Supplementary Material
Distilling *What* and *Why*: Enhancing Driver Intention Prediction with MLLMs

Sainithin Artham* Avijit Dasgupta* Shankar Gangisetty C. V. Jawahar
CVIT, IIIT Hyderabad, India

A. Additional Implementation Details

Video-LLaMA [8]. We use the LLaMA-2-Chat backbone under multimodal supervision. The architecture comprises a frozen Vision Transformer (ViT) [1] encoder and a Q-Former [4] module with 32 query tokens. Video inputs are processed through a frame-level visual encoder with a maximum of 32 temporal positions and are fused using a two-layer sequential transformer. Training was conducted over 3 epochs, each consisting of 1000 iterations, with a batch size of 2. A linear warmup followed by a cosine decay learning rate schedule was used, starting with an initial learning rate of 3×10^{-5}, warming up from 1×10^{-6}, and decaying to a minimum of 1×10^{-5}.
Qwen 2.5-VL [6]. The vision encoder is kept frozen during training, while the multimodal MLP and language model [2] components are updated. The model is trained for 3 epochs with a batch size of 4, using a cosine learning rate schedule starting at 2×10^{-7} and a warmup ratio of 3%. To reduce memory usage, the input sequence length is capped at 8192 tokens. Checkpoints are saved every 1000 steps, with only the most recent checkpoint retained.

Both models were trained solely on the vision branch using 8 frames per video, with each frame resized to 224×224 pixels. Pre-fine-tuned checkpoints were used for initialization. All training and inference were conducted on an NVIDIA RTX A6000 GPU with 40 GB of memory.

B. End-to-end inference latency

As shown in Table 1, our Qwen2.5-VL-ED achieves both higher accuracy and F_1 while maintaining significantly lower latency compared to other baselines. The codebases for action anticipation models are unavailable, preventing direct comparison, though their inference latency is likely lower given our larger parameter sizes.

C. Prompt for Scene Context

Given a driving video input V, we prompt the model M(Videollama) to gener- ate a description of the scene. We design a prompt that guides the model to focus on critical elements such as ego-vehicle behavior, road semantics, and interactions with surrounding entities (see Figure 1).

*Both authors have contributed equally to this research.

Model	Params (B)	Acc. (%)	F_1 (%)	Latency (ms/video)
InternVL2	8	34.57	27.54	634 ± 95
LLaVA-NeXT	7	10.86	2.16	639 ± 85
Video-LLaMA3	7	25.73	23.91	524 ± 70
Qwen2.5-VL-ED (Ours)	**7**	**72.28**	**73.81**	329.77 ± 55

Table 1. **Performance–latency trade-offs on Brain4Cars.**

Prompt Template

System Instruction: You are able to understand the visual content that the user provides. Follow the instructions carefully and explain your answers in detail.
User Input: You are given a driving video clip from an egocentric perspective. Your task is to generate a **detailed and structured caption** that clearly describes the **maneuver and scene dynamics**.
Focus Areas:
- Movement and trajectory of the ego vehicle (e.g., lane change, turn, stop, acceleration)
- Changes in direction (e.g., turning left, merging right)
- Road context (e.g., traffic signals, lane markings, road type)
- Presence and behavior of surrounding vehicles or obstacles

Constraints: Only output the summary. Do not include any commentary or additional explanation. Be concise but informative.

Figure 1. Prompt template for generating driving scene descriptions.

D. Prompt for Generating Maneuvers and Explanations

We use multimodal inputs—including road and lane masks ($s_{t,i}$), optical flow ($d_{t,i}$), scene context ($\mathcal{C}_f$, $\mathcal{C}v$), and surrounding context ($\mathcal{C}src$)—as prompts to classify driving maneuvers and generate corresponding explanations. Eight frames are

LLM Prompt Template: Multimodal Driving Maneuver Classification and Explanation Generation

Basic Instruction: Classify the driving maneuver based on Road/Lane Masks, Optical Flow, Frame-wise and Video Captions, and Surrounding Context.

Constraints:

- Choose one label only: *right turn, right lane change, left turn, left lane change, forward*.
- Use only the input information.
- Do not reference specific input sources in your explanation.
- Follow the output format exactly.

Label Definitions:

- **right turn:** Sharp/significant turn to the right.
- **right lane change:** Shift into adjacent right lane.
- **left turn:** Sharp/significant turn to the left.
- **left lane change:** Shift into adjacent left lane.
- **forward:** Continue ahead or stop (e.g., halt, red light, park).

