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

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

Abstract

Predicting a drivers' intent (e.g., *turns, lane changes) is a critical capability for modern Advanced Driver Assistance Systems (ADAS). While recent Multimodal Large Language Models (MLLMs) show promise in general vision-language tasks, we find that zero-shot MLLMs still lag behind domain-specific approaches for Driver Intention Prediction (DIP). To address this, we introduce DriveXplain, a zero-shot framework based on MLLMs that leverages rich visual cues such as optical flow and road semantics to automatically generate both intention maneuver (what) and rich natural language explanations (why). These maneuver–explanation pairs are then distilled into a compact MLLM, which jointly learns to predict intentions and corresponding explanations. We show that incorporating explanations during training leads to substantial gains over models trained solely on labels, as distilling explanations instills reasoning capabilities by enabling the model to understand not only what decisions to make but also why those decisions are made. Comprehensive experiments across structured (Brain4Cars, AIDE) and unstructured (DAAD) datasets demonstrate that our approach achieves state-of-the-art results in DIP task, outperforming zero-shot and domain-specific baselines. We also present ablation studies to evaluate key design choices in our framework. This work sets a direction for more explainable and generalizable intention prediction in autonomous driving systems. Project webpage:* https://avijit9.github.io/DriveXplain/

1. Introduction

ADAS are essential for improving driver safety and enabling autonomous driving [24] that supports drivers in navigation and managing safety-critical situations. A key emerging capability is predicting a driver's intention [13] before the execution of a maneuver (*e.g.*, lane changes, turns, slow-stop), enabling timely interventions and collision avoidance. Though prior DIP methods such as CNN-LSTM [10, 32] and Transformer-based [20, 45] primarily focused on accurately predicting *what* maneuver the driver will take. They lacked the ability to explain *why* that maneuver was chosen.

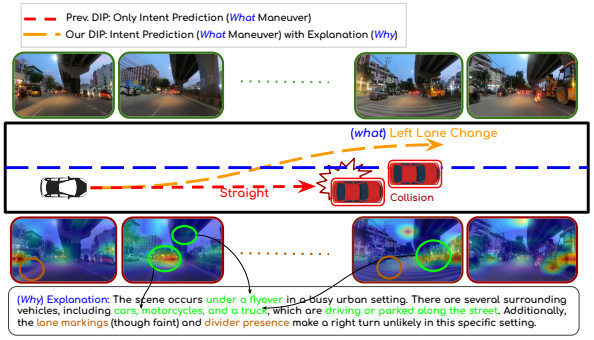

Figure 1. Illustration of a driving scenario where the ADAS vehicle predicts a left lane change (*what*) to avoid slower traffic ahead (*why*). Existing DIP [10, 20, 32, 45] models lacking reasoning may miss such cues, while our framework jointly learns and distills both maneuver and explanation, improving decision quality.

Without accompanying reasoning, such predictions lack contextual grounding, limiting their reliability in real-world scenarios, particularly in ambiguous or safety-critical situations.

Recently, MLLMs [1, 4, 6, 7, 12, 17, 34, 50, 52, 53] have emerged as a powerful and general-purpose solution for a wide range of computer vision tasks, including image captioning [17], visual question answering [28], and video understanding [53]. These models combine vision encoders [36, 39] with language decoders [3, 38], allowing them to interpret and reason visual inputs through natural language. Due to their strong capabilities in visual understanding and natural language generation, MLLMs have become a natural choice for complex tasks like DIP. Despite their generalization strengths, zero-shot MLLMs often struggle in DIP settings because they lack the domain-specific context and driving-relevant cues necessary for accurate intention prediction. This shortfall limits their performance compared to specialized DIP models [10, 20, 32, 45] and prevents them from fully utilizing their reasoning potential in safety-critical driving scenarios (refer Sec. 4).

To address this gap, we introduce *DriveXplain*, a zero-shot framework that leverages a novel prompting strategy with specialized visual inputs such as optical flow, road semantics and surrounding context extracted from raw videos to provide driving-specific contextual cues to an LLM (*e.g.*, LLaMA-3.1-8B [9]), enabling accurate prediction of driving maneuvers. Moreover, DriveXplain is also capable of generating explanations that accompany

*Both authors have contributed equally to this research.

its predicted maneuvers. We empirically demonstrate that DriveXplain achieves state-of-the-art results despite not being trained on any driving video data. However, the approach is somewhat impractical for real-world deployment due to its large model size and dependence on specialized visual inputs extracted from raw videos.

