From 2D attention maps to 3D visualizations in deep driving models

Josep Bonet i Sáez

June 28, 2026

Abstract

The adoption of end-to-end deep learning models in autonomous driving has led to significant performance gains, but their "black-box" nature hinders safety verification, debugging and user trust. This project addresses that challenge by developing a novel explainability system that visualises the driving model's own attention in 3D. Recent advances in vision-language models (VLMs) have enabled the prediction of not only where a human driver would look, but also the semantic reasoning behind that attention. We repurpose such a VLM, fine-tuning it on attention maps extracted from a state-of-the-art end-to-end driving model, so that it produces semantic explanations aligned with the model's actual focus, not a human's gaze. The core contribution lies in projecting these model-derived attention maps into a coherent 3D representation of the driving scene and enriching that 3D view with the VLM's textual annotations. The final output will allow end-users to observe, in full spatial context, which elements of the environment (e.g. pedestrians, traffic signs, road boundaries) the model is attending to, along with human-readable descriptions of what those elements are and why they are important. By transforming abstract 2D heatmaps and text into spatially grounded 3D explanations, this project aims to significantly enhance the interpretability of autonomous driving agents, ultimately contributing to safer and more trustworthy AI systems for the road.

Keywords: Driver attention prediction, Explainable AI (XAI), Autonomous Driving, 3D Visualization, Deep Learning, Large Language Models (LLMs), Computer Vision, End-to-End Driving, Human-Centered AI.

1 Introduction - Context of the Project

1.1 Context and Motivation

The rapid advancement of deep learning has propelled autonomous driving from a futuristic concept to an imminent reality. End-to-end (E2E) driving models, which directly map raw sensor inputs to driving actions, have demonstrated remarkable performance in complex driving tasks [1, 2]. By learning driving policies directly from data, these models avoid the complexities and potential error propagation of traditional modular pipelines [3]. A diagram can be seen in Fig. 1.

However, this performance comes at a significant cost: opacity. The intricate neural networks that power E2E systems function as "black boxes", making it exceedingly difficult for engineers, regulators, and end-users to understand why a vehicle makes a particular decision, such as breaking suddenly or executing a lane change. In contrast, modular systems decompose the driving task into separate stages (e.g. perception, localization, planning and control), each with well-defined inputs and outputs. While the individual modules may themselves be deep learning models and thus partially opaque, this decomposition allows engineers to isolate and inspect each component's behaviour independently. As a result, if the vehicle brakes unexpectedly, engineers can trace the issue back to a specific module by checking its outputs, something much harder to do with a monolithic E2E model. Additionally, E2E models are not typically trained on explicit object-detection objectives like recognising pedestrians or stop signs; instead, their attention to such elements emerges purely from the driving task itself. This means a model may focus on a traffic light without ever having been taught the concept of a traffic light, further complicating the interpretation of its internal attention.

This lack of interpretability in E2E systems poses a critical barrier to the widespread deployment of autonomous vehicles. Without insight into a model's reasoning, validating its safety, debugging its failures, and building public confidence become huge challenges. This has brought up

• Contact E-mail: josep.bonet@autonoma.cat
• Supervised by: Antonio Manuel Lopez Peña (Departament de Ciències de la Computació)
• Academic Year 2025/26

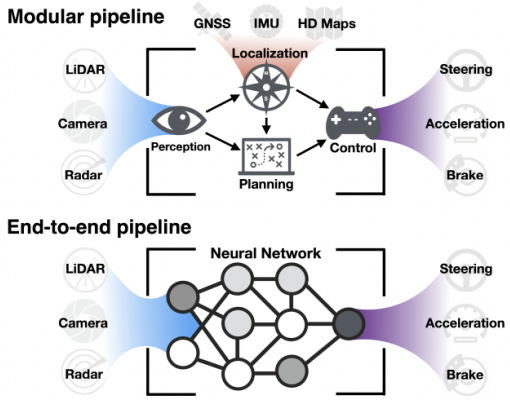

Fig. 1: Comparison of modular and end-to-end pipelines for autonomous driving. The modular pipeline consists of many interconnected modules (perception, prediction, planning, control), whereas the end-to-end approach treats the entire pipeline as one learnable machine learning task that maps raw sensor inputs directly to driving actions. Adapted from [1].

the emergence of Explainable Artificial Intelligence (XAI), a field dedicated to making AI decision-making transparent and understandable to humans [4].

1.2 Problem Definition: The limits of 2D Explanations

In the context of autonomous driving, a prominent XAI approach involves visualizing the internal ”attention” of a model. Attention mechanisms, originally designed to improve model performance, can be repurposed to show which regions of an input image the model deems more important for its decision. These are typically rendered as 2D heatmaps overlaid on the camera view, as explored in works like [2, 5]. While useful, these 2D visualizations have two inherent limitations.

First, they collapse the 3D world into a flat image, losing the depth and spatial relationships that are essential for driving. As Figure 2 shows, a bright blob on a pedestrian does not tell us how far away the pedestrian is, whether they are on a collision course, or how they are positioned relative to other objects.

Second, and perhaps more important, a raw heatmap provides no semantic explanation. It cannot convey what the model is focusing on (a pedestrian? a traffic sign? a shadow?) nor why that element matters. The user sees a hotspot but does not know if the model has recognised a danger, is simply tracking a lane marking, or is distracted by an irrelevant detail. For instance, in Fig. 2 the bright region around the traffic light and its post could indicate that the model is not only focusing on the traffic light but also on the region before crossing the road to check for pedestrians about to cross.

The LLada framework proposed by Zhou et al. [6] (see Fig. 6) addresses this semantic gap but with a different objective. LLada predicts where a human driver would look (a 2D attention map) and also generates natural language descriptions: what the driver is looking at and why it is important. This turns an abstract heatmap into a human-readable explanation. However, LLada is designed for driver attention prediction, which means it explains human gaze, not the internal decision-making of an autonomous driving model. In our work, we need to explain what our own E2E model is attending to. Therefore, we will repurpose and fine-tune LLada so that its textual outputs describe the model’s focus, not a human’s.

Even with such model-focused semantic explanations, a deeper problem remains: these explanations are still locked to the 2D image plane. The “what” and “why” text sits as flat captions or overlays, completely detached from the 3D world the vehicle actually navigates. An engineer cannot see the distance to the attended object, its position relative to the car’s path, or the spatial layout of the scene. The open challenge, then, is how to combine rich semantic descriptions of a driving model’s attention with the 3D spatial context that driving demands, so that the explanation is both meaningful and spatially complete.

1.3 Project Objectives and Contributions

This work addresses this problem by developing a novel system that elevates 2D explainable attention into an interactive 3D space. The core hypothesis is that projecting model attention onto a 3D representation of the scene will provide more interpretability, allowing users to intuitively grasp not only what the model is looking at, but also the spatial context of that attention. To achieve this, the project is structured around the following concrete objectives:

- **Foundation - Model Replication:** To establish a performance baseline by fine-tuning the LLada model [6] on the W^3DA dataset, replicating the foundational results as a starting point for subsequent work.
- **Generating Semantic Explanations with a VLM:** Instead of manually annotating data, we will leverage a large, pre-trained Vision-Language Model to automatically produce the ”what” and ”why” descriptions for the regions highlighted by the target driving model’s attention maps. By feeding the VLM the camera image and the model’s attention hot-spots, we can obtain human-readable text that explains the model’s focus without expensive fine-tuning. Fig. 5 shows an example of the input to the model. This process will yield a dataset of driving scenes annotated with model attention maps and semantically rich, natural-language explanations.
- **Model Adaptation:** To adapt the pre-trained LLada model to the specific data distribution and driving scenarios of our generated dataset, ensuring the textual explanations faithfully reflect what the driving model is attending and why.
- **3D Visualization Tool:** To develop an interactive 3D visualization tool (e.g. using Open3D) where the predicted driver attention is volumetrically highlighted on the scene geometry, allowing users to explore the model’s focus from arbitrary viewpoints.

By achieving these objectives, this project will contribute:

Fig. 2: Activation maps of a state-of-the-art model called CIL++ at an intersection in Town02 [2] of the CARLA simulator [7]. Three image areas from different views are highly activated: the traffic light at the right image, the crossing pedestrians at the central one, and the lane shoulder at the left one. While these 2D heatmaps show where the model attends, they fail to convey critical 3D spatial information such as the distance to the pedestrians, their exact position relative to the vehicle's path, or the depth relationships between scene elements.