Optical Flow Example:

```
Frame90:  right, right, ..., right, left
Frame120:  right, left, ..., right, left
Frame150:  right, right, ..., left, left
```

Label: **right turn**

Road/Lane Mask Example:

```
Frame90:  road, lane, ..., lane
Frame120:  road, road, ..., lane
```

Label: **right lane change**

Multimodal Inputs:

Frame-wise captions:	$\{\mathcal{C}_f\}$
Video-wide captions:	$\{\mathcal{C}_V\}$
Surrounding context:	$\{\mathcal{C}_{\text{src}}\}$
Optical Flow:	$\{d_{t,i}\}$
Road/Lane masks:	$\{\ s_{t,i}\}$

Output Format:

```
Label: <one of the 5 maneuver labels>
Explanation1: <Justification>
Explanation2: <Justification>
Explanation3: <Justification>
Explanation4: <Justification>
Explanation5: <Justification>
```

Figure 2. Prompt for generation of zero-shot maneuver and explanation in motion context using multimodal inputs.

sampled to select in-context examples for OFM and RLM. The maneuver classes includes: *right turn*, *right lane change*, *left turn*, *left lane change*, and *forward* for Brain4Cars [3] and AIDE [7]. For the DAAD [5], we extend the maneuver classes set by adding *slow down* where the vehicle decelerates or prepares to stop and *U-turn* where the vehicle makes a U-turn.

Prompt Template for Explanation Ordering

Task: You are provided with a driving video and its associated high-level driving maneuver label. Based on the video and the given label, evaluate each of the following five explanations. Assign a label indicating how well each explanation is supported by the visual evidence and how consistent it is with the provided label.

Scoring Guidelines:

`strongly supported`: Clearly and directly aligned with both video and label.

`supported`: Generally consistent but lacks some detail.

`moderately supported`: Partially aligned; somewhat ambiguous.

`weakly supported`: Minimal relevance to video or label.

`not supported`: Unrelated or contradictory to video and label.

Important Constraint: Do not assign the same label to any two explanations. You are expected to **rank** the explanations by their relevance and support.

Input:

label:$\{\mathbf{y}\}$
explanation1:$\{\mathbf{E_1}\}$
explanation2:$\{\mathbf{E_2}\}$
explanation3:$\{\mathbf{E_3}\}$
explanation4:$\{\mathbf{E_4}\}$
explanation5:$\{\mathbf{E_5}\}$

Output Format: Return a list of five labels in the same order as the explanations.

Example Output: `[strongly supported, supported, weakly supported, moderately supported, not supported]`

Figure 3. Prompt for explanation ordering.

The prompt designed for the generation of maneuvers and explanations in motion context is shown in Figure 2.

E. Prompt for Explanation Ordering

Given the maneuver and explanations generated by DriveXplain, we input them into the VLM using the prompt shown in Figure 3, enabling the model to rank the explanations based on the video content and their contextual relevance, according to predefined scoring criteria.

F. Additional Qualitative Analysis

In Table 2, we presents a qualitative comparison of maneuver predictions and visual-textual explanations across Brain4Cars, AIDE, and DAAD datasets using DriveXplain , Qwen 2.5-VL-ED and Video-LLaMA-ED. Despite receiving the same

visual inputs, each model demonstrates varying explanation and maneuver accuracy. In particular, DriveXplain often aligns less accurately with ground-truth maneuver classes, while ED models, particularly VideoLLaMA, produce more contextually grounded and maneuver-consistent predictions. The visual rows further highlight how ED enhances alignment between predicted actions and corresponding visual semantics.

In Table 3, we present a qualitative comparison between DriveXplain and Qwen 2.5-VL on Brain4Cars and DAAD datasets under zero-shot settings (see Sec: Can VLMs be used for generating explanations?). While Qwen2.5-VL fails to predict the correct maneuver in both cases, misclassifying the right turn as forward or lane change, DriveXplain accurately identifies the maneuver and provides grounded explanations. Its outputs reference relevant visual cues such as intersections, road curvature, and surrounding vehicles, reflecting a stronger alignment with driving semantics.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 1

[3] Ashesh Jain, Hema S. Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *ICCV*, 2015. 2, 4

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 2023. 1

[5] Abdul Wasi, Shankar Gangisetty, Shyam Nandan Rai, and C. V. Jawahar. Early anticipation of driving maneuvers. In *ECCV*, 2024. 2, 4

[6] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and et al. Qwen2.5 technical report. *CoRR*, 2024. 1

[7] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, and et al. AIDE: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *ICCV*, 2023. 2, 4