To this end, we propose a knowledge distillation [11] approach that transfers the knowledge acquired by DriveXplain into a smaller, more efficient MLLM (*e.g.*, Video-LLaMA [53], Qwen2.5-VL [4]). Additionally, to transfer the reasoning capabilities of DriveXplain, we distill its generated explanations into a smaller MLLM. Consider an illustration in Figure 1, where an autonomous vehicle performs a left lane change maneuver (*what*) to avoid slower traffic ahead by utilizing a clear left lane (*why*). Capturing such reasoning can enhance safety, reduce mispredictions, and improve generalization in diverse scenarios. Existing DIP methods [10, 20, 32, 45] focus primarily on predicting *what* maneuver occurs, often overlooking the reasoning of *why* that maneuver was taken. These models, typically built on modality-specific encoders and sequential architectures, are limited in their ability to capture broader context, multi-agent dynamics, or causal factors, hindering their capacity to jointly predict and explain drivers' intent. In summary, our framework DriveXplain jointly generates maneuvers and corresponding explanations, which are distilled into a single MLLM [4, 53]. These distilled models, *Video-LLaMA-ED* and *Qwen2.5-VL-ED*, capture both decisions and their rationale, offering improved inference efficiency (refer Table 1) and scalability for DIP.

The main contributions of our work are:

1. We are the first to conduct a comprehensive evaluation of multiple general-purpose and driving-specific MLLMs [6, 7, 17, 21, 34, 42, 50, 53] on the DIP task [45], experimentally demonstrating that these MLLMs consistently fail to reliably predict drivers' intentions.
2. We introduce DriveXplain, a framework that enhances maneuver prediction by incorporating driving-specific contextual information.
3. We present an explanation-guided distillation strategy to transfer both the *what* (maneuver) and the *why* (reasoning) from DriveXplain into a single, unified MLLM for enhanced maneuver prediction.
4. We perform extensive quantitative and qualitative evaluations of the MLLMs on structured (Brain4Cars [13], AIDE [49]) and unstructured (DAAD [45]) DIP datasets. Our method consistently outperforms prior approaches, demonstrating the effectiveness of proposed framework.

2. Related Work

2.1. Driver Intention Prediction

Traditional DIP methods have relied on bidirectional RNNs [26] and CNN-LSTM architectures [10, 13, 14, 16, 32], that primarily focus on spatial features while offering limited temporal modeling. This restricts their effectiveness in capturing long-range dependencies essential for accurate intent prediction. To overcome these limitations, transformer-based architectures [40] were introduced, offering improved performance by modeling long-term temporal context. Recently memory-augmented methods such as CEMFormer [20] and M^{2}MVIT [45] have further enhanced temporal consistency and anticipation robustness. Despite these advances, existing DIP methods treat the task as a classification problem, predicting *what* maneuver will be performed without addressing *why* that maneuver is expected. This lack of explanation limits their applicability in safety-critical, real-world ADAS scenarios. To address this limitation, we propose a DIP framework that improves maneuver prediction and provides human-understandable reasoning.

2.2. MLLMs for Driving

Recent progress in MLLMs, encompassing both closed-source models [1, 12] and their open-source counterparts [4, 6, 7, 17, 34, 50, 52, 53], has significantly advanced the state of computer vision, including domain-specific tasks such as autonomous driving [21]. These models leverage large-scale pretraining on image-text or video-text pairs to capture spatial, semantic, and, in some cases, temporal dependencies. The integration of LLMs into vision tasks has enabled more expressive multimodal interactions and strengthened visual understanding capabilities. Though powerful in multimodal understanding, their ability to generalize to fine-grained, causal reasoning tasks remains limited [28] and largely unexplored especially in the context of DIP.

Recent works have begun incorporating reasoning as additional modality, to enhance model performance [19, 25, 30, 44, 46]. In autonomous driving [8, 55], this has led to applications such as driving captioning [2, 15], question answering [23, 31, 35, 43], conversational driver assistants [29], driving actions [22], as well as scene understanding and planning [37, 47]. However, these models mainly focus on *what* is the maneuver and rarely address *why* specific driving decisions are made. In contrast, we propose a unified framework that jointly models both the maneuver prediction (*what*) and its underlying rationale (*why*).