- A curated dataset that pairs E2E driving model attention maps with VLM-generated semantic explanations ("what" and "why"), capturing what the model is "thinking" in each scene.
- A fine-tuned version of the LLada model adapted to the dataset exposed above.
- A proof-of-concept 3D visualization tool that translates abstract AI reasoning into an understandable spatial format, demonstrating the value of 3D explanations over traditional 2D maps.

1.4 Document Structure

The remainder of this document is organized as follows. Section 2 reviews the state of the art in E2E driving, explainable AI for autonomous vehicles, positioning this work within the existing research landscape. Section 3 details the methodology, work plan, and resources required to achieve the stated objectives. Section 4 will describe the development process, followed by the presentation and discussion of results in Section 5. Finally, Section 6 will present the conclusions and outline directions for future work.

2 State-of-the-art

2.1 End-to-end (E2E) Autonomous Driving

Traditional autonomous driving systems are often decomposed into a modular pipeline, consisting of distinct components for perception, prediction, planning, and control [1] as seen in Fig. 1. By design, these modules avoid the "black box" problem: each component tackles a well-defined task and produces explicit, inspectable outputs (e.g. detectected object lists, predicted trajectories), so engineers can trace the reasoning behind any decision and pinpoint failures to a specific stage, enabling transparent debugging and safety verification. However, these modules are engineered separately and can suffer from error accumulation, as inaccuracies in one stage propagate to the next. E2E driving emerged as a paradigm to overcome these limitations by training a single neural network to map raw sensor inputs directly to low-level driving actions [1, 3]. Furthermore, this approach simplifies the data generation and training process. Instead of requiring precisely annotated datasets for each individual module (e.g. segmented images for perception or rule-based trajectories for planning), E2E models can be trained directly on large corpora of human demonstrations, learning to imitate expert behavior from raw sensor data alone. These advantages come at the cost of making the model less interpretable by design since you are no longer able to see the intermediary stages' outputs like a detected object list.

As surveyed by Tampuu et al. [1], E2E architectures have evolved significantly, leveraging advances in deep learning such as convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) for temporal modelling. These models learn driving policies implicitly from large-scale human driving data, demonstrating an ability to handle complex scenarios in simulation and controlled real-world settings. The promise of E2E driving lies in its simplicity and its potential to learn holistic driving behaviors that are difficult to capture with hand-crafted rules.

The Computer Vision Center (CVC) has been at the forefront of this research. Xiao et al. [2] proposed CIL++, an enhanced version of Conditional Imitation Learning that leverages multi-view attention learning and transformer-based feature encoding. By integrating multi-camera perspectives and advanced attention mechanisms, CIL++ overcomes the limitations of early imitation learning baselines, demonstrating how dynamically attending to distinct spatial

regions across different camera views can significantly improve driving performance and generalization in complex urban environments. Building on this, Porres et al. [5] explored methods for guiding the attention of such models during training, showing that incorporating human gaze data can lead to more robust and safer driving policies. These works establish a strong local foundation in E2E driving and, crucially, highlight the importance of attention as both a mechanism for performance and a window into model behavior.

2.2 Explainable AI (XAI) in Autonomous Driving

The performance gains of deep learning models, including those for E2E driving, have come at the cost of interpretability. This "black box" nature is particularly problematic in safety-critical domains like autonomous driving, where understanding the "why" behind a decision is as important as the decision itself. This has motivated the rise of XAI, a field dedicated to creating techniques that render AI models transparent and their outputs understandable to humans [4].

XAI methods can be broadly categorized. A common distinction is between intrinsic methods, which involve building models that are inherently interpretable (e.g. decision trees), and post-hoc methods, which aim to explain a pre-trained black-box model. Post-hoc methods are particularly relevant for explaining complex deep neural networks and can be further divided into model-specific or model-agnostic techniques. In the context of autonomous driving, a comprehensive overview by Atakishiyev et al. [8] discusses various XAI approaches, including visual explanations (saliency maps), concept-based explanations, and natural language explanations.

One of the most prevalent post-hoc techniques for vision-based models is the use of attention mechanisms, where attention weights can be visualized as 2D heatmaps that provide a coarse indication of where in the input image the model is "looking" when making a decision. A visualization of such maps can be seen in Fig. 2, Fig. 3, and Fig. 5. Works like [2, 5] have used these maps to gain insights into their E2E driving models. However, a significant limitation of these 2D attention maps is that they project the world into a flat plane. They fail to capture the critical depth information and spatial relationships that define a driving scene. For example, a user can see that a model is attending to a pedestrian, but not how close that pedestrian is or if it is seeing the pedestrian as a danger or not.

This limitation highlights the need for explainability methods that move beyond raw, low-level spatial data toward human-centric semantic context. Works like driver attention prediction work towards fixing it. Within the framework of XAI, predicting where a human driver would look in a given scene serves as more than just a behavioral model; it acts as an interpretability baseline. By aligning an autonomous agent's attention with a human's cognitive focus, driver attention prediction provides a benchmark for verifying whether an AI's internal reasoning matches safe, human-like driving logic.

The integration of driver attention into XAI is best exemplified by the "Where, What, Why" framework, embodied in the LLada (Large Language model-driven attention for driver attention) model proposed by Zhou et al. [6]. LLada's key innovation represents a paradigm shift in autonomous driving XAI, moving beyond simple spatial tracking to provide multi-faceted, human-readable explanations. As its name suggests, it predicts not only where a driver is looking (a 2D attention map) but also generates textual explanations detailing what they are looking at (e.g., "a pedestrian") and why it is important to the driving task (e.g., "because they are about to cross the street").

Architecturally, LLada functions as a vision-language model, employing a vision encoder to process the driving scene and a Large Language Model (LLM) to generate textual explanations. The model is trained to jointly predict the attention map and the corresponding text, effectively learning to articulate the cognitive reasoning behind a driver's gaze. This rich, multi-modal output makes LLada an ideal mechanism for addressing the semantic deficiencies of traditional visual XAI, transforming abstract attention weights into explicit, human-interpretable context. This capability inspires the present work, in which we leverage such human-interpretable explanations—though repurposed from driver attention to model attention—as described in the subsection 2.4.

2.3 Monocular Depth Estimation

To transition from 2D image-plane explanations to a fully realized 3D spatial context, recovering the scene's geometric structure from standard monocular camera feeds is essential. This task, known as Monocular Depth Estimation (MDE), involves predicting a continuous depth map from a single RGB image. In the context of autonomous driving, MDE acts as a computationally efficient alternative or complement to active sensors like LiDAR, enabling the mapping of pixel-level features to real-world 3D coordinates.

Early deep learning approaches treated MDE as a standard supervised regression task trained on ground-truth depth data from LiDAR or simulation [9]. However, gathering pixel-perfect ground truth is notoriously complex. To circumvent this, a prominent paradigm shifted toward self-supervised learning frameworks, which leverage spatial-temporal consistency across video sequences or stereo pairs to learn depth implicitly via photometric loss without requiring explicit annotations [10, 11]. Concurrently, modern architectures have advanced significantly by incorporating Vision Transformers (ViTs) as backbones to capture fine-grained structural boundaries and long-range dependencies across complex driving environments [12, 13].

Despite these advancements, standard MDE frameworks historically struggled with scale ambiguity, outputting relative depth maps that lack true physical dimensions. For autonomous driving applications, this requires secondary alignment or ground-truth scaling factors. To overcome this limitation, the field has recently transitioned toward Monocular Depth Foundation Models [14, 15]. Notably, architectures like Depth Anything V2 [15] leverage massive unannotated data combined with synthetic dataset scaling to perform robust, zero-shot metric depth estimation. By producing a highly accurate, metric-calibrated absolute depth profile directly from outdoor environments, these foundation models eliminate scale ambiguity.