[8] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP (Demos)*, 2023. 1

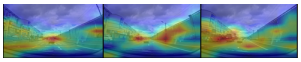
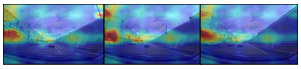
	Brain4Cars (GT: Left lane change)	AIDE (GT: Forward)	DAAD (GT: Left turn)
Original			
Qwen 2.5-VL-ED			
Video-LLaMA-ED			
Model	**Brain4Cars**	**AIDE**	**DAAD**
DriveXplain	**Predicted Maneuver:** Left lane change **Explanation:** The vehicle's movement into the left lane, without any indication of a turn or slowing down, is consistent with a lane change rather than a turn. The surrounding environment, including the trees and power lines, suggests a structured road layout that allows for deliberate lane changes.	**Predicted Maneuver:** forward **Explanation:** The ego vehicle maintains a consistent position in the middle of the highway, indicating a deliberate choice to avoid unnecessary lane changes or turns. The presence of surrounding vehicles and heavy traffic suggests that the driver is prioritizing safe distance and smooth flow	**Predicted Maneuver:** left turn **Explanation:** The ego vehicle prepares for and executes a left turn maneuver, as indicated by its changing lane position, interaction with surrounding vehicles, and alignment with road features such as intersections and turn indicators. These elements suggest a planned turn to the left based on traffic flow and navigation intent.
Qwen 2.5-VL-ED	**Predicted Maneuver:** Left lane change **Explanation:** The vehicle transitions left within a residential road, likely to adjust positioning or prepare for a turn. The lateral movement is evident from the vehicle's trajectory, and the environment with parked cars and roadside features—supports a controlled lane change.	**Predicted Maneuver:** Forward **Explanation:** The vehicle appears to be continuing straight on a multi-lane road in moderate traffic, with no indication of turning or lane-changing. The presence of commercial buildings and overcast weather suggests typical urban driving conditions	**Predicted Maneuver:** Left turn **Explanation:** The ego vehicle prepares for a left turn, navigating through urban traffic with multiple vehicles and pedestrians. The maneuver is consistent with a turn into a perpendicular route, supported by city street context and vehicle alignment.
Video-LLaMA-ED	**Predicted Maneuver:** Left lane change **Explanation:** The ego vehicle performs a deliberate shift to the adjacent left lane, which is evident from its lateral movement relative to lane markings and nearby vehicles. This maneuver appears to be made for better positioning or to overtake a slower vehicle ahead, consistent with typical left lane change behavior in urban driving scenarios.	**Predicted Maneuver:** Forward **Explanation:** The vehicle maintains a steady trajectory through a city street without any observable lane changes or turns. Its consistent speed and lane position, along with typical urban surroundings, indicate the driver's intent to continue moving forward.	**Predicted Maneuver:** Left turn **Explanation:** The ego vehicle initiates a left turn at an intersection, as indicated by the shift in vehicle orientation and trajectory. The maneuver aligns with surrounding traffic patterns and road layout, confirming the leftward deviation from the original path.

Table 2. Qualitative analysis of predicted maneuvers and visual explanations for DriveXplain (*ours*), Qwen 2.5-VL-ED (*ours*), and Video-LLaMA-ED (*ours*) models on Brain4Cars [3], AIDE [7], and DAAD [5] datasets.

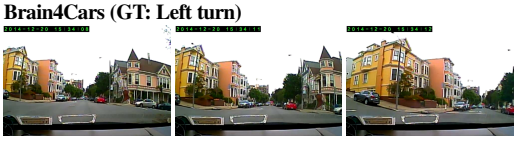
Model	Brain4Cars (GT: Left turn)	DAAD (GT: Left lane change)
		
Qwen 2.5-VL	**Predicted Maneuver:** Forward **Explanation:** The vehicle is moving straight down the street, as indicated by the consistent direction of the road and the lack of any sharp turns or lane changes.	**Predicted Maneuver:** Forward **Explanation:** The vehicle continues moving straight ahead on a busy urban street with moderate traffic and pedestrians.
Dolphins	**Predicted Maneuver:** Forward **Explanation:** The scene captures a highway segment with multiple vehicles traveling in both directions, viewed from a forward-facing camera mounted inside a moving vehicle."	**Predicted Maneuver:** Forward **Explanation:** The video captures an urban road scene with diverse traffic, including cars, motorcycles, and auto-rickshaws. The camera, positioned inside a moving vehicle, progresses forward, documenting both oncoming and forward-moving traffic
DriveXplain	**Predicted Maneuver:** Left turn **Explanation:** The vehicle's turning motion is consistent with a left turn, as it involves a significant change in direction. The surrounding environment, including traffic signals and signage, also supports the intention to turn left."	**Predicted Maneuver:** Left lane change **Explanation:** The vehicle transitions from the right to the left lane during the sequence, indicating a leftward lane change. While detailed behaviors of nearby vehicles aren't explicitly stated, the presence of other vehicles and pedestrians suggests potential interactions. Urban elements like street lights and possible road features are observed, even if traffic signs aren't clearly mentioned. The directional shift to the left reflects a deliberate maneuver, and the movement remains smooth and stable throughout.

Table 3. Qualitative comparison of predicted maneuvers and explanations for DriveXplain and Qwen 2.5-VL in zero-shot setting across Brain4Cars and DAAD datasets.