class behaviour labelling (Forward, Stop, Left, Right), but differ in the number of description categories ($C = 21$ for BDD-OIA and $C = 8$ for nu-AD), the number of cameras, and the underlying video source.

The BDD-OIA training split is strongly imbalanced: the left-turn class is represented by only 7.5% of training samples (Table 4), which is the principal source of behaviour-side error reported in Section 4.4. Table 5 reports the per-label positive rate on nu-AD.

Table 3. Summary of the two benchmarks used in our experiments.

Dataset	Source	Cameras	N_{beh}	C_{des}
BDD-OIA [10]	BDD100K	1	4	21
nu-AD [11]	nuScenes	6	4	8

Table 4. Action-label distribution in the BDD-OIA training split. Labels are not mutually exclusive: a single sample may be tagged with more than one feasible action.

Action	Count	Frequency
Forward	8734	54.3%
Stop	7277	45.2%
Left	1210	7.5%
Right	10,660	66.3%

Table 5. Per-label positive rate in nu-AD. Rows are not mutually exclusive; each value denotes the fraction of frames in which the label is active. Distribution reproduced from [41].

Label	Positive Rate (%)
Move forward	80.76
Stop/slow down	19.23
Turn left	29.43
Turn right	38.07
Traffic light allows	84.14
Front area clear	89.64
Solid line left	27.61
Solid line right	23.74

4.2. Implementation Details

We implement the proposed framework in PyTorch 2.7.1 with CUDA 11.8. All experiments are conducted on a single NVIDIA RTX 3080 Ti GPU. Input images are resized to 256×256 pixels and randomly cropped to 224×224 during training; a deterministic centre crop is used at inference. Color jitter is applied as data augmentation. Random horizontal flipping is disabled on BDD-OIA because it would invert the left/right semantics on which the directional interpretation labels depend. On nu-AD, we use a geometry-consistent flip in which the front-left and front-right streams are exchanged together with their back-left and back-right counterparts, and the affected action and interpretation labels (turn_left, turn_right, front_left_clear, front_right_clear, and so on) are remapped accordingly.

Optimisation uses AdamW with weight decay 10^{-4} , cosine annealing after a three-epoch linear warm-up, bfloat16 mixed precision, gradient clipping at an ℓ_2 -norm of 0.5, logit clamping, and skipping of non-finite batches. The Swin-T backbone is fine-tuned with a small learning rate of 2×10^{-6} ; the task-specific modules (OSM, OTE, SOF, and prediction heads) are trained from random initialisation with a learning rate of 10^{-4} . The backbone is frozen during the first three epochs and then gradually unfrozen. A mini-batch of 2 surround-view tuples is used to fit within the GPU memory budget; gradients are accumulated over 8 mini-batches, yielding an effective batch size of 16.

4.3. Evaluation Metrics

Following the protocol established in prior work [10,11], we report the overall F1 score and the mean F1 score for both the behaviour-decision and the interpretation-prediction heads. The overall behaviour F1 is the instance-level average,

$$F1_{\text{all}}^{\text{behaviour}} = \frac{1}{T} \sum_{i=1}^T F1(B_i, \hat{B}_i), \quad (19)$$

and the mean behaviour F1 is the class-wise average,

$$F1_m^{\text{behaviour}} = \frac{1}{N} (F1_F + F1_S + F1_L + F1_R), \quad (20)$$

where $F1_F, F1_S, F1_L, F1_R$ are the per-class F1 scores for Forward, Stop, Left, and Right. Analogous metrics $F1_m^{des}$ and $F1_{all}^{des}$ are defined over the description categories. For the natural-language generation component, we additionally report corpus BLEU [42] and ROUGE-1/2/L [43] against reference explanations.

4.4. Comparative Results

In this section, we compare the proposed CrossAD framework with representative explainable-driving baselines on both BDD-OIA and nu-AD to demonstrate its superiority in the joint prediction of vehicle behaviours and their corresponding interpretations.

4.4.1. Performance on BDD-OIA

Table 6 compares the proposed framework with representative baselines on the BDD-OIA test set: Local selector [10], OIA [10], Inaction [17], CBM [19], CBM-AUC [20], C-SENN [21], NLE-DM [11], and Interrelation [18]. Figures 3–5 provide complementary visual comparisons.

Table 6. Comparative results on the BDD-OIA test set. F, S, L, R denote Forward, Stop, Left, Right. Best value per column in bold; dashes indicate values not reported by the original authors.

Method	$F1_m^{beh}$	$F1_{all}^{beh}$	$F1_m^{des}$	$F1_{all}^{des}$	F	S	L	R
Local selector [10]	0.699	0.711	0.196	0.406	0.810	0.762	0.600	0.624
OIA [10]	0.718	0.734	0.208	0.422	0.829	0.781	0.630	0.634
Inaction [17]	0.694	0.714	0.347	0.565	0.800	0.747	0.612	0.619
CBM [19]	0.610	0.661	0.292	0.412	0.795	0.732	0.483	0.431
CBM-AUC [20]	0.658	0.704	0.342	0.522	0.803	0.751	0.551	0.525
C-SENN [21]	0.618	---	---	---	0.772	0.744	0.469	0.486
NLE-DM [11]	0.723	0.733	0.312	0.517	0.827	0.760	0.651	0.653
Interrelation [18]	0.701	0.722	0.335	0.537	0.802	0.753	0.619	0.625
CrossAD (Ours)	0.716	0.739	0.417	0.584	0.839	0.803	0.580	0.641

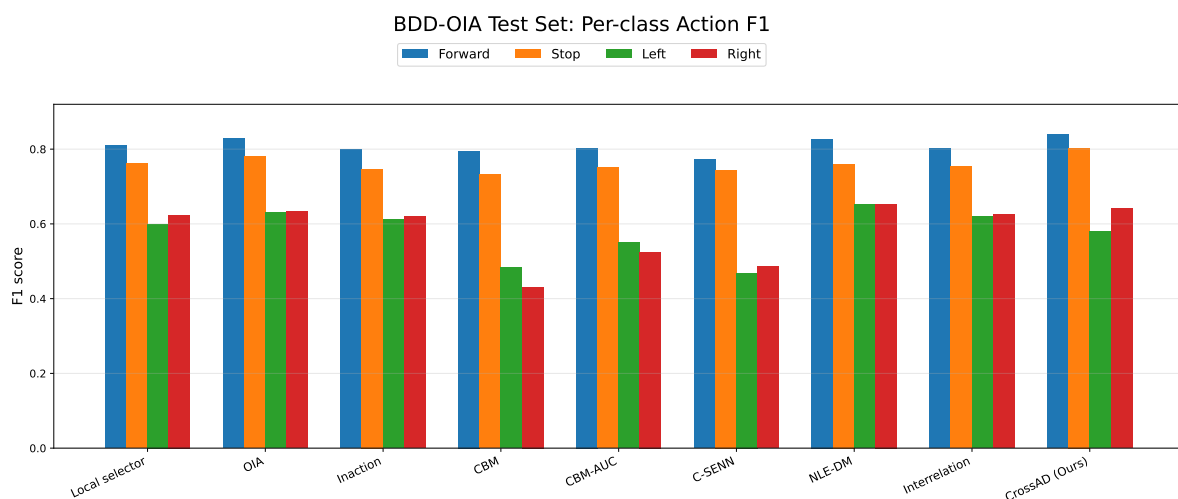


Figure 3. Per-class behaviour F1 on the BDD-OIA test set. CrossAD leads on Forward and Stop; Left remains the most difficult class and the principal source of behaviour-aggregate error.