In this work, rather than utilizing depth strictly for stan-

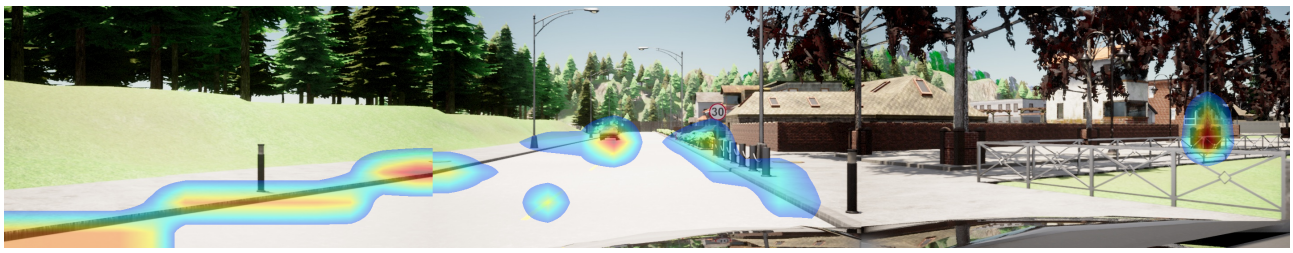

Fig. 3: Input images from the three cameras of the CARLA simulator [7], with attention heatmaps extracted from the guided-attention variant of the CIL++ model from Porres et al. [5] overlaid using the JET colormap (red/yellow = high attention). The heatmaps reveal where the model directs its focus before any depth projection or semantic annotation is added.

dard vehicle control, geometric reconstruction, or trajectory path planning, we leverage a metric MDE foundation model to unproject abstract 2D attention maps into a synchronized 3D structural point cloud. This grounding allows us to bind semantic natural-language explanations to authentic physical coordinates (X, Y, Z) relative to the ego-vehicle center.

2.4 Positioning of this work

The convergence of the fields discussed above reveals a clear opportunity. Research in E2E driving has established the importance of attention for both performance and interpretability [1, 2, 5]. XAI has provided frameworks for generating these explanations, with attention maps being the primary tool [4, 8]. Meanwhile, driver attention prediction has advanced to generate rich, semantic explanations, as exemplified by the LLada model [6]. Concurrently, the emergence of zero-shot metric depth foundation models [15] provides a robust mechanism to extract absolute spatial dimensions directly from monocular vision.

However, a critical gap remains. Both traditional attention visualizations and state-of-the-art driver attention explanations are ultimately confined to the 2D image plane. They are disconnected from the three-dimensional environment in which the vehicle operates. An end-user cannot easily grasp the spatial relationships that are fundamental to driving: the distance to an attended object, its position relative to the vehicle's path, or the layout of the scene.

This project directly addresses that gap by lifting model explanations into 3D. Importantly, we draw a clear distinction: our goal is not to explain where a human driver would look (driver attention prediction), but to explain where our autonomous driving model actually focuses its attention (model attention prediction). We therefore repurpose the human-centered LLada framework for model-centered explainability. Concretely, we fine-tune LLada on attention maps extracted from our E2E driving model (CIL++), paired with crafted textual descriptions of "what" and "why." In this way, LLada learns to generate natural language that articulates the reasoning behind the model's internal attention (not a human's).

At inference, we retain only this linguistic output. The 2D attention map that LLada would normally predict is discarded, because we already possess the driving model's own attention maps as the genuine signal of model focus. These model attention maps are projected into 3D using absolute physical distances extracted via a metric monocular depth estimation model [15], and LLada's textual "what" and "why" annotations are rendered as spatial labels on the resulting 3D view. The final output allows users to intuitively explore what the driving model is attending to, why that element matters, and (for the first time) where those objects are located in full spatial context. This work thus bridges the gap between state-of-the-art 2D/text explainability and the critical need for 3D spatial understanding, creating a novel and more powerful tool for building trust and verifying the behavior of intelligent driving agents.

3 Methodology and Work Plan

3.1 Overall Methodology

To achieve the objectives outlined in Section 1.3, this project will follow a phased, experimental methodology. This section details the specific tasks within each phase, the tools and resources required, and a realistic timeline. The methodology is designed to be iterative, allowing refinement based on findings in each phase.

1. **Setup & Foundation:** The initial phase focuses on dataset familiarisation, development environment setup, and a background replication of the LLada model [6] on its original W^3DA dataset. This step builds a solid understanding of driver attention prediction but is not a prerequisite for the core visualisation work.

2. **3D Visualisation Prototyping:** This is the project's core creative phase. Multiple prototype pipelines are constructed to lift 2D attention maps, extracted directly from a state-of-the-art E2E driving model (more specifically: [5]), into a 3D representation. Different projection techniques are explored, iterating on visual clarity and spatial interpretability using Open3D. By the end of this phase, a preferred pipeline is identified.

3. **Dataset creation and Model Adaptation:** First, we build a custom dataset that pairs the attention maps of the driving model with semantic "what" and why explanations. The attention maps are extracted from an E2E driving model, and the textual annotations are generated automatically by a large VLM prompted with the camera image and the model's attention map. This yields a dataset where each scene is annotated with the model's actual focus and a human-readable description of what it is attending to and why. Second, we fine-tune the LLada architecture on this dataset so

that it aligns with the driving model's attention rather than a human driver's gaze.

4. **Evaluation & Synthesis:** A qualitative evaluation assesses the interpretability gains of the final pipeline, comparing standard 3D attention views with the LLada-enriched version. The final report is written concurrently with this evaluation, synthesising all findings and conclusions.

3.2 Phase 1: Data Preparation and Model Setup (Weeks 1-6)

This phase corresponds to the first mandatory follow-up session in week 6 (March 16-20, 2025).

3.2.1 Task 1.1: Dataset Familiarization

The first task is a thorough analysis of the two primary datasets.

- **W^3DA Dataset:** The original dataset used to train LLada will be studied to understand its structure, annotations (attention maps, textual explanations), and driving scenarios. This is essential for replicating the model's baseline performance.
- **CIL++ model:** A detailed analysis of the model provided will be conducted. Key characteristics to be identified include: camera setups, and the nature of available annotations (e.g., pre-existing attention data).

3.2.2 Task 1.2: LLada Model Acquisition and Baseline Replication

The official codebase and pre-trained weights for the LLada model [6] will be obtained. The development environment will be set up using Python and PyTorch on the available GPU workstation. The first milestone will be to replicate the key results reported in the LLada paper on the W^3DA dataset's validation set. This step is crucial to verify the correct setup and to establish a performance baseline for comparison with later adaptations.

3.3 Phase 2: 3D Visualization Prototyping (Weeks 7-13)

This phase contains the project's core creative work and leads to the second mandatory follow-up session in week 13 (May 11-15, 2025).

3.3.1 Task 2.1: Extraction of Driving Model Attention Maps

We will use RGB images captured from the CARLA simulator [7] as input to the guided-attention variant of the CIL++ E2E driving model proposed by Porres et al. [5]. The model's internal 2D attention maps will be extracted from these images. These maps, which reflect the model's own focus rather than a separate predictor, will serve as the primary input for the subsequent 3D lifting process and will form the spatial basis of the final visualisations.

3.3.2 Task 2.2: Development of Multiple 3D Visualization Prototypes

Several parallel pipelines will be prototyped to project the 2D attention maps into 3D space, exploring different visualization strategies:

- **Full point cloud:** Plotting all depth-projected points from the attention map to see the raw spatial distribution.
- **Subsampled point cloud:** Plotting only a selected subset of points to reduce clutter and highlight salient regions.
- **Mesh-based markers:** Placing simple textured meshes (e.g. a square with an icon or label) at the 3D locations of key scene elements (such as a traffic light symbol over a detected traffic light) to provide an abstract, easily interpretable annotation layer within the 3D scene.

Depth maps for lifting the attention into 3D will be obtained using a Depth-Anything-V2 model [16, 17] (specifically *depth-anything/ Depth-Anything-V2-Metric-Outdoor-Small-hf*), which provides metric depth estimates suitable for outdoor driving scenes. This model is fine-tuned on the Virtual KITTI dataset [18]. All prototypes will be implemented with Open3D. Development will proceed iteratively, first on single static frames to evaluate visual clarity and spatial interpretability, then on short frame sequences to assess temporal coherence. By the end of Phase 2, the most effective pipeline (in terms of intuitive spatial representation and minimal visual clutter) will be selected as the base visualization system.