2.3. Knowledge Distillation

Knowledge distillation (KD) [11] has been widely adopted to compress large teacher models into smaller, efficient student models to improve the efficiency of inference. Distillation approaches can generally be categorized into three types, namely, logit-based [5], feature-based [51], and explanation-based [27]. Logit-based KD [5] aligns the output distributions (logits) of the teacher and student models. Feature-based KD [51], on the other hand, encourages the student to mimic the teacher's intermediate activation maps. Recently, explanation-based distillation has gained traction; for example, Parchami et al. [27] proposed a simple, parameter-free method where the student is trained

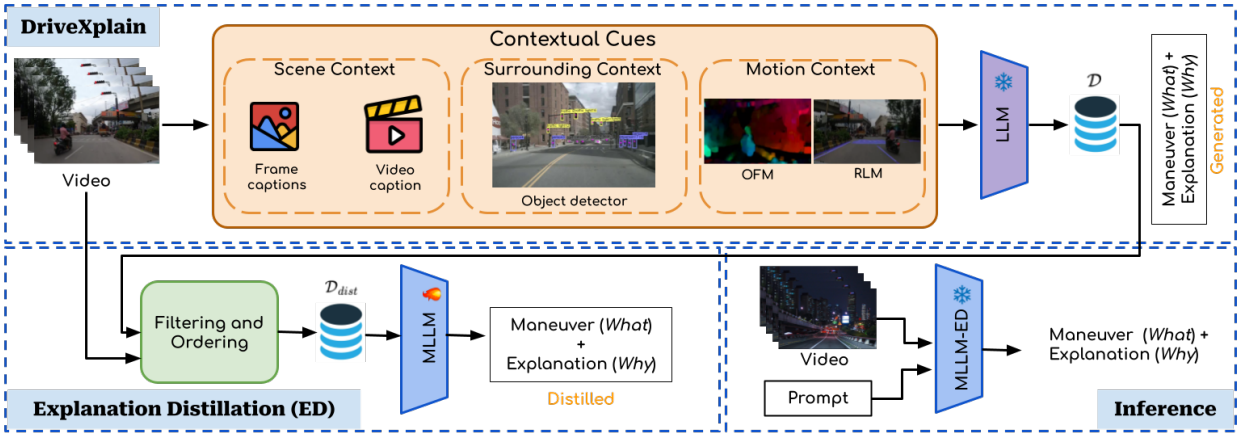

Figure 2. **Our proposed framework for the DIP task. DriveXplain** generates natural language explanations alongside maneuvers and **Explanation Distillation** distills these explanations into a single MLLM to enhance DIP performance at **inference**.

to produce explanations similar to those of the teacher.

In the context of MLLMs, prior work has explored distilling Chain-of-Thought (CoT) [46] reasoning into smaller MLLMs [19]. However, the challenge of distilling specialized visual inputs such as those required in driving scenarios remains largely underexplored. In particular, existing methods do not address how to compress large MLLMs that rely on rich, domain-specific visual inputs into smaller models capable of both accurate prediction and reasoning. In this work, we address this gap by distilling both the intended maneuver (*what*) and the corresponding natural language explanations (*why*), generated by a large, zero-shot MLLM framework into a single, compact MLLM trained to jointly predict what the driver will do and why.

3. Methodology

We propose a DIP framework that goes beyond classifying driver maneuvers (*what*) from video, addressing limitations of prior approaches [10, 20, 32, 45]. Our framework not only predicts driver maneuvers (*what*) but also generates natural language explanations (*why*) to justify each decision, thus enhancing transparency, trust, and applicability in safety-critical driving scenarios. Given a video input $V \in \mathbb{R}^{T\times H\times W\times C}$ consisting of T frames from a driving perspective, the objective is to predict the driver's maneuver $\hat{\mathbf{y}} \in Y$, where Y represents the set of possible maneuver classes (e.g., turns, lane changes, straight, slow-stop), and simultaneously generate a explanation $\mathbf{E}$ that reasons the predicted maneuver.