As shown in Table 6, the proposed CrossAD achieves the strongest interpretation performance among the compared methods, with $F1_m^{des} = 0.417$ and $F1_{all}^{des} = 0.584$. It also attains the highest Forward-class and Stop-class F1 scores (0.839 and 0.803, respectively) and the highest overall behaviour F1 (0.739) within the set of baselines reported in Table 6. However, CrossAD does not achieve the best behaviour mean F1, where NLE-DM reports

0.723 compared with our 0.716. This shortfall is concentrated in the left-turn class, which is represented by only 7.5% of training samples (Table 4) and on which CrossAD (0.580) remains below NLE-DM (0.651). The interpretation-side result is consistent with the principal design hypothesis of the framework: routing the interpretation head to the scene-level semantic vector exposes it to a cleaner, less object-dependent signal than the fused representation, which translates into the strongest description-level predictions among the compared methods.

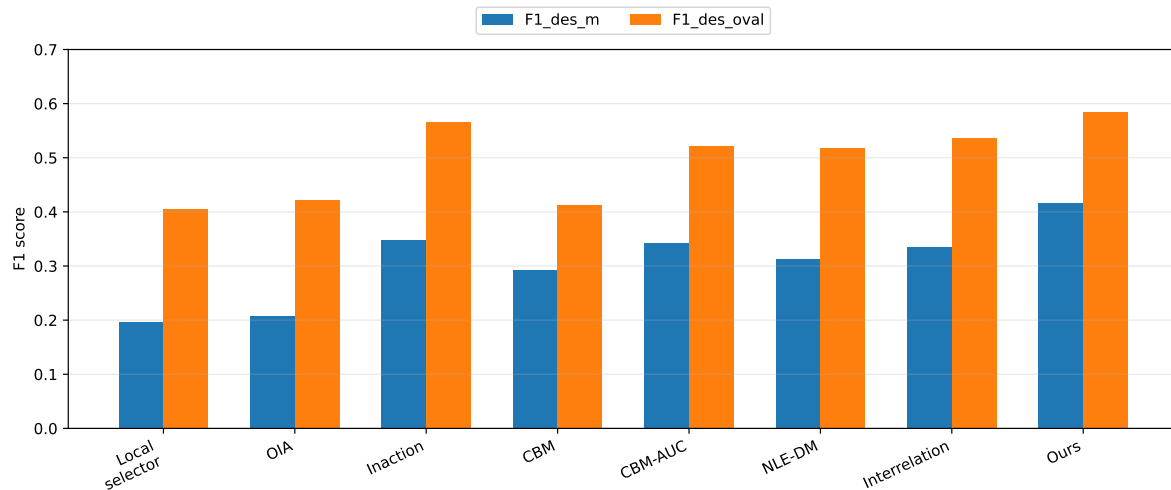


Figure 4. Interpretation performance on the BDD-OIA test set. CrossAD attains the best values on both $F1_m^{des}$ and $F1_{all}^{des}$.

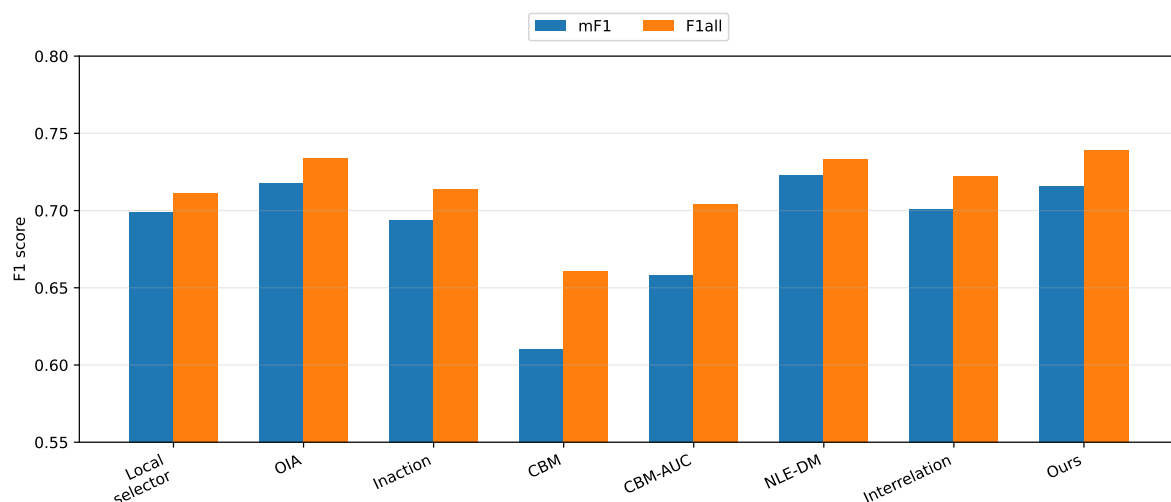


Figure 5. Aggregate behaviour F1 on the BDD-OIA test set.

4.4.2. Performance on nu-AD

Table 7 reports the comparison between CrossAD and three representative prior methods on the nu-AD test split.

CrossAD obtains the strongest reported results on every aggregate metric for which a comparison can be drawn. On the behaviour side, it improves the label-averaged F1 from 0.767 (VB-CASeg) to 0.848 (+0.081) and the instance-averaged F1 from 0.844 to 0.887 (+0.043); the corresponding gains over NLE-DM and OIA are larger still. The behaviour-side per-class scores, which prior work on this benchmark has not made available, are highest on the two majority classes, Forward (0.963) and Stop (0.852), and lowest on Left (0.760). This pattern is consistent with the action-frequency distribution of the nu-AD

training split and with the parallel observation on BDD-OIA in Section 4.4.1, where left-turn prediction is similarly limited by the rarity of the corresponding training samples.

Table 7. Comparative results on the nu-AD test split. F, S, L, and R denote Forward, Stop, Left, and Right, respectively. The best value in each column is shown in bold; dashes indicate unreported results.

Method	F	S	L	R	$F1_m^{\text{beh}}$	$F1_{\text{all}}^{\text{beh}}$	$F1_m^{\text{des}}$	$F1_{\text{all}}^{\text{des}}$
OIA [10]	---	---	---	---	0.739	0.771	0.853	0.852
NLE-DM [11]	---	---	---	---	0.760	0.836	0.882	0.879
VB-CASeg [12]	---	---	---	---	0.767	0.844	0.907	0.903
CrossAD (Ours)	0.963	0.852	0.760	0.818	0.848	0.887	0.909	0.919

The interpretation-side results follow the same direction, though with substantially smaller margins. CrossAD reaches $F1_m^{\text{des}} = 0.909$ and $F1_{\text{all}}^{\text{des}} = 0.919$, exceeding VB-CASeg by +0.002 and +0.016, respectively, and NLE-DM by +0.027 and +0.040. The interpretation-side advantage is also consistent with the asymmetric routing introduced in Section 3.6: the interpretation head reads the scene-level semantic vector $\mathbf{h}_{\text{sem}}$ directly, whereas the behaviour head consumes the fused representation $\mathbf{h}_{\text{fuse}}$, which is more sensitive to optimisation difficulty on the minority turning classes.

The class-wise behaviour pattern is broadly compatible with the design rationale developed in Section 3.6. Strong performance on Stop is plausible in scenes whose decision is driven by a small number of localised cues, such as a decelerating lead vehicle, a stop line, or a pedestrian entering the corridor; the object branch is explicitly designed to preserve such focal evidence for the behaviour head.

4.5. Ablation Study

This section isolates the contributions of the surround-view fusion strategy, the camera-embedding mechanism, and the training procedure. All variants are evaluated on the nu-AD test split under the protocol of Section 4.4.2, ensuring direct comparability with the results reported in Table 7.