3.4 Phase 3: Model Adaptation and Semantic Enhancement (Weeks 14-21)

This phase enriches the base 3D visualization with human-like semantic explanations and runs concurrently with the final evaluation and report writing.

3.4.1 Task 3.1: Generation of a Model-Focused Explainability Dataset

A custom dataset is built by pairing RGB images from the CARLA simulator [7] with the corresponding attention maps extracted from an E2E driving model (e.g., CIL++ [2]). To automatically annotate driving scenes with "what" and "why" explanations, we require a vision-language model that can accurately identify objects and regions in complex driving images, follow structured multi-step prompts, and generate concise, context-aware reasoning. We selected Qwen2.5-VL-7B-Instruct [19] for the following reasons:

- **Strong visual reasoning ability:** Qwen2.5-VL is a state-of-the-art multimodal large language model that excels at detailed image captioning, visual question answering, and chain-of-thought reasoning. Its instruct-tuned variant reliably produces structured outputs that match our "Number of regions / What / Why" format.

- **Open-weight and locally deployable:** Unlike proprietary APIs, Qwen2.5-VL-7B can be run entirely on local GPUs. This eliminates API costs, avoids data-privacy concerns, and gives us full control over inference parameters.
- **Better alignment with the LLada annotation strategy:** The original W³DA dataset was annotated using Qwen-VL-Max [6]. Qwen2.5-VL-7B is a more recent and capable successor that preserves the same instruction-following philosophy while offering improved accuracy.
- **Sufficient scale for the task:** The 7B parameter size provides an excellent trade-off between performance and resource requirements. In half-precision (`torch.float16`), the model fits within the combined memory of two TITAN Xp GPUs (12 GB each), enabling rapid annotation of the full dataset on our available hardware without the need for external cloud APIs.

While other VLMs (e.g., LLaVA-1.6 [20], CogVLM2 [21]) were considered, Qwen2.5-VL-7B-Instruct was ultimately chosen for its combination of open availability, strong instruction-following, and compatibility with the LLada framework.

3.4.2 Task 3.2: Fine-Tuning LLada for Model Attention Explanation

The LLada architecture [6] is fine-tuned on the dataset created in Task 3.1. The objective is to adapt LLada so that it generates "what" and "why" explanations that describe the driving model's attention, rather than a human driver's gaze. During training, the model learns to produce coherent textual explanations from the driving scene and the associated model attention map. At inference, only the textual output is retained—the LLada-predicted attention map is discarded, as the true model attention maps are already available from the driving model.

3.4.3 Task 3.3: Integration of Semantic Overlays and Evaluation

The fine-tuned LLada's textual explanations are rendered as spatial annotations within the 3D visualisation, creating a semantic overlay on top of the driving-model attention. This yields a complete system that shows where the driving model is looking (volumetric attention highlights), what it sees, and why that matters, all in full 3D context. A qualitative evaluation compares the base 3D attention views (driving-model only) with the LLada-enriched version, focusing on interpretability gains, clarity, and user understanding, conducted on a set of representative driving scenes. The final project report is written in parallel, documenting the development process, design choices, results, and conclusions, ensuring readiness by the submission deadline.

3.5 Tools and Resources

The project will be developed using the following tools and resources:

- **Hardware:** Access to a powerful workstation equipped with NVIDIA GPUs (4 NVIDIA GeForce RTX 3090, 4 NVIDIA TITAN Xp and 1 NVIDIA TITAN Xp COLLECTORS EDITION).
- **Software:**
  - **Core Languages/Libraries:** Python, PyTorch, OpenCV.
  - **LLM/Transformers:** Hugging Face Transformers and Datasets libraries.
  - **3D Visualization:** Open3D for 3D processing and rendering in the prototypes.
- **Data:**
  - The publicly available W^3DA dataset [6].
  - A custom driving dataset consisting of RGB images captured in the CARLA simulator [7] and attention maps extracted from an E2E driving model (CIL++ [5]).
  - Publicly available model checkpoints from the Hugging Face Hub.

3.6 Work Plan and Timeline

The project timeline is designed to align with the mandatory follow-up sessions and final submission deadline (June 28, 2025). Table 1 presents a month-by-month breakdown of the tasks.

4 Development

This section details the technical realisation of the project, following the phased plan described in Section 3.1.

4.1 Phase 1 – Environment Setup and Preliminary Work

The development environment was set up on a shared GPU cluster that hosts a mix of NVIDIA RTX 3090 (24 GB) and TITAN Xp (12 GB) cards. A Python virtual environment was created with PyTorch, Hugging Face Transformers, and other required libraries, and all experiments were run under CUDA 12.8.

The initial task was to replicate the LLada model [6] on its original W^3DA dataset. After considerable debugging—the code provided by the authors contained several compatibility issues and did not run out-of-the-box on our cluster—we were able to successfully fine-tune the model and obtain the expected textual "what" and "why" explanations. This replication confirmed our understanding of the architecture and the training pipeline.

4.2 Phase 2 – 3D Visualisation Prototyping

To lift the 2D driving-model attention into 3D, we used the pre-trained `Depth-Anything-V2-Metric-Outdoor-Small` model [16] to infer dense metric depth maps from the central-camera frames of the CARLA [7] dataset (See more in detail in the Appendix B).

Period	Main Activities and Deliverables
Weeks 1-6 (Feb 16 - Mar 20)	Phase 1: Setup & Foundation - Literature review consolidation - Dataset analysis (W^3DA, CVC) - Environment setup and LLada baseline replication **Deliverable: Progress report for 1st mandatory follow-up (Week 6)**
Weeks 7-13 (Mar 21 - May 15)	Phase 2: 3D Visualization Prototyping - Develop different 3D visualization pipelines **Deliverable: Progress report for 2nd mandatory follow-up (Week 13)**
Weeks 14-21 (May 16 - Jun 28)	Phase 3 & 4: LLada fine-tuning, integration & evaluation - Choose final 3D visualization pipeline - Fine-tune LLada on the custom dataset - Final evaluation of all components - Writing the final report **Deliverable: Final project submission (June 28)**
July 2-8, 2025	Final Defense: Preparation and public presentation

TABLE 1: Proposed work plan and timeline for the project.

Each depth map was back-projected to a point cloud via the pinhole camera model. The point cloud was coloured with the original image pixels and filtered to remove points with invalid depth. The CIL++ attention maps were averaged across layers, upsampled to the image resolution, and transferred to the 3D vertices using the same depth-validity mask. Three prototypes were built and compared:

1. **Full point cloud** retained all points from the image, but was too dense and cluttered. Additionally, the noise from some bad estimations makes everything difficult to understand and the sheer quantity of pixels (even after downsampling with voxels) consumed excessive GPU memory and made interactive rendering impractical.

2. **Filtered point cloud:** retain only some points based on some ranking like specific distances or heatmap importance.

3. **Mesh-based markers:** an abstract representation that places simple 3D icons at the centroids of high-attention clusters.

The filtered point cloud was chosen as the base pipeline because it offers the best trade-off between interpretability and computational cost. In this variant, only points whose attention value (derived from the E2E driving model's heatmap) exceeds a threshold are kept, ensuring that the visualisation focuses on the most salient scene elements. A detailed technical description of the pipeline (model loading, depth estimation, coordinate transformation, attention colouring, and viewpoint setup) is provided in Appendix B.

4.3 Phase 3 — Model-Focused Dataset Generation and Semantic Integration

4.3.1 3.1 – Dataset Generation with Qwen2.5-VL-7B-Instruct

To provide human-readable explanations of the driving model's attention, we built a custom dataset that pairs the CIL++ attention maps with "what" and "why" textual annotations. The annotations were produced automatically using the **Qwen2.5-VL-7B-Instruct** vision-language model [19].

The generation proceeded in two stages:

1. **Single-camera annotations.** Initially, the pipeline processed only the central camera view. For each frame, the RGB image from the CARLA simulator [7] was loaded and the corresponding CIL++ attention map was normalised, converted into a JET heatmap, and blended at 50% opacity onto the original image. A structured prompt, very similar to the one used for the W^3DA dataset [6], asked the model to list the attention regions (*what*) and explain why they are important (*why*). The generated text was saved as a plain-text file. This stage produced a first collection of central-camera annotations.

2. **Multi-camera (stitched) annotations.** Because the CIL++ model receives three synchronised views (left, center, right) and its attention is distributed across them, a single-camera explanation cannot fully capture the model's reasoning. We therefore extended the annotation pipeline to merge all three views into one composite image. For each frame, the left, center, and right RGB images were each overlaid with their respective attention heatmaps and then concatenated horizontally. A dedicated prompt informed the VLM that the image contained three side-by-side perspectives and asked it to identify attention regions across all views, specifying to which camera each region belonged. The VLM outputs were saved alongside the

composite images, yielding a multi-view explainability dataset.

Throughout both stages the model was loaded in half-precision (`torch.float16`) and run on TITAN Xp GPUs. Inference used greedy decoding with a maximum of 300 new tokens per frame (central) and 400 tokens for the stitched variant. In total, around 2 000 central-camera frames were annotated, and the same frames were processed in the multi-camera setup, producing a unified collection of frame-aligned attention maps and semantic explanations.

4.3.2 3.2 – Integration of Semantic Explanations into the 3D Visualisation

Instead of the originally planned LLada fine-tuning, which had to be deferred due to persistent issues with the official codebase (see Phase 1), the VLM-generated annotations were directly incorporated into the 3D visualisation. Two complementary outputs were created:

- **Static four-panel composite.** The single-frame pipeline was extended to read the corresponding annotation file and render it in the top-right quadrant of the display. The resulting view shows the three input views (stitched horizontally) in the top-left panel, the "what" and "why" text in the top-right panel, the bird's-eye projection at the bottom-left, and the side-profile projection at the bottom-right (Fig. 4).
- **Animated sequence video.** Using the multi-camera 3D pipeline, all annotated frames were compiled into a video (2 fps, each frame appears for 0.5 seconds) that animates the attention distribution over time. Each frame follows the same four-panel layout, allowing the viewer to observe how the model's focus shifts as the driving scene evolves.

This integration gives the user an immediate, spatially grounded understanding of the model's behaviour: the volumetric highlights show *where* the model is looking across all three camera views, while the adjacent text panel explains *what* it sees and *why* that matters — all within a single, easy-to-interpret view, either as a static image or as a continuous sequence. An example composite frame is shown in Fig. 4.

5 Results

The project has progressed through the three planned phases, and the main deliverables have been produced according to the adjusted timeline. This section summarises the concrete outcomes of each phase, highlights the final integrated visualisation, and comments on the achieved interpretability gains.

5.1 Phase 1: LLada Replication and Environment Setup

The official LLada model [6] was successfully fine-tuned on the W^3DA dataset using the shared GPU cluster. After resolving numerous compatibility issues with the original codebase, the model produced the expected "what" and "why" textual explanations, confirming that the training pipeline was correctly understood. The intention was always to follow the paper's original recipe: first annotate a custom dataset with a vision-language model (Qwen-VL), and then fine-tune LLada on that dataset. The first part (annotation with Qwen2.5-VL-7B-Instruct [19]) proceeded smoothly, but the second part (fine-tuning LLada on our new data) could not be completed because the official code proved fragile and the large memory footprint made training on our hardware impractical within the project timeline. With more time, the LLada fine-tuning step would still be pursued; for this project, the VLM-generated annotations were used directly, achieving the same semantic enrichment without the need for an additional fine-tuned model.

5.2 Phase 2: 3D Visualisation Pipeline

A fully functional 3D visualisation pipeline was built. Using metric depth maps estimated by Depth-Anything-V2 [16], 2D attention maps from the CIL++ model [2] were lifted to a three-dimensional point cloud. Three variants were prototyped and compared:

- **Full point cloud** – contained all projected points but was too dense and noisy for interactive exploration.
- **Filtered point cloud** – retained only points within a 20 m radius from the ego-vehicle, dramatically reducing clutter while preserving the attention signal in the most safety-relevant region.
- **Mesh-based markers** – required a separate object detector and complex geometric modelling; it was abandoned due to time constraints and the risk of shifting the project's focus.

The filtered point cloud was selected as the base pipeline. It offers a clear spatial representation of the model's focus and runs smoothly on a single GPU. Figures 9 and 10 in the Appendix illustrate the difference in visual clarity.

5.3 Phase 3: Semantic Dataset and Integrated Composite View

A dataset of around 2 000 central-camera frames (Town02, CARLA [7]) was annotated with "what" and "why" explanations using Qwen2.5-VL-7B-Instruct [19]. The VLM was prompted with the RGB image and the model's attention heatmap blended as a grayscale overlay. The generated text was saved in plain-text files, forming a reusable model-focused explainability dataset.

These semantic explanations were integrated into the single-frame visualisation as a four-panel composite (Fig. 4):

1. Top-left: original camera image.
2. Top-right: the VLM's "what" and "why" text, automatically scaled.
3. Bottom-left: bird's-eye view of the attention-filtered point cloud.
4. Bottom-right: side-profile view.

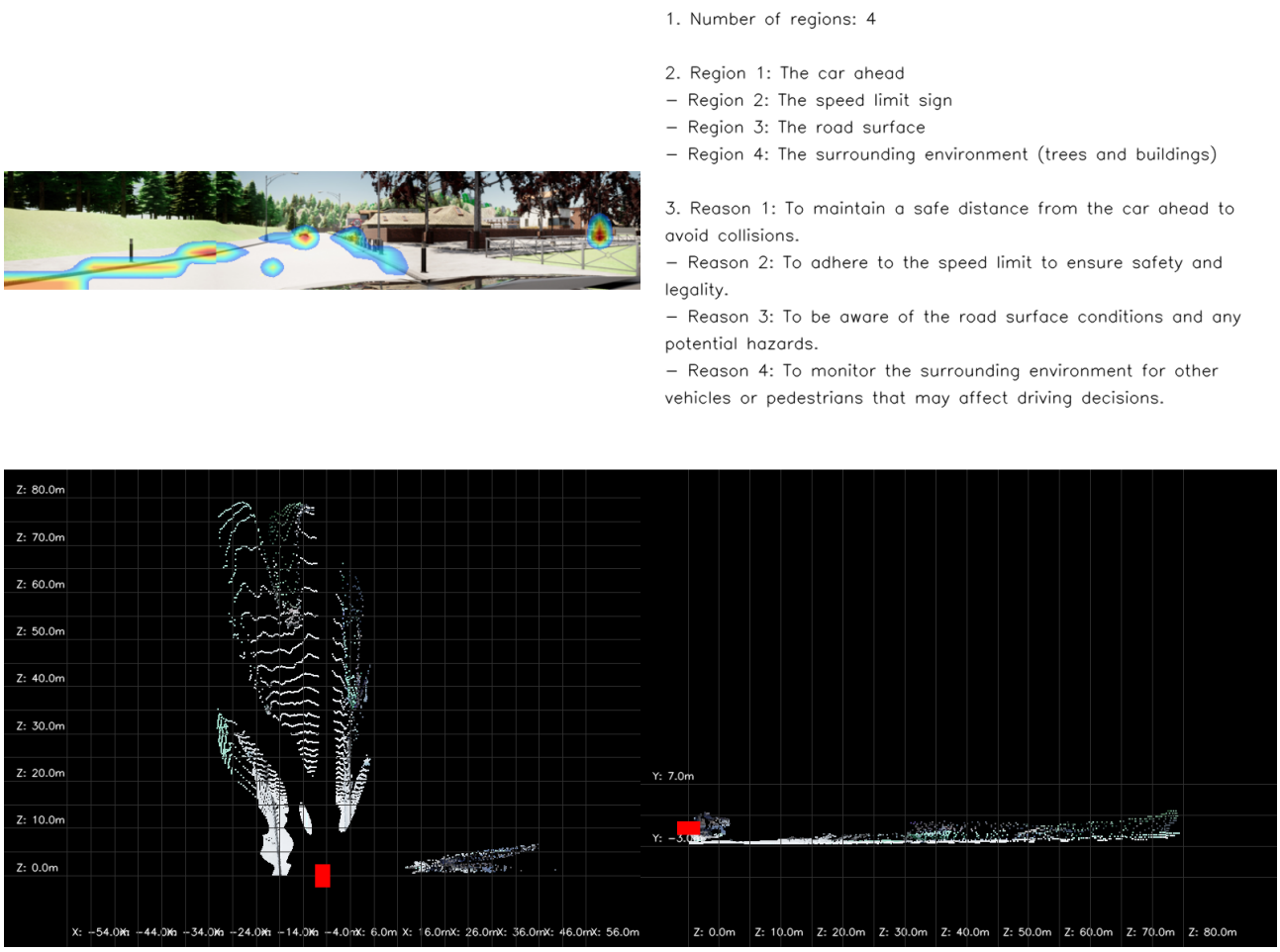

Fig. 4: Example composite frame illustrating the integrated 3D explanation. **Top-left:** stitched left, center, and right camera views with attention overlays (Fig. 3 shows it in full resolution). **Top-right:** the VLM-generated *what* and *why* textual explanation, automatically scaled to fit the panel. **Bottom-left:** bird's-eye projection of the 3D point cloud, filtered by attention threshold. **Bottom-right:** side-profile projection.

The integration required only a few additional lines of Python code, and the resulting composite can be generated in under two seconds per frame on the available hardware since the text has already been generated.

5.4 Qualitative Assessment of Interpretability

An internal, qualitative inspection of several representative scenes confirmed that the composite view conveys information that is impossible to obtain from a 2D heatmap alone:

- The **BEV and side views** reveal the distance, lateral position, and height of the attended elements, answering *where* in the 3D world the model's focus lies.
- The **text panel** describes *what* the model sees and *why* that element is relevant, transforming an abstract hotspot into a human-readable justification.

At the same time, the BEV and side projections can be somewhat noisy: sparse point clouds and depth-estimation artifcats occasionally produce scattered points that make it hard to quickly grasp the exact spatial layout, especially for users unfamiliar with the visualisation.

For instance, in Fig. 4 the model attends to a car ahead amongst other things. The BEV shows that the car is roughly 55 m in front of the ego-vehicle; the textual explanation correctly identifies the attended object (the car ahead) and reads: "To maintain a safe distance from the car ahead to avoid collisions."

However, not all explanations were accurate. For instance, Fig. 5 shows a scene where the VLM produced the following annotation:

```
Number of regions: 3
Region 1: Cyclist (view: left)
Region 2: Cyclist (view: center)
Region 3: Cyclist (view: right)
Reason 1: To avoid collision with
the cyclist ahead (left view).
Reason 2: To maintain a safe
distance from the cyclist (center
view).
Reason 3: To ensure the cyclist is
not obstructed by the vehicle (right
view).
```

This output illustrates two common failure modes observed in the internal study:

- **Incorrect region count:** the model reported three attention regions, but the heatmap clearly shows four, a traffic sign in the left view is overlooked.

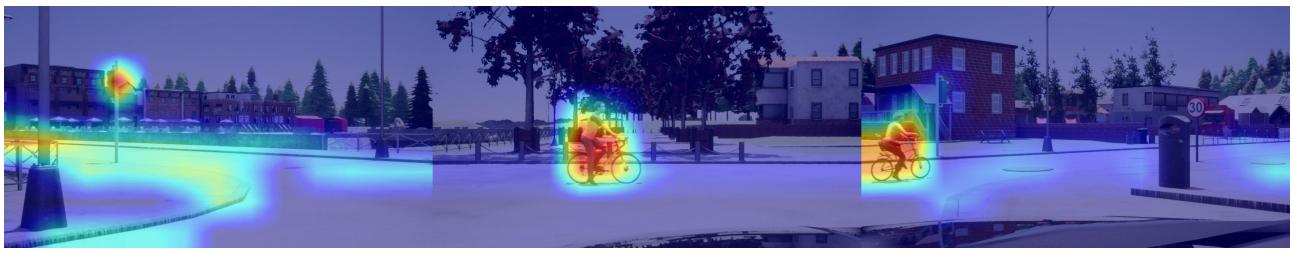

Fig. 5: Example of the input images (stitched for clarity) for the VLM-generated annotation. The frame shows four distinct attention regions (the traffic sign (the back of a traffic light) on the left is clearly highlighted), yet the model reports only three regions—all labelled as 'cyclist'. Also in the left view, the attended object is actually a fence, not a cyclist. This illustrates the kind of mistake that can occur when context is limited and the heatmap overlayed over the image.

- **Mislabeled objects:** all three regions are labelled as 'cyclist', yet the attended object in the left view is a fence, not a cyclist.

Such errors illustrate that the VLM can struggle to correctly count and identify objects, especially when the merged image is small and context is limited.

It is important to note that this assessment was conducted on a small set of scenes recorded under similar, benign weather conditions (sunny, daytime) and without highly complex traffic interactions. Even in these relatively simple scenarios, the VLM-generated explanations occasionally contained minor inaccuracies, such as misattributing an attended region to a generic "road ahead" when a more specific element (e.g., a vehicle) was present. A larger and more diverse evaluation, including adverse weather and dense traffic, would be needed to fully characterise the robustness of the semantic explanations and the 3D visualisation.

5.5 Adherence to the Work Plan

All milestones of the revised work plan (Table 1) were met on schedule. The only significant deviation from the original proposal—the replacement of LLada fine-tuning with direct VLM annotation—did not compromise the core contribution; on the contrary, it yielded a simpler, more robust pipeline that still provides semantic explanations aligned with the driving model's attention. The project therefore successfully bridges the gap between 2D attention maps and spatially grounded, human-interpretable 3D explanations.

6 Conclusions and Future Work

6.1 Conclusions

This project set out to bridge the gap between opaque 2D attention heatmaps and spatially grounded, human-interpretable explanations of an E2E driving model. The main achievements can be summarised as follows:

- A functional 3D visualisation pipeline was built that lifts 2D attention maps from the CIL++ model [2] into a metric point cloud using the Depth-Anything-V2 foundation model [16].

- A custom dataset of driving scenes was automatically annotated with "what" and "why" semantic explanations by the Qwen2.5-VL-7B-Instruct VLM [19], capturing the reasoning behind the model's focus.

- A four-panel composite view was created (RGB image, textual explanation, bird's-eye view, side view) that presents, in a single frame, *what* the model attends to, *why* it matters, and *where* it is located in the three-dimensional world.

Qualitative inspection of several driving scenes confirmed that the combined spatial and semantic information provides a deeper level of interpretability than 2D heatmaps alone. The bird's-eye and side views immediately reveal distances, lateral positions, and depth ordering, while the text panel transforms an abstract hotspot into a meaningful, human-readable justification.

Nevertheless, several limitations temper these achievements. The LLada fine-tuning step, originally intended to produce a fully end-to-end explanation model, could not be completed due to the fragility of the official codebase and hardware constraints. The current pipeline also exhibits practical weaknesses:

1. The BEV and side projections can become noisy and cluttered, specially when depth estimation produces sparse or scattered points, making rapid spatial interpretation difficult.

2. The VLM-generated annotations occasionally miscount or mislabel attended regions, reflecting the limitations of a generic VLM prompted with limited context.

3. The evaluation was performed on a small set of scenes under benign, constant conditions (sunny, urban, daytime), so the ability to generalise of the findings to adverse weather or dense traffic remains unexplored.

4. Finally, the evaluation itself is purely qualitative and internal, no user study or quantitative metric has yet been applied.

6.2 Future Work

The results obtained open several clear avenues for improvement and extension:

- **LLada fine-tuning and data-scale study.** Completing the fine-tuning of the LLada architecture on our

model-focused dataset and evaluating its explanation quality remains a primary goal for future work. An important open question is *how many* annotated samples are required for the fine-tuned LLada to faithfully capture the reasoning of a particular driving model; a systematic study varying the dataset size would provide valuable guidelines for practical deployment.

- **Richer context for VLM prompting.** At present, the prompt fed uses handcrafted metadata (e.g. morning, urban, sunny). Providing real situational data such as the ego-vehicle's speed, navigational commands, and weather/temperature could yield more precise and context-aware textual explanations, improving the overall quality of the generated annotations.

- **Exploring alternative explainable architectures.** The modular design of our pipeline makes it straightforward to replace the annotation engine. More recent multimodal LLMs could be evaluated as the semantic backbone. Comparing their explanation quality, inference speed, and resource consumption would help identify the most suitable candidate for a real-time, on-board explainability module.

- **Improved multi-camera composite layout.** The stitched RGB views in the top-left panel are currently too small to see fine details clearly. A redesigned layout with larger camera images would improve the practical usability of the visualisation.

- **User study and quantitative evaluation.** A formal user study with end users, driving engineers, and AI researchers would quantify the interpretability gain, measuring how quickly and accurately participants can answer questions about the model's behaviour using the 2D heatmaps versus the 3D composite views.

- **Temporal coherence and real-time performance.** The current pipeline operates on individual frames. Adding temporal smoothing and displaying attention dynamics in a video would give insight into how the model's focus shifts over time. Optimising the depth estimation and rendering steps for real-time operation is a necessary step toward integrating such a tool into a live debugging or monitoring dashboard.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Antonio M. López Peña, for proposing this project and for accepting me as his student. His guidance and insight have been invaluable throughout this work.

I am deeply thankful to Alexandre François Levy for his continuous help and support during the whole project. His expertise and willingness to discuss ideas have greatly enriched my research experience.

I also wish to extend my thanks to the ADAS group at the Computer Vision Center (CVC) for providing me with ideas and a stimulating environment that helped shape the direction of this project.

REFERENCES

[1] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A survey of end-to-end driving: Architectures and training methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1364–1384, 2022.

[2] Y. Xiao, F. Codevilla, D. Porres, and A. M. López, "Scaling vision-based end-to-end autonomous driving with multi-view attention learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1586–1593.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[4] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[5] D. Porres, Y. Xiao, G. Villalonga, A. Levy, and A. M. López, "Guiding attention in end-to-end driving models," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2353–2360.

[6] Y. Zhou, J. Tang, X. Xiao, Y. Lin, L. Liu, Z. Guo, H. Fei, X. Xia, and C. Gou, "Where, what, why: Towards explainable driver attention prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2675–2685.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[8] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *IEEE Access*, vol. 12, pp. 101 603–101 625, 2024.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Predicting depth, surface normals and semantic labels from a single image using a deep convolutional network," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2366–2374.

[10] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838.

[11] T. t. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1851–1858.

[12] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 173–12 183.

[13] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4009–4018.

[14] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[15] L. Yang, B. c. Kang, Z. Huang, L. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[16] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv:2406.09414*, 2024.

[17] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[18] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[19] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wang, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[20] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024.

[21] W. Hong, G. Wang, *et al.*, "Cogvlm2: Visual language models for next-generation applications," *arXiv preprint arXiv:2405.15170*, 2024.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[23] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[24] J. Xie, Z. Yang, and M. Z. Shou, "Show-o2: Improved native unified multimodal models," *arXiv preprint arXiv:2506.15564*, 2025.

A Overall Architecture of LLada

The LLada model [6] is a vision-language framework that jointly predicts a driver's spatial attention map and generates natural-language explanations. At its core, a vision encoder (CLIP) processes the input driving image into visual tokens, which are then fed into a large language model (LLM) together with tokenized contextual prompts. To bridge the two modalities and force the LLM to reason about visual saliency, a special learnable token `[ATTN]` is introduced. During training, the model is optimised with a multi-task loss that combines a saliency prediction objective (e.g., Kullback–Leibler divergence against human gaze fixations) with the standard language-modelling loss for explanation generation.

The attention map is not a by-product of the LLM's internal self-attention; instead, the final representation of the `[ATTN]` token is used to query the visual features via cross-attention in a dedicated "cognitive-aware attention decoder". This decoder produces a 2D heatmap that highlights the regions most relevant to the generated explanation. Consequently, the model simultaneously answers "where" a human driver looks (the attention map), "what" they are looking at (semantic description), and "why" it matters (causal justification). This tight coupling between visual grounding and language generation is what makes LLada suitable as a source of semantic overlays in our 3D visualisation pipeline. The overall architecture is illustrated in Fig. 6.

B Detailed 3D Visualisation Prototype Pipeline

The code is organised as a single script that sequentially builds and evaluates the different variants described below.

B.1 Environment and Depth Estimation

The environment was built with `torch`, `transformers`, `opencv-python`, `matplotlib`, `open3d`, and `plotly`. The pre-trained model `Depth-Anything-V2-Metric-Outdoor-Small` [16] was loaded from Hugging Face Hub. Each central-camera frame was resized to a width of 250 px before inference to keep memory usage and processing time low. The model outputs a dense metric depth map (Fig. 8).

B.2 Full Point Cloud (Pipeline 1)

Using a pinhole camera model with focal lengths $f_x = f_y = 0.8\,W$ and principal point at the image centre, every pixel (u, v) was back-projected to a 3D point:

$$X = \frac{(u - c_x)\,Z}{f_x}, \quad Y = \frac{(v - c_y)\,Z}{f_y}, \quad Z = \mathrm{depth}(u, v).$$

All points with $Z > 0$ were stored in an Open3D `PointCloud` together with the original image colours. This produces a full 3D reconstruction of the scene (Fig. 9), but the enormous number of points and depth-estimation artefacts create severe visual clutter and slow down rendering, making interactive exploration impractical.

B.3 Range-Filtered Point Cloud (Pipeline 2)

To focus the visualisation in the important parts rather than the whole image thus improving performance, we just need to focus on some important parts. To know which points are the most relevant a couple of methods arose: thresholding

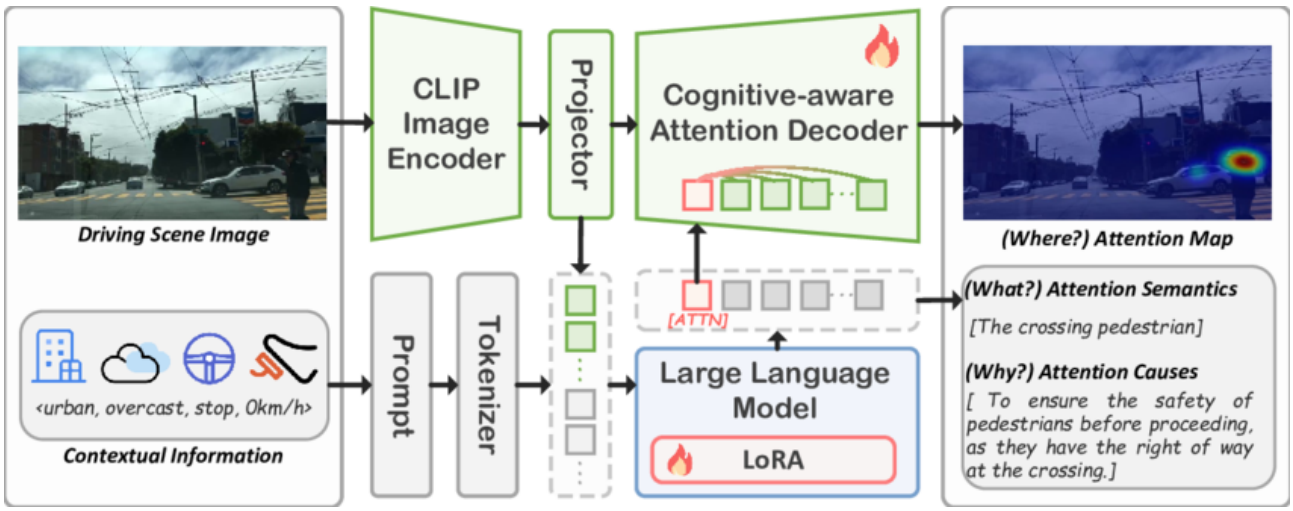

Fig. 6: Overall architecture of LLada. Given a driving scene represented as an image with contextual information, the image is processed by the visual encoder (CLIP encoder and projector) to extract visual tokens, while the contextual information, along with prompt sequences, is tokenized by the LLM tokenizer. The visual and text tokens are then fed into the LLM. To adapt the LLM for explainable attention prediction tasks, a special attention token [ATTN] is introduced to encode high-level cognitive cues. Finally, [ATTN] performs cross-attention with the visual tokens, generating the attention map (answering Where?) through a cognitive-aware attention decoder, while the LLM produces the semantic descriptions (answering What?) and causal explanations (answering Why?) of attended regions. (Figure adapted from Zhou et al. [6])

Fig. 7: Input image from CARLA [7].

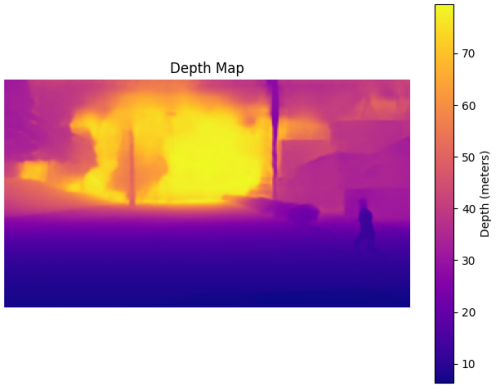

Fig. 8: Estimated depth map.

the attention map to select the regions the model attended or select a distance to the car and plot all the points that are closer. Since the attention map provided wasn't from a fully trained model and it didn't focus on the important parts, the implementation focused on the distance based one.

In the implementation case, all points with $Z > 20$ m were discarded. After this depth filtering the point count drops dramatically (Fig. 10), allowing smooth interactive manipulation and revealing the spatial distribution of attention clearly.

B.4 Why Mesh-Based Markers Were Discarded (for Now)

A third approach, placing symbolic 3D meshes (e.g., pedestrian icons, traffic-light symbols) at the centroids of high-attention clusters, was designed but not fully implemented. Three reasons led to this decision:

1. **Object detector requirement:** the mesh approach requires an independent module that segments or detects specific object instances in the scene. Building such a module would shift the project's focus away from attention visualisation and further add computational cost.
2. **Object complexity:** some objects are difficult to render as simple standard meshes (they may have curves and other complexities). Which complicates a scene reconstruction.
3. **Development time:** Phase 2 had a fixed deadline; implementing a robust mesh-annotation system would have delayed the prototyping of the point-cloud pipelines.

For these reasons, the mesh concept is noted as not working properly and the current pipeline focuses on the range-filtered point cloud with attention colours.

C Prompts for Data Creation

Two distinct prompts were used to guide the Qwen2.5-VL-7B-Instruct model in generating seman-

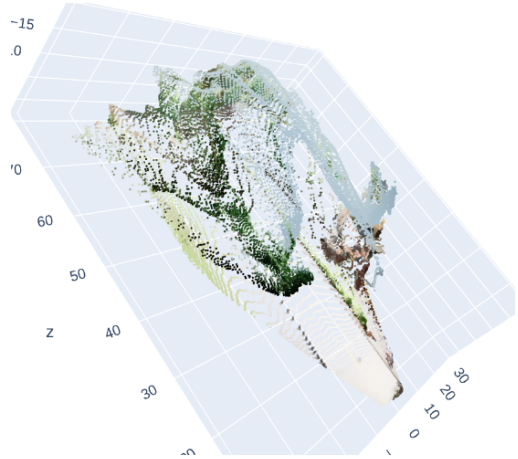

Fig. 9: Full point cloud coloured with the original image. The high density and noise obscure important regions. The point cloud has been voxelized to downsample more the number of points.

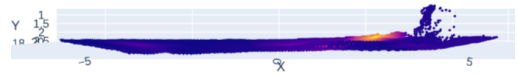

Fig. 10: Range-filtered point cloud ($Z \leq 20$ m) coloured by driving-model attention. The scene is much easier to interpret, and the viewpoint can be rotated interactively at the cost of not seeing the full scene.

tic annotations. They are based on the ones used in the paper [6]

Single-Camera Prompt

```
Imagine you are driving in the provided image.
<Contextual Information>
The driving scenario is as follows: It is {
    time_of_day} in a {location} area with {
    weather}
</Contextual Information>
The image shows your visual attention
    distribution. Please answer the following
    questions briefly:
<Question and Output Format >
1. How many regions of attention are present in
    the image?
- Format your response as: 'Number of regions: [
    number]'
2. What are the specific regions where the driver
    's attention is focused?
- For each region, format your response as
    follows:
- 'Region 1: [Name of the first region]'
- - (If applicable) 'Region 2: [Name of the
    second region]'
- - (If applicable) 'Region 3: [Name of the third
     region]'
3. Why is the driver focusing on these regions?
- For each reason, consider the impact on
    upcoming driving decisions, start with 'To
    ... ', and format your response as follows:
- 'Reason 1: [Explanation for the first region]'
- - (If applicable) 'Reason 2: [Explanation for
    the second region]'
- - (If applicable) 'Reason 3: [Explanation for
    the third region]'
Important Notes for Your Response:
- Avoid assumptions about elements not visible in
     the image.
</Question and Output Format >
```

Multi-Camera (Stitched) Prompt

```
Imagine you are driving in the provided image.
<Contextual Information>
The driving scenario is as follows: It is {meta['
    time_of_day']} in a {meta['location']} area
    with {meta['weather']}.
\</Contextual Information\>
The image is a merged view of three camera
    perspectives (left, center, right) placed
    side-by-side. The image also shows visual
    attention distributions (heatmaps) for each
    view.
Please analyze the entire scene and answer the
    following questions briefly:
<Question and Output Format >
1. How many distinct regions of attention are
    present across all views?
- Format your response as: 'Number of regions: [
    number]'
2. What are the specific regions where the driver
    's attention is focused? Indicate which view
    (left, center, or right) each region belongs
    to.
- For each region, format your response as
    follows:
- 'Region 1: [Name of the region] (view: [left/
    center/right])'
- - (If applicable) 'Region 2: [Name of the
    region] (view: [left/center/right])'
- - (If applicable) 'Region 3: [Name of the
    region] (view: [left/center/right])'
3. Why is the driver focusing on these regions?
- For each reason, consider the impact on
    upcoming driving decisions, start with 'To
    ... ', and format your response as follows:
- 'Reason 1: [Explanation for the first region]'
- - (If applicable) 'Reason 2: [Explanation for
    the second region]'
- - (If applicable) 'Reason 3: [Explanation for
    the third region]'
Important Notes for Your Response:
- Avoid assumptions about elements not visible in
     the image.
</Question and Output Format >
```

D Adjustments to the Original Work Plan

Since the submission of the initial report, several modifications have been made to the objectives, methodology, and timeline of the project. These changes were driven by practical constraints encountered during early implementation and by a sharper focus on the core contribution: the 3D visualisation of driving attention. All adjustments were discussed with and approved by the supervisor.

D.1 Removal of Comparative Benchmarking

The original plan called for a comparative study of multiple VLMs (BLIP-2 [22], EMU-3 [23], Show-o [24]) on driver attention prediction. However, the focus of the project lies in the 3D visualization and the LLada should be added on top. After revision of the timeline, looking into different architectures for the language task was dropped and the focus

was put in making the LLada work correctly for the generated dataset. This is specially important since the W^3DA dataset is focused on driver attention prediction and not explaining the model's outputs.

D.2 Replacement of the Learned Distance Prediction Module

The initial proposal intended to extend a model with a custom head that predicts the distance to the attended object. However, this required time and making sure that the LLada model worked, thus reducing the time the 3D pipeline was worked upon. Meanwhile, the pre-trained Depth-Anything-V2 model [16] offered a robust, off-the-shelf alternative that produces metric depth maps from single images without any task-specific fine-tuning. Using this model allowed us to lift 2D attention into 3D space immediately, without the risk and overhead of training a dedicated module. Consequently, the distance prediction and its associated benchmarking were dropped from the objectives, and the 3D pipeline now relies entirely on Depth-Anything-V2 for depth information.

D.3 Reordering of Development Phases

The work plan has been reorganised to prioritise the core 3D visualisation. In the initial plan, model fine-tuning and benchmarking (Phase 2) preceded the visualisation work (Phase 3). In the revised plan, Phase 2 is exclusively devoted to prototyping the 3D visualisation pipeline using attention maps from existing CIL++ models [5] and depths from Depth-Anything-V2. The fine-tuning of LLada on the generated dataset is now Phase 3, where its outputs will add an optional semantic overlay on top of an already functional 3D viewer. This reordering reduces risk: a working visualisation is delivered early, while the LLada enrichment becomes an enhancement rather than a bottleneck.

D.4 Updated Objectives

As a result, the official objectives have been simplified (See Objectives 1.3 for more detail).

These revised objectives are fully reflected in the methodology description and in the updated work plan (Table 1). The core contribution (spatial interpretability of autonomous driving agents) remains unchanged, and the project is on track to deliver a functional 3D explanation tool within the original deadlines.