Our proposed framework as shown in Figure 2 consists of two key stages:

- **DriveXplain** (§ 3.1): A 15B-parameter model that generates maneuver predictions and corresponding explanations using a novel structured prompting strategy in a zero-shot manner. Although DriveXplain delivers improved predictions and explanations, its large model size poses challenges for efficient inference, especially in latency-sensitive or resource-constrained settings.
- **Explanation Distillation (ED)** (§ 3.2): To mitigate the above limitation, we introduce an Explanation Distillation stage, where the reasoning capabilities of DriveXplain are compressed into a single 7B MLLM. This distilled model retains predictive and explanatory performance while enabling faster and more scalable inference.

3.1. DriveXplain

One straightforward approach to DIP is to prompt existing MLLMs [6, 7, 17, 34, 42, 50, 53] in a zero-shot setting. However, as we demonstrate empirically in § 4.2, these MLLMs struggle to accurately predict driver intentions, even when pre-trained on driving-related video data. This limitation can be primarily attributed to their lack of access to structured contextual information, such as temporal cues, traffic semantics, and driver-centric observations, which are critical to reliably anticipating maneuvers. In this work, we aim to address the core question: *How can we effectively provide relevant driving context to MLLMs to enable accurate DIP?*

Given an input video V, we uniformly sample a set of T frames $F = \{f_1, f_2, \cdots, f_T\}$. The DriveXplain stage comprises three key modules: scene context, surrounding context, and motion context, detailed below.

3.1.1. Scene Context

Captures the broader environmental and semantic context of the driving scene by combining high-level video captions that summarize the overall scene (e.g., "*approaching a busy urban intersection*") with fine-grained frame-level captions that provide temporal details (e.g., "*traffic light turns green*", "*pedestrian crossing ahead*"). This fusion enables the model to reason about road layout, static infrastructure, traffic signs, and temporal scene evolution, key elements for accurately anticipating driver maneuvers. We use a MLLM ($\mathcal{M}$) to extract frame-wise captions $\mathcal{C}_f = \{\mathcal{M}(f_1), \mathcal{M}(f_2), \cdots, \mathcal{M}(f_T)\}$ and a video-level caption $\mathcal{C}_v = \mathcal{M}(V)$. The prompt for $\mathcal{M}$ is provided in the supplementary. The scene context

information is aggregated into $\mathcal{C}_{sc}$ which is defined as:

$$\mathcal{C}_{sc} = \{\mathcal{C}_f, \mathcal{C}_v\} \tag{1}$$

3.1.2. Surrounding Context

To better capture the driver's interaction with surrounding objects, while objects such as *traffic lights* may be identified in scene context, the underlying behavioral cues, such as *slowing down* or *preparing to stop*, are often not explicitly modeled. To address this, we sample frames uniformly from each video. For each frame f_t, we apply an object detection model $\mathcal{O}$ to extract vehicle surrounding cues. The model returns a set of detections, where each detection consists of an object class $\mathbf{o}_{t,i}$, a confidence score $\mathbf{c}_{t,i} \in [0,1]$, and a 2D position $\mathbf{p}_{t,i} \in \mathbb{R}^2$. The surrounding context information ($\mathcal{C}_{src}$) of the video is represented as:

$$\mathcal{C}_{src} = \left\{ \{\mathbf{o}_{t,i}, \mathbf{c}_{t,i}, \mathbf{p}_{t,i}\}_{i=1}^{N} \right\}_{t=1}^{T}. \tag{2}$$

3.1.3. Motion Context

Captures ego-vehicle's dynamic behavior over time, and reflects patterns such as lateral shifts (e.g., *left-to-right*) and abrupt directional changes signaling maneuvers. These cues carry essential semantic information that conventional language models and MLLMs are not equipped to handle directly. To address this, we use two modules: Optical Flow Masking (OFM), which highlights motion regions via pixel-wise frame displacements, and Road and Lane Masking (RLM), which isolates road and lane boundaries to provide spatial context for vehicle movement.