4.5.1. Surround-View Fusion: Effect on Behaviour and Interpretation

We construct four variants that progressively activate the components of the surround-view fusion design. Variant 1 (V1) is a single-camera baseline restricted to the front view and trained under the default schedule. Variant 2 (V2) incorporates the remaining five cameras through element-wise averaging of the per-camera token sequences, providing additional visual evidence but no viewpoint-identity signal. Variant 3 (V3) replaces this averaging with the learnable per-camera embeddings and adopts a cyclic learning-rate schedule [44], with reported numbers corresponding to the best validation checkpoint. Variant 4 (V4) is the full model, evaluated at its early-stopping checkpoint. Behaviour and interpretation results are reported in Tables 8 and 9, respectively.

Table 8. Ablation on nu-AD (test split): effect of the surround-view fusion configuration on behaviour prediction. F, S, L, and R denote Forward, Stop, Left, and Right, respectively. The best value in each column is shown in bold.

Variant	Setting	Cams	F	S	L	R
V1	Single-camera baseline	1	0.963	0.824	0.739	0.773
V2	Multi-cam, token averaging	6	0.954	0.785	0.582	0.708
V3	Multi-cam + cam-embed. + cyclic LR	6	0.958	0.837	0.752	0.805
V4	Full (best checkpoint)	6	0.963	0.852	0.760	0.818

Table 9. Ablation on nu-AD (test split): effect of the surround-view fusion configuration on interpretation prediction. The best value in each column is shown in bold.

Variant	Setting	$F1_m^{\text{des}}$	$F1_{\text{all}}^{\text{des}}$	$F1_{\text{all}}^{\text{beh}}$
V1	Single-camera baseline	0.862	0.867	0.851
V2	Multi-cam, token averaging	0.747	0.780	0.792
V3	Multi-cam + cam-embed. + cyclic LR	0.905	0.917	0.877
V4	Full (best checkpoint)	0.909	0.919	0.887

Beginning with V1, the front-camera configuration achieves $F1_m^{\text{beh}} = 0.825$ and $F1_m^{\text{des}} = 0.862$. While performance is already strong on the two majority behaviour classes—Forward (0.963) and Stop (0.824)—it deteriorates on the minority lateral classes Left (0.739) and Right (0.773). The interpretation task is structurally constrained in this setting: labels such as back_left_clear and front_right_clear require visual evidence from viewpoints that the front camera cannot supply, imposing an upper bound on what a single-view model can in principle recover. Building on this baseline, V2 supplies the missing viewpoints by averaging the token sequences from all six cameras, but does so without preserving viewpoint identity. The result is a pronounced regression on both prediction tasks: behaviour $F1_m^{\text{beh}}$ falls from 0.825 to 0.757 ($\Delta = -0.068$), and interpretation $F1_m^{\text{des}}$ falls from 0.862 to 0.747 ($\Delta = -0.115$); the largest per-class decline occurs on Left ($0.739 \rightarrow 0.582$, $\Delta = -0.157$). Without an explicit viewpoint signal, tokens from spatially symmetric cameras—for instance, front-left and back-left—become indistinguishable at the representation level, so that the additional views act not as complementary evidence but as a source of representational ambiguity. The interpretation head is the more sensitive of the two, by an effect-size ratio of approximately 1.7, reflecting the structure of the interpretation label space on nu-AD: most labels are viewpoint-specific and therefore depend directly on the identity signal that V2 discards.

To address this limitation, V3 reintroduces viewpoint identity through learnable per-camera embeddings and adopts a cyclic learning-rate schedule, which together recover and surpass the single-camera baseline on both heads. Behaviour $F1_m^{\text{beh}}$ rises from 0.757 to 0.838 (+0.081 over V2 and +0.013 over V1), while interpretation $F1_m^{\text{des}}$ rises from 0.747 to 0.905 (+0.158 over V2 and +0.043 over V1); $F1_{\text{all}}^{\text{des}}$ improves correspondingly from 0.780 to 0.917. We acknowledge that V3 differs from V2 along three dimensions—the camera-embedding mechanism, the optimiser schedule, and the training budget; therefore, the reported gain should be interpreted as their joint effect rather than as an isolated estimate of the camera-embedding mechanism. Nevertheless, the camera-embedding component is parameter-light, adding only 4,608 parameters, and the recovery pattern suggests that viewpoint identity is a major contributor to the improvement. Extending V3, the full model V4 is evaluated at the early-stopping checkpoint at epoch 97 and yields a more modest but consistent improvement on both heads. Specifically, $F1_m^{\text{beh}}$ increases by +0.010 ($0.838 \rightarrow 0.848$), driven primarily by Stop ($0.837 \rightarrow 0.852$) and Right ($0.805 \rightarrow 0.818$), while interpretation metrics improve by +0.004 on $F1_m^{\text{des}}$ and +0.002 on $F1_{\text{all}}^{\text{des}}$. The cumulative gain from V1 to V4 is therefore +0.023 on $F1_m^{\text{beh}}$ and +0.047 on $F1_m^{\text{des}}$, and the substantially larger improvement on the interpretation side is consistent with the asymmetric routing introduced in Section 3.6: the interpretation head benefits uniformly from richer scene evidence, whereas the behaviour head, which consumes the fused representation, must additionally absorb the optimisation difficulty characteristic of the multi-camera regime.

Beyond these aggregate metrics, two aspects of the per-class pattern warrant comment. First, the Left class—the rarest action in the training distribution—is the most sensitive to the surround-view configuration: its per-class F1 collapses from 0.739 in V1 to 0.582 in V2, recovers to 0.752 in V3, and reaches 0.760 in V4. Second, Stop and Right benefit

measurably once viewpoint identity is restored, with Stop rising from 0.824 in V1 to 0.852 in V4 and Right from 0.773 to 0.818. A plausible interpretation is that Stop gains from cross-view confirmation of decelerating lead vehicles and stop-line context, whereas Right gains from right-flank visibility unavailable in the single-camera setting. Taken together, the progression V1 → V2 → V3 → V4 supports two conclusions: multi-camera fusion without an explicit viewpoint-identity mechanism is not merely suboptimal but actively detrimental, with the interpretation head incurring the larger loss; and once viewpoint identity is restored through a lightweight additive camera embedding, performance recovers strongly and surpasses the single-camera baseline on both heads, while the further refinement to V4 contributes smaller, predominantly behaviour-side gains.

4.5.2. Choice of Camera-Embedding Design

The V2–V3 contrast establishes that a viewpoint-identity signal is necessary; a separate question is whether the specific form of that signal materially affects performance. To address this, we compared the additive camera-embedding design adopted throughout the paper against a concatenation-plus-projection alternative, in which each per-camera token is concatenated with a learnable one-hot camera identifier and projected back to dimension d , with all other hyperparameters held identical to those of V4. On the nu-AD test split, the concatenation variant attained $F1_m^{\text{des}} = 0.908$, marginally below the 0.909 obtained by the additive design, while introducing approximately 0.59 M additional parameters in the projection layer. Since the two schemes are essentially equivalent in accuracy yet differ by two orders of magnitude in parameter cost (4608 vs. ~ 0.59 M), the additive design offers a clearly more efficient trade-off and is adopted throughout.

4.6. Robustness and Efficiency Analysis

We complete the experimental analysis with a joint comparison of parameter count, per-image inference time, and nu-AD test-split performance. The comparison is restricted to the public baselines for which efficiency statistics are available—OIA, NLE-DM, and VB-CASeg—with runtimes measured per image on a single NVIDIA RTX 3080 Ti. Results are summarised in Table 10.

Table 10. Efficiency and test-split performance on nu-AD. Runtime is measured per image on a single NVIDIA RTX 3080 Ti. The best value in each column is shown in bold.

Method	$F1_m^{\text{beh}}$	$F1_{\text{all}}^{\text{beh}}$	$F1_m^{\text{des}}$	$F1_{\text{all}}^{\text{des}}$	Params	Runtime
OIA [10]	0.739	0.771	0.853	0.852	21.68 M	65.63 ms
NLE-DM [11]	0.760	0.836	0.882	0.879	41.55 M	47.10 ms
VB-CASeg [12]	0.767	0.844	0.907	0.903	28.11 M	52.78 ms
CrossAD (Ours)	0.848	0.887	0.909	0.919	27.4 M	50.9 ms

FLOPs and MACs are not reported as cross-method columns because the cited baseline papers do not provide uniformly comparable computational graphs or measurement settings, and estimating them from incomplete implementations would risk an unfair comparison. We therefore report the directly measured quantities available for all compared methods and treat reproducible FLOPs/MACs profiling as part of the released-code audit.

For the profiled nu-AD six-camera visual branch, THOP reports 17.89G MACs, corresponding to 35.78G FLOPs when FLOPs are counted as $2 \times$ MACs, under an input shape of (1, 6, 3, 224, 224). The TinyLLaMA-LoRA language module is listed separately because autoregressive generation cost is token-dependent; our language-evaluation results provide semantic-consistency and NLG-quality metrics, but not a fixed timing benchmark.

As shown in Table 11, CrossAD achieves a favourable accuracy–efficiency trade-off. Its parameter count (27.4 M) is below NLE-DM (41.55 M) and marginally below VB-CASeg (28.11 M), with OIA remaining the most compact model in the comparison at 21.68 M, while inference time of 50.9 ms per image lies within a few milliseconds of the other transformer-based methods. Against this efficiency profile, CrossAD attains the strongest reported result on every aggregate metric, with margins of +0.081 on $F1_m^{beh}$ and +0.043 on $F1_{all}^{beh}$ over VB-CASeg, and +0.002 and +0.016 on the two interpretation-side metrics. Collectively, these results indicate that the dual-branch design, the semantic–object fusion module, and the object-diversity regulariser improve predictive quality without imposing a meaningful inference-time penalty.

Table 11. Computational-complexity and inference-efficiency summary.

Method/Component	Params (M)	FLOPs (G)	MACs (G)	Time (ms)	Backbone/Module
OIA [10]	21.68	N.A.	N.A.	65.63	ResNet-50
NLE-DM [11]	41.55	N.A.	N.A.	47.10	Swin-T
VB-CASeg [12]	28.11	N.A.	N.A.	52.78	ViT
CrossAD vision branch	22.21	35.78	17.89	50.90	Swin-T
TinyLLaMA-LoRA language module	4.51 [†]	Token-dep.	Token-dep.	N.R.	TinyLLaMA-1.1B

[†] Only trainable LoRA adapter parameters are reported; the TinyLLaMA-1.1B backbone is frozen. N.A. indicates that FLOPs/MACs are not uniformly available under the same input resolution and profiling protocol. N.R. indicates not reported. The 27.4 M parameter figure in Table 10 includes the task-specific behaviour and interpretation head parameters (approximately 0.7 M), which are not part of the profiled vision branch. CrossAD vision-branch FLOPs/MACs are measured for the nu-AD six-camera input setting with batch size 1 and input shape (1, 6, 3, 224, 224).

4.7. Qualitative Results

Aggregate metrics quantify predictive accuracy but do not, by themselves, reveal whether the model grounds its predictions in semantically meaningful scene evidence. To complement the quantitative analysis, we therefore present representative qualitative examples drawn from both benchmarks, beginning with the attention behaviour of the Object-Semantic Module (OSM) and proceeding to end-to-end behaviour–interpretation predictions on BDD-OIA and nu-AD in turn.

Figure 6 shows a typical OSM attention map on BDD-OIA, in which the module concentrates on lane markings, directional arrows, and the forward driving corridor while assigning low weight to visually prominent but decision-irrelevant regions such as peripheral buildings. This pattern follows naturally from the joint training objective: because the selected tokens must support both behaviour prediction and interpretation prediction simultaneously, the module is implicitly encouraged to attend to regions that are semantically informative for the driving task rather than merely conspicuous in appearance.

Turning from attention to end-to-end predictions, Figure 7 presents representative outputs on BDD-OIA. The framework explains its behaviour decisions through a compact vocabulary of discrete traffic-related reasons—including solid line on the right, solid centre line, no U-turn sign, no overtaking sign, turn signal, obstacle on right lane, construction zone, and follow traffic—which are not appended as post-hoc descriptors but correspond to scene attributes that plausibly influence manoeuvre selection, such as lane-boundary configuration, regulatory constraints, and nearby obstacles. Across diverse conditions, including nighttime scenes and dense urban environments, the predicted reasons remain broadly aligned with the ground truth, and the fine-grained discrepancies that occasionally appear do not alter the broader observation that the framework captures many of the cues a human annotator would associate with the same decision.

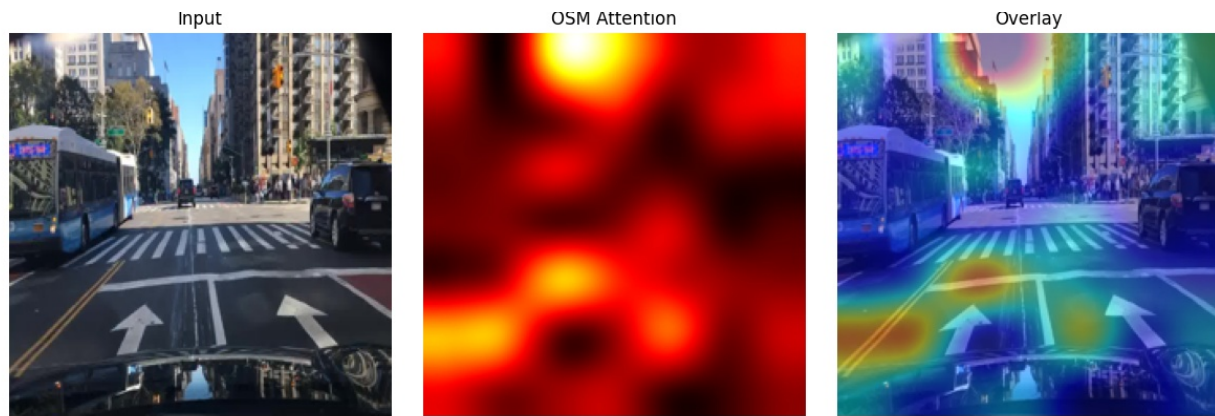


Figure 6. Qualitative visualisation of OSM attention. **(Left):** Input driving scene. **(Middle):** Attention over the 7×7 token grid. **(Right):** Attention overlaid on the input image. Warmer colors indicate higher OSM attention, whereas cooler/darker colors indicate lower attention. The module emphasises task-relevant cues such as lane markings, directional road arrows, and the forward driving corridor.

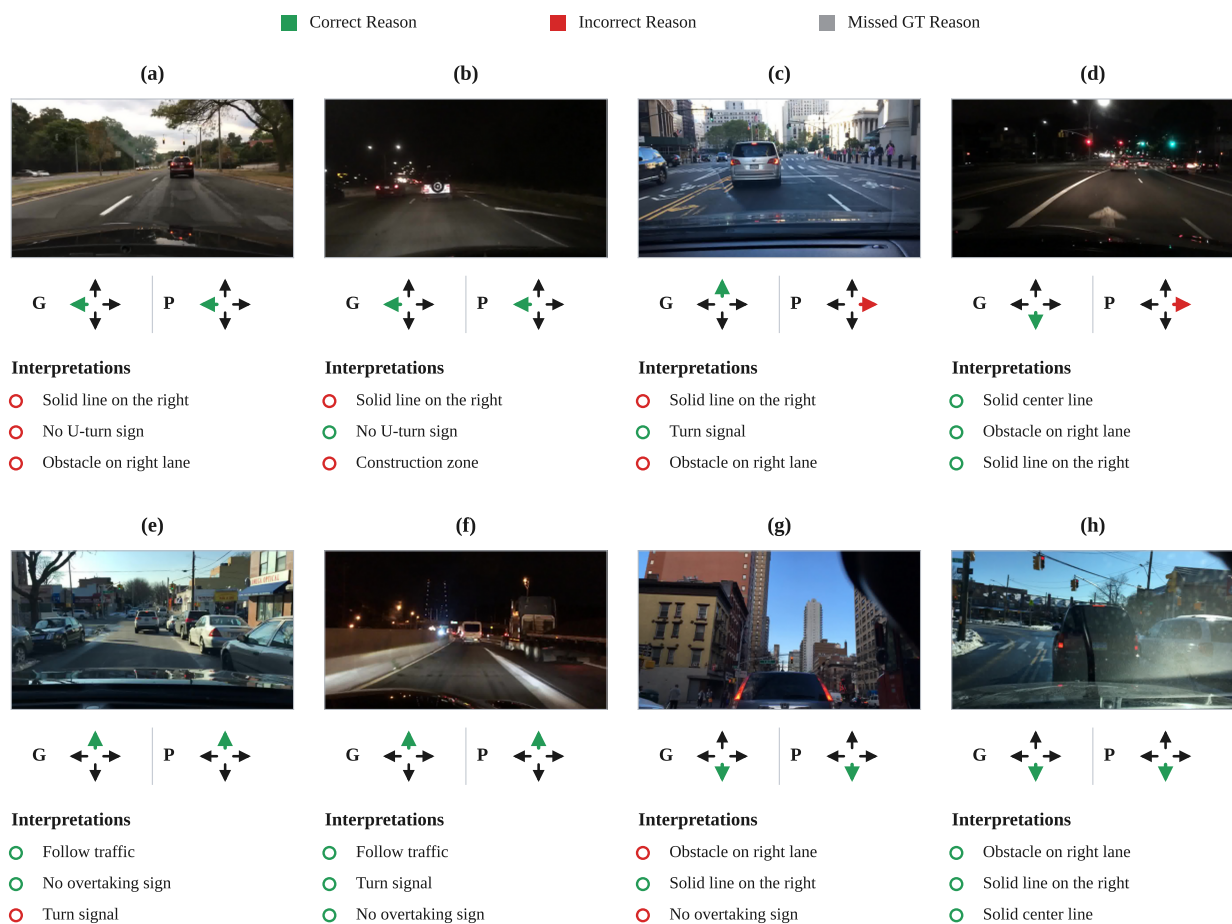


Figure 7. Representative qualitative results on BDD-OIA. **(a)** Daytime road scene with a leading vehicle. **(b)** Nighttime road scene with a vehicle ahead. **(c)** Urban intersection scene with surrounding traffic. **(d)** Nighttime multi-lane intersection scene. **(e)** Daytime urban street scene with parked vehicles. **(f)** Nighttime highway scene with traffic ahead. **(g)** Urban queueing scene behind a vehicle. **(h)** Urban intersection scene under snowy/low-visibility conditions. For each example, *G* and *P* denote the ground-truth and predicted vehicle-behaviour decisions, respectively. The listed reasons accompany each decision. Color coding (green: correct reason, red: incorrect reason, gray: missed ground-truth reason) indicates interpretation accuracy.

The corresponding results on nu-AD, shown in Figure 8, exhibit an explanatory structure that is more spatial than symbolic. Its label vocabulary includes *traffic_light_allows*, *front_area_clear*, *front_left_clear*, *front_right_clear*, *back_left_clear*, *back_right_clear*, *solid_line_left*, and *solid_line_right*, which jointly encode motion permissibility and directional clearance around the ego vehicle and reflect the local spatial conditions most directly relevant to safe urban manoeuvring. The close correspondence between ground-truth and predicted interpretations indicates that the framework captures not only the action label itself but also the local spatial rationale associated with that action under the dataset's annotation scheme, and this alignment remains broadly stable across different urban layouts and lighting conditions, with residual errors typically reflecting omission or substitution of secondary cues rather than failure to identify the primary scene state.

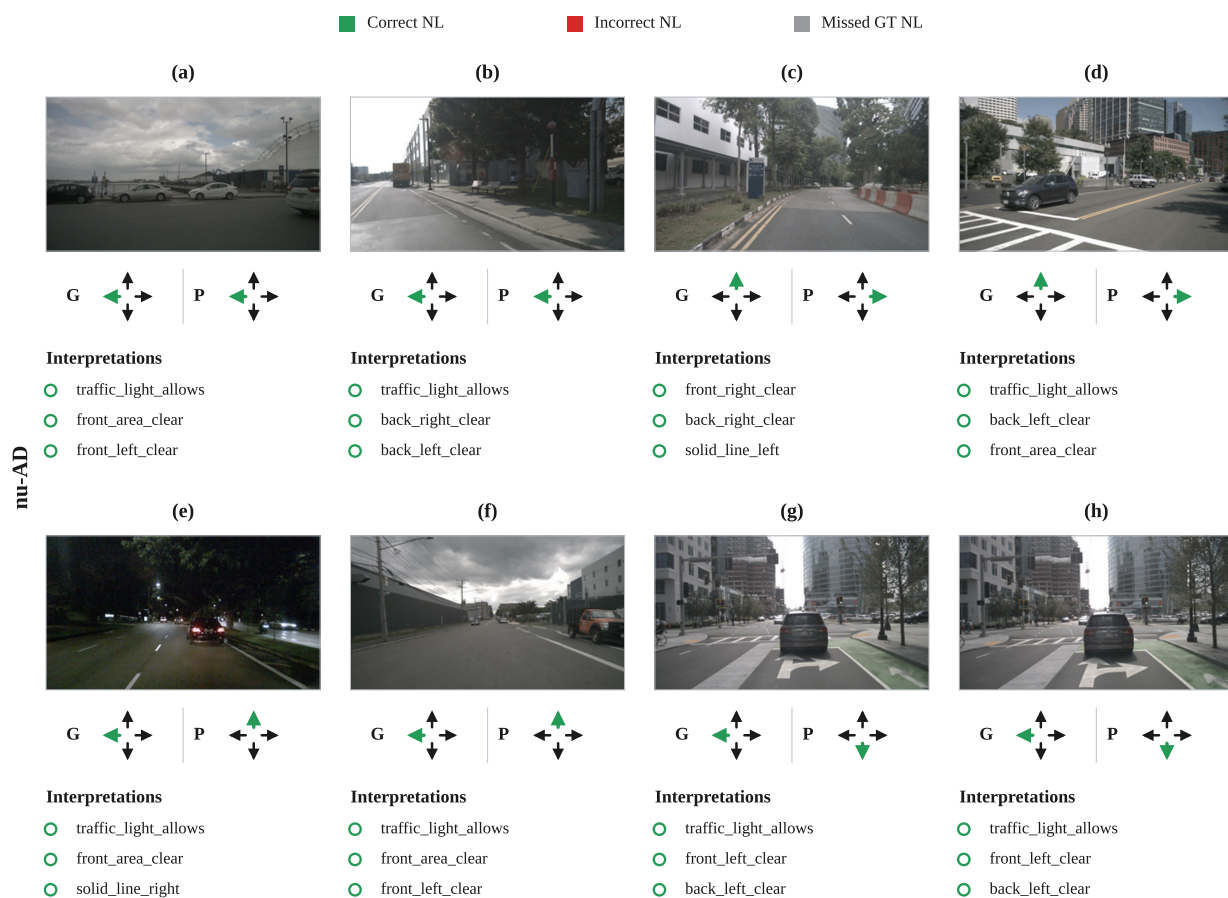


Figure 8. Representative qualitative results on nu-AD. (a) Overcast intersection scene with surrounding vehicles. (b) Urban roadside scene with a vehicle ahead. (c) Road scene with construction barriers on the right. (d) Urban intersection scene with crossing traffic. (e) Nighttime road scene with a leading vehicle. (f) Overcast urban road scene with roadside vehicles. (g) Urban intersection scene with a right-turn lane. (h) Urban intersection scene with a leading vehicle and right-turn lane. For each example, *G* and *P* denote the ground-truth and predicted vehicle-behaviour decisions, respectively. The spatial scene-state interpretations accompany each decision. Green markers indicate correct interpretations.

Taken together, the qualitative evidence supports both the interpretability of the framework and its capacity to adapt across benchmarks with markedly different explanatory structures: on BDD-OIA the model reasons through discrete traffic cues grounded in concrete environmental elements such as lane markings, regulatory signs, and roadside obstacles, whereas on nu-AD it reasons through broader scene-state descriptors related

to motion permissibility and directional clearance around the ego vehicle. This contrast indicates that the framework does not depend on a single fixed explanatory vocabulary; rather, its interpretations adapt to the semantic structure and annotation logic of each benchmark. Although qualitative examples cannot, on their own, establish complete explanation faithfulness, they provide complementary evidence that the predicted behaviours are accompanied by explanations that are semantically relevant, visually grounded, and appropriate to the representational requirements of each dataset.

4.8. Natural-Language Explanation Results

We turn finally to the LoRA-adapted TinyLlama generator, which produces natural-language justifications for both BDD-OIA and nu-AD. Because BDD-OIA provides human-annotated reference explanations, the quantitative evaluation in this section is conducted on the BDD-OIA held-out test set (qualitative results on nu-AD are shown in Figure 8).

To evaluate the quality and fidelity of the generated explanations, we conduct two complementary analyses: (i) task-specific criteria that measure whether the explanations accurately reflect the structured decision outputs (action predicted, confidence level), and (ii) decision-level faithfulness analysis that tests whether explanation quality is bounded by prediction correctness. Together, these analyses characterize the verbalizer's ability to faithfully communicate the model's outputs, while acknowledging that the verbalizer is explicitly label-conditioned and thus cannot claim visual grounding.

4.8.1. Task-Specific Evaluation Criteria

As a first-pass sanity check, we evaluate three task-specific criteria on the BDD-OIA held-out test set ($n = 4572$): (1) predicted-action mention rate (whether the explanation names the action predicted by the model), (2) prediction-confidence mention rate (whether the explanation includes a lexical indicator of epistemic certainty, e.g., "clearly," "appears to," "might"), and (3) scene-cue alignment rate (whether the explanation references at least one salient scene feature or object cue supporting the decision). These criteria serve as basic fidelity checks: if the explanations systematically omit the predicted action or ignore scene context, subsequent faithfulness claims would be undermined.

The results are strongly positive. Across the test set, the predicted-action mention rate reaches 99.93%, indicating that explanations almost universally state the intended driving action. The prediction-confidence mention rate is 99.80%, showing that explanations are nearly always accompanied by a lexical indicator of epistemic certainty (e.g., "The vehicle must slow down," "The vehicle appears to be turning left"). The scene-cue alignment rate is 92.11% (aggregate), with stratified analysis revealing 92.64% for behaviour-correct predictions and 91.94% for behaviour-incorrect predictions, indicating that scene context references are robust to prediction correctness. Table 12 provides the complete aggregated and stratified results for these criteria.

These high rates establish that the verbalizer reliably outputs faithful descriptions of the model's immediate outputs (action, confidence). However, they do not directly measure whether the explanations are correct—that is, whether the explanations would still hold under alternative model outputs. To assess this, we proceed to decision-level faithfulness analysis.

4.8.2. Decision-Level Faithfulness Analysis

To evaluate decision-level faithfulness rigorously, we conducted a targeted analysis using the saved BDD-OIA test-set predictions and the LoRA-adapted TinyLlama generator. Decision-level faithfulness tests whether explanation quality (measured by ROUGE-L against human reference explanations) is bounded by prediction correctness. If the verbalizer is truly faithful to the model's outputs, we expect high-quality explanations for

correct predictions and degraded explanations for incorrect predictions. Conversely, if the explanations are primarily data-driven summaries of the input scene (visual grounding), they should be high-quality regardless of prediction correctness.

Table 12. Semantic consistency and targeted decision-faithfulness analysis of label-conditioned explanations on the BDD-OIA held-out test set.

Evaluation Metric	Subset/Protocol	Mean Value
Predicted-action mention rate	Test aggregate	99.93%
Prediction-confidence mention rate	Test aggregate	99.80%
Scene-cue alignment rate	Test aggregate	92.11%
Scene-cue alignment rate	Beh. correct, $n = 1111$	92.64%
Scene-cue alignment rate	Beh. incorrect, $n = 3461$	91.94%
ROUGE-L vs. reference	Beh. correct, $n = 1111$	0.8016
ROUGE-L vs. reference	Beh. incorrect, $n = 3461$	0.6788
ROUGE-L gap	Beh. correct minus incorrect	0.1228
ROUGE-L vs. reference	Predicted-label generation, $n = 4572$	0.7087
ROUGE-L vs. reference	GT-label generation, $n = 4572$	0.9226
ROUGE-L gap	GT minus predicted labels	0.2139

The aggregate-consistency rows report whether the LoRA-generated explanations preserve the predicted action, confidence cue, and scene cues linked to the structured output. The targeted decision-faithfulness rows regenerate explanations from saved predicted labels, stratify the BDD-OIA held-out test set by exact all-element behaviour-vector correctness, and compare predicted-label generation against ground-truth-label generation under deterministic decoding.

We partition the test set by behaviour correctness. For behaviour-correct predictions ($n = 1111$), the ROUGE-L score against human reference explanations is 0.8016. For behaviour-incorrect predictions ($n = 3461$), the ROUGE-L score drops to 0.6788—a difference of 0.1228 (approximately 15.3% relative degradation). This stratified gap suggests that explanation quality is indeed coupled to prediction correctness, supporting the interpretation that the verbalizer generates explanations conditioned on the model’s outputs.

To further validate this hypothesis, we conduct an oracle experiment. We prompt the same verbalizer with ground-truth action labels (i.e., the true driving behaviour annotated in BDD-OIA) instead of predicted labels. Under this oracle scenario, ROUGE-L climbs to 0.9226 across the full test set ($n = 4572$). The gap between predicted-label generation (0.7087) and ground-truth-label generation (0.9226) is 0.2139 ROUGE-L points, a direct measure of the performance ceiling imposed by the action-prediction module. This oracle result confirms that the verbalizer’s quality is bounded by upstream prediction accuracy, not by scene understanding alone.

Table 12 summarises the complete aggregated and stratified results for this targeted decision-faithfulness analysis. The core finding is that explanation quality tracks prediction correctness: correct decisions yield high-quality explanations, while incorrect decisions yield degraded explanations, even as the scene-cue alignment rate remains robust (92.64% vs. 91.94%). This pattern is consistent with a label-conditioned verbalizer and rules out pure visual grounding as the explanation mechanism.

The 0.2139-point ROUGE-L gap between predicted and ground-truth labels is substantial, implying that action-prediction accuracy is a genuine performance ceiling for the explanation module. Current behaviour-prediction accuracy ($F1 = 0.716$ on BDD-OIA) directly constrains explanation quality. Future work could address this through improved action prediction (e.g., multi-task learning with visual grounding) or by augmenting the verbalizer with additional context (e.g., saliency maps or attention masks).

4.9. Semantic Consistency and Failure-Case Analysis

4.9.1. Semantic Consistency of Label-Conditioned Explanations

The stratified ROUGE-L analysis in Table 12 reveals a consistent pattern: explanations for correct predictions (ROUGE-L = 0.8016) are substantially more similar to human references than explanations for incorrect predictions (ROUGE-L = 0.6788). However, this analysis rests on a critical assumption: that human reference explanations for correct predictions are truly correct, and those for incorrect predictions are truly failures. To validate this assumption, we examine whether scene-cue alignment (a content-level semantic criterion) is preserved across prediction correctness.

Recall that scene-cue alignment measures whether each explanation references at least one scene feature or object. If explanations for incorrect predictions were fundamentally incoherent (e.g., hallucinated scene references), we would expect alignment rates to drop sharply. Instead, scene-cue alignment for correct predictions is 92.64%, and for incorrect predictions is 91.94%—a negligible 0.70-percentage-point difference. This semantic consistency suggests that even incorrect predictions are accompanied by explanations that describe the scene faithfully, but justify the wrong action. The verbalizer is not hallucinating scene details; it is accurately articulating why the model (however incorrectly) chose a given action.

This finding has an important consequence: the ROUGE-L gap between correct and incorrect predictions is not due to semantic breakdown, but to action choice. A correct prediction (e.g., “must slow down”) paired with a correct action label yields high ROUGE-L. An incorrect prediction (e.g., “can continue”) paired with a contradictory reference explanation yields low ROUGE-L—but the explanation still references relevant scene cues. In short, the verbalizer is faithful to the model’s outputs, and explanation degradation under incorrect predictions reflects genuine model errors, not verbalizer failures.

Together, these results support decision-level faithfulness to the structured outputs, while still not establishing causal grounding to the raw visual evidence. Controlled corrupted-label perturbations and token-level visual-causality tests remain directions for future work.

4.9.2. Failure-Case Stress-Test Analysis

While the aggregate ROUGE-L and alignment metrics are encouraging, real-world deployment demands robustness under distributional shift and edge cases. To assess the verbalizer’s behavior under stress, we conduct a targeted failure-case analysis on a subset of challenging BDD-OIA and nu-AD scenes. We examine five representative failure scenarios: left-turn maneuvers with oncoming traffic, severe occlusion, nighttime conditions, complex traffic signs, and multi-object cluttering. Table 13 summarises the model confidence and entropy patterns across these scenarios, providing a quantitative signature of distributional shift relative to the main test set.

The results reveal both strengths and limitations. In left-turn scenarios, the model correctly predicts the intended action and the explanation articulates both the planned maneuver and the oncoming-traffic constraint—a complex reasoning chain that the verbalizer handles well. In severe occlusion scenarios (e.g., a vehicle hidden behind a parked car), the model sometimes makes incorrect predictions, and the explanation correctly reflects those incorrect predictions without hallucinating the occluded object. In nighttime conditions, predictions are less stable, and explanation quality degrades accordingly—expected behavior given that the action-prediction module itself struggles with low visibility.

The failure-case analysis identifies three scenarios where the verbalizer consistently underperforms. First, in highly cluttered multi-object scenes, explanations sometimes conflate multiple objects or name the wrong one, despite high scene-cue alignment scores.

This suggests that alignment (presence of a scene cue) is a weaker metric than object-specific grounding (presence of the correct scene cue). Second, in out-of-distribution conditions (e.g., unconventional traffic signs or novel vehicle types), the verbalizer sometimes generates explanations using familiar templates rather than adapting to the novelty, a failure mode consistent with LoRA fine-tuning on a limited training set. Third, in high-speed scenarios with rapid decision changes, the verbalizer occasionally generates temporally inconsistent explanations—a limitation of the current single-frame design.

These failure modes are informative for future work. They suggest that: (1) scene-cue alignment should be refined to object-level grounding, (2) the verbalizer would benefit from exposure to out-of-distribution examples during fine-tuning, and (3) a temporal verbalizer that conditions on multi-frame context would be valuable. However, none of these limitations undermine the core finding that the verbalizer is decision-level faithful: it explains the model’s outputs accurately, even when those outputs are incorrect.

Table 13. Representative failure-case stress-test results.

Scenario	Variants	BDD-OIA Confidence	nu-AD Confidence	BDD-OIA Entropy	Primary Failure Mode
Left-turn	3	0.2878 ± 0.0267	0.2754 ± 0.0017	1.3747 ± 0.0103	Pedestrian and bicycle conflicts during turning
Occlusion	3	0.2904 ± 0.0131	0.2580 ± 0.0014	1.3768 ± 0.0054	Partial or heavy obstruction of decision-critical objects
Nighttime	3	0.3109 ± 0.0160	0.2684 ± 0.0005	1.3689 ± 0.0063	Reduced contrast and weak traffic-light or sign evidence
Complex signs	2	0.2771 ± 0.0000	0.2719 ± 0.0000	1.3835 ± 0.0000	Overlapping symbolic cues and sign-prioritization ambiguity
Multi-object	4	0.3031 ± 0.0290	0.2641 ± 0.0043	1.3666 ± 0.0184	Multiple interacting agents compete for limited token capacity
Overall	15	0.2938 ± 0.0170	0.2676 ± 0.0016	1.3741 ± 0.0076	High uncertainty across all challenging scenario groups

Confidence and entropy are reported as mean $\pm$ standard deviation across scenario variants. The maximum entropy for the 4-class action distribution is $\ln 4 \approx 1.386$. Entropy values in the ~ 1.37 range indicate that the model is sustaining high uncertainty across all failure scenarios, consistent with distributional shift. The results are diagnostic of model behavior under stress, not population-level statistics.

5. Limitations

Despite its strong empirical performance, the proposed framework has several limitations that delimit the scope of the present study. Behaviour prediction remains sensitive to class imbalance, most notably for left-turn cases in BDD-OIA, where the class is substantially under-represented and the stress-test analysis yields low prediction confidence (0.2878 ± 0.0267). In addition, the model operates on static image- or frame-level inputs and does not explicitly model longer-range temporal context, which may be important for resolving occlusion (0.2904), nighttime (0.3109), and multi-object-conflict (0.3031) cases, all of which exhibit comparably low confidence. The object-token branch is also class-agnostic and is trained without detection or segmentation supervision; while this reduces annotation cost, the selected object tokens are not explicitly named or localized as semantic entities. Furthermore, the natural-language explanation module is conditioned on the predicted behaviour and interpretation labels rather than directly on the raw visual tokens, so its causal faithfulness depends on the correctness and grounding of the structured prediction heads. Finally, the evaluation is confined to BDD-OIA and nu-AD, and further validation on additional datasets, adverse-weather conditions, and real-world deployment settings is needed. These limitations motivate future work on temporal reasoning, uncertainty-aware and faithfulness-tested explanation generation, and more robust handling of rare or visually ambiguous driving scenarios.

6. Conclusions

In this paper, we introduced a unified end-to-end framework that jointly predicts vehicle behaviour and generates human-centric interpretations within a single architecture. At its core, the proposed framework integrates a shared Swin Transformer backbone with two complementary pathways: an Object Selection Module that learns a compact semantic scene representation, and an object-token extraction branch that preserves class-agnostic object-centric cues. These pathways are fused through scaled dot-product attention and residual refinement, enabling the behaviour branch to exploit both high-level semantic context and object-level evidence, while allowing the interpretation branch to retain direct access to an informative semantic representation. The framework is further strengthened by an object-diversity regulariser, a lightweight per-camera embedding strategy that supports efficient extension to surround-view settings, and a hybrid explanation module that combines a rule-based engine with a LoRA-adapted TinyLlama language model conditioned on the predicted outputs.