Optical Flow Masking. To encode spatial layout and motion dynamics we compute the optical flow between each pair of consecutive frames. Let I_t and I_{t+1} denote two such frames from video V. The optical flow map is denoted as $\mathbf{F}_t \in \mathbb{R}^{H \times W \times 2}$, where each vector $\mathbf{F}_t(x, y)$ captures the horizontal and vertical displacement at pixel location (x, y). We divide each optical flow map into P non-overlapping patches of size $p \times p$. For each patch i in frame t, we compute the average flow vector as:

$$\vec{v}_{t,i} = \frac{1}{p^2} \sum_{(x,y) \in \text{patch}_i} \mathbf{F}_t(x, y). \tag{3}$$

To model directionality $\mathbf{d}_{t,i}$ for each patch i for frame f_t, we define a motion label based on $\vec{v}_{t,i}$. The direction label is defined as:

$$\mathbf{d}_{t,i} = \begin{cases} \text{left} & \text{if } \vec{v}_{t,i} < -\theta, \\ \text{right} & \text{if } \vec{v}_{t,i} > \theta, \\ \text{none} & \text{otherwise,} \end{cases} \tag{4}$$

where θ is a predefined threshold for horizontal motion. In practice, we consider only the horizontal component of $\vec{v}_{t,i}$, as vehicle movement occurs primarily in the lateral direction on the road.

Road and Lane Masking. We also obtain semantic segmentation maps for each frame f_t, denoted as $\mathbf{S}_t \in \{0,1,2\}^{H \times W}$, where 0 denotes background, 1 indicates road, and 2 represents lane markings. For each patch i of frame f_t, a semantic label is assigned as:

$$\mathbf{s}_{t,i} = \begin{cases} \text{lane} & \text{if majority of pixels are labeled 2,} \\ \text{road} & \text{if majority are labeled 1,} \\ \text{road and lane} & \text{if significant mix of 1 and 2,} \\ \text{none} & \text{otherwise} \end{cases} \tag{5}$$

Thus, each patch is described by a tuple $\{\mathbf{d}_{t,i}, \mathbf{s}_{t,i}\}$, which captures both spatial and directional semantics. These patch-level descriptors are aggregated across the sampled frames F, forming the structured representation as:

$$\mathcal{C}_{mc} = \left\{ \{\mathbf{d}_{t,i}, \mathbf{s}_{t,i}\}_{i=1}^{P} \right\}_{t=1}^{T}. \tag{6}$$

The contextual cues ($\mathcal{C}_{sc}$,$\mathcal{C}_{src}$,$\mathcal{C}_{mc}$) are text-based representations derived from visual inputs, providing driving-specific context. These are fed into an LLM to infer the driver's intent and predict the maneuver $\hat{\mathbf{y}} \in Y$. A sample prompt is illustrated in Figure 3. Since LLM may not inherently understand the structure of our proposed motion cues $\mathcal{C}_{mc}$, such as OFM and RLM, we incorporate in-context examples within the prompt to facilitate better comprehension and interpretation.

3.2. Explanation Distillation

Our proposed framework, DriveXplain, operates in a zero-shot setting. Although DriveXplain demonstrates strong performance, its large size (15B parameters) and reliance on modality-specific inputs such as optical flow and lane masks limit its practicality for real-world deployment. To address this, we distill its knowledge into a more compact 7B MLLM for efficient inference.

In particular, when a training set of videos $(V, \mathbf{y})$ is available, we leverage DriveXplain to generate both the maneuver classes $\mathbf{y}$ and corresponding explanations $\mathbf{E}$ (refer supplementary for the prompt). This knowledge is then distilled into a MLLM to enable end-to-end prediction of driver intents directly from raw visual inputs. However, the generated explanations $\mathbf{E}$ may be noisy, as they are produced by the model and may not always be accurate. We identify two primary sources of noise: (a) incorrect maneuver predictions, which render the entire set of explanations unreliable (§ 3.2.1); and (b) partial inconsistency, where the predicted maneuver is correct, but some explanations are informative while others are flawed (§ 3.2.2).