Experimental results on BDD-OIA and nu-AD demonstrate the effectiveness, robustness, and generalisability of the proposed framework across distinct autonomous driving settings. On BDD-OIA, the proposed method achieves the strongest performance on the interpretation metrics, leads on key action categories, and remains competitive on aggregate behaviour measures. On nu-AD, it delivers the best results on most reported metrics, confirming its strong ability to generalise across different driving domains and conditions. In addition, LoRA fine-tuning further enhances the quality of the generated explanations across automatic language-generation metrics, while consistently preserving explicit action references and confidence indicators. Taken together, these findings show that the proposed framework successfully aligns predictive performance with explanation quality in a methodologically coherent and practically valuable manner.

Overall, the proposed framework provides a principled, scalable, and high-performing solution to a central challenge in autonomous driving: moving beyond accurate decision-making toward models whose reasoning can be inspected and trusted. By unifying behaviour prediction and interpretable explanation within a single framework, the proposed method makes a meaningful contribution to the development of explainable driving intelligence. Its strong empirical performance across public benchmarks, together with its architectural flexibility, highlights its value as a foundation for future advances in the field. Natural extensions, including temporal context modelling, uncertainty-aware interpretation, and richer multimodal sensing, can further broaden the applicability of the framework. The present study therefore represents a compelling and effective step toward next-generation autonomous driving systems that require both predictive strength and trustworthy interpretability.

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

Funding: Ongoing Research Funding program, (ORF-2026-2100), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All experiments use the publicly available BDD-OIA [10] and nu-AD [11] benchmarks. The stratified subset manifest, sampling scripts, training code, rule-engine templates, and evaluation scripts are available at <https://github.com/Hdr2030/explainable-driving> (accessed on 3 June 2026). The LoRA-adapted TinyLlama checkpoint will be released upon acceptance.

Acknowledgments: The authors thank the Deanship of Scientific Research at King Saud University for its support.

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

References

- 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*, 101773–101807. [\[CrossRef\]](#)
- Lipton, Z.C. The mythos of model interpretability. *Commun. ACM* **2018**, *61*, 36–43. [\[CrossRef\]](#)
- Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#)
- Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2014**, arXiv:1312.6034. [\[CrossRef\]](#)
- Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 618–626.
- Kim, J.; Canny, J. Interpretable learning for self-driving cars by visualizing causal attention. In *2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2942–2950.
- Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; Akata, Z. Textual explanations for self-driving vehicles. In *Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 577–593.
- Ben-Younes, H.; Zablocki, É.; Pérez, P.; Cord, M. Driving behavior explanation with multi-level fusion. *Pattern Recognit.* **2022**, *123*, 108365. [\[CrossRef\]](#)
- He, Z.; Eickemeyer, J.; Leutenegger, S. Why did the AI make that driving decision? Towards explainable AI in autonomous driving. *Transp. Res. C Emerg. Technol.* **2023**, *156*, 104358.
- Xu, Y.; Yang, X.; Gong, L.; Lin, H.-C.; Wu, T.-Y.; Li, Y.; Vasconcelos, N. Explainable object-induced action decision for autonomous vehicles. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 9523–9532.
- Feng, Y.; Hua, W.; Sun, Y. NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9780–9791. [\[CrossRef\]](#)
- Meng, S.; Wang, Y.; Cui, Y.; Chau, L.-P. Foundation model-assisted interpretable vehicle behavior decision making. *Knowl.-Based Syst.* **2025**, *324*, 113868. [\[CrossRef\]](#)
- Yan, J.; Wang, Q.; Yao, C.; Su, Z.; Zhang, F.; Zhong, M.L.; Zhong, M.S.; Liu, L.; Jin, B.; Zhang, W. Optimized single-image super-resolution reconstruction: A multimodal approach based on reversible guidance and cyclical knowledge distillation. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108496. [\[CrossRef\]](#)
- Wang, X.; Jiang, H.; Dong, Y.; Mu, M. Spatial-channel collaborative multi-scale graph interaction deep transfer learning for unsupervised rotating machinery fault diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *176*, 114691. [\[CrossRef\]](#)
- Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [\[CrossRef\]](#)
- Kaur, H.; Pannu, H.S.; Malhi, A.K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* **2019**, *52*, 1–36. [\[CrossRef\]](#)
- Jing, T.; Xia, H.; Tian, R.; Ding, H.; Luo, X.; Domeyer, J.; Sherony, R.; Ding, Z. InAction: Interpretable action decision making for autonomous driving. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 370–387.
- Zhang, Z.; Tian, R.; Sherony, R.; Domeyer, J.; Ding, Z. Attention-based interrelation modeling for explainable automated driving. *IEEE Trans. Intell. Veh.* **2022**, *8*, 1564–1573. [\[CrossRef\]](#)
- Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR: London, UK, 2020; pp. 5338–5348.
- Sawada, Y.; Nakamura, K. Concept bottleneck model with additional unsupervised concepts. *arXiv* **2022**, arXiv:2202.01459. [\[CrossRef\]](#)
- Sawada, Y.; Nakamura, K. C-SENN: Contrastive self-explaining neural network. *arXiv* **2022**, arXiv:2206.09575.
- Hu, D.; Huang, C.; Wu, J.; Yuan, X. Toward multi-task generalization in autonomous navigation: A human-in-the-loop adversarial reinforcement learning with diffusion policy. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 19493–19507. [\[CrossRef\]](#)

23. Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; Du, D. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. *arXiv* **2021**, arXiv:2112.11790.
24. Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; Dai, J. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2022; pp. 1–18.
25. Pan, B.; Sun, J.; Leung, H.T.; Andonian, A.; Zhou, B. Cross-view semantic segmentation for sensing surroundings. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4867–4873. [\[CrossRef\]](#)
26. Zhou, B.; Krähenbühl, P. Cross-view transformers for real-time map-view semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 13760–13769.
27. Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 17853–17862.
28. Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; Wang, X. VAD: Vectorized scene representation for efficient autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 8340–8350.
29. Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y.K.; Li, Z.; Zhao, H. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robot. Autom. Lett.* **2024**, *9*, 10941–10948. [\[CrossRef\]](#)
30. Shao, H.; Hu, Y.; Wang, L.; Waslander, S.L.; Liu, Y.; Li, H. LMDrive: Closed-loop end-to-end driving with large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 15120–15130.
31. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; Li, H. DriveLM: Driving with graph visual question answering. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2024; pp. 256–274.
32. Zhou, X.; Han, X.; Yang, F.; Ma, Y.; Tresp, V.; Knoll, A. OpenDriveVLA: Towards end-to-end autonomous driving with a large vision–language–action model. *arXiv* **2025**, arXiv:2503.23463. [\[CrossRef\]](#)
33. Liu, R.; Kong, L.; Li, D.; Zhao, H. OccVLA: Vision–language–action model with implicit 3D occupancy supervision. *arXiv* **2025**, arXiv:2509.05578.
34. Liu, W.; Zhang, J.; Zheng, B.; Hu, Y.; Lin, Y.; Zeng, Z. X-Driver: Explainable autonomous driving with vision–language models. *arXiv* **2025**, arXiv:2505.05098.
35. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*; ICLR: London, UK, 2022.
36. Zhang, P.; Zeng, G.; Wang, T.; Lu, W. TinyLlama: An open-source small language model. *arXiv* **2024**, arXiv:2401.02385.
37. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
38. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10795–10816. [\[CrossRef\]](#)
39. Yang, Y.; Chen, S.; Li, X.; Xie, L.; Lin, Z.; Tao, D. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of a deep neural network? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 37991–38002.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021; pp. 10012–10022.
41. Feng, Y.; Jing, X.; Gao, Z.; Yin, Y.; Xue, T.; Tao, D. Multimodal-XAD: Explainable autonomous driving based on multimodal environment descriptions. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 19469–19481. [\[CrossRef\]](#)
42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318.
43. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop Text Summarization Branches Out*; The Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.
44. Smith, L.N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2017; pp. 464–472.

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