3.2.1. Explanations Filtering

To ensure explanation quality, we retain only samples where the predicted maneuver $\hat{\mathbf{y}}$ matches the ground-truth $\mathbf{y}$ from the DIP dataset $\mathcal{D}$, discarding those with incorrect predictions and their associated explanations. We denote the filtered dataset as $\mathcal{D}_{clean}$, defined as:

$$\mathcal{D}_{clean} = \{(V, \mathbf{y}, \mathbf{E}) \in \mathcal{D} \,|\, \hat{\mathbf{y}} = \mathbf{y}\}. \quad (7)$$

3.2.2. Explanation Ordering

Filtering samples based on correct maneuver predictions reduces maneuver class noise but does not guarantee the quality of associated explanations. To further refine the dataset $\mathcal{D}_{clean}$, we use a frozen VLM $\mathcal{J}$ as a judge to evaluate the alignment between the video V, its ground-truth $\mathbf{y}$, and each candidate explanation $\mathbf{E}_i$, assigning a relevance score (s_i) that reflects explanation consistency, defined as:

$$s_i = \mathcal{J}(V, \mathbf{y}, \mathbf{E}_i). \quad (8)$$

To refine explanation alignment, we explored two strategies: numeric scoring and categorical ranking. Empirically, we observed that categorical ranking (e.g., "strongly supported", "moderately supported", "supported", "weakly supported", and "not supported") proved more stable and yielded better generalization than raw numerical scores. Explanations with strong alignment are retained, and the corresponding tuple $(V, \mathbf{y}, \mathbf{E}_i)$ is included in the distilled training set $\mathcal{D}_{dist}$, defined as:

$$\mathcal{D}_{\text{dist}} = \left\{ (V, \mathbf{y}, \mathbf{E}^*) \,|\, \mathbf{E}^* = \underset{i}{\operatorname{argmax}}\, s_i \right\}, \quad (9)$$

where $\mathbf{E}^* \in \mathbf{E}$ is the explanation with the highest alignment score to the context. Refer to the supplementary for detailed prompt examples.

3.2.3. Distillation

Finally, we distill the triplet $(V, \mathbf{y}, \mathbf{E}^*) \sim \mathcal{D}$ into a smaller MLLM $\mathcal{Q}_{\text{VLM}}$ [4, 53]. The model is trained via next-token prediction using a cross-entropy loss, with the objective of maximizing the likelihood of generating both the explanation and maneuver tokens conditioned on the video and prompt. By jointly training on the maneuver (*what* the driver's intention is) and the explanation (*why* that intention is inferred), the model is encouraged to internalize both the decision outcomes and the underlying reasoning process. We denote the models trained with Explanation Distillation as Video-LLaMA-ED and Qwen2.5-VL-ED.

3.3. Inference

During inference, we prompt the explanation distilled model $\mathcal{Q}_{\text{VLM}}$ with the question "*What is the maneuver being performed?*" and a test video. The model produces a unified output that contains both the predicted maneuver and its explanation. For evaluation, we extract the maneuver from this response and compare it with the ground truth.

4. Results and Analysis

4.1. Experimental Setting

Datasets. We evaluate our unified framework, comprising of DriveXplain and Explanation Distillation on two structured datasets, Brain4Cars [13] and AIDE [49], and one unstructured dataset, DAAD [45]. All performance evaluations are conducted on the respective test sets of these datasets. Notably, none of these datasets include ground-truth explanation annotations corresponding to the maneuver classes. Consequently, our evaluation focuses solely on intent classification metrics, and exclude metrics related to semantic explanation quality.

Basic instruction: Analyze the following multimodal inputs to classify the driving maneuver: Frame-level captions, video caption, surrounding context, RLM, and OFM.

Constraints: Respond strictly with one of the maneuvers: *right turn*, *right lane change*, *left turn*, *left lane change*, *forward*.

Task and label descriptions:

- **Right turn**: The vehicle makes a sharp or significant turn to the right.
- **Right lane change**: The vehicle shifts into the right lane.
- **Left turn**: The vehicle makes a sharp or significant turn to the left.
- **Left lane change**: The vehicle shifts into the left lane.
- **Forward**: The vehicle continues straight, completes a maneuver, or comes to a stop.

OFM example:{`example`} Label: Right turn
RLM example:{`example`} Label: Right lane

Multi-modal input representations:
Frame-level captions:$\{\mathcal{C}_f\}$; Video caption:$\{\mathcal{C}_{\text{v}}\}$;
Surrounding context:$\{\mathcal{C}_{\text{src}}\}$; OFM:$\{d_{t,i}\}$; RLM:$\{s_{t,i}\}$

Output: <maneuver>

Figure 3. Prompt for driving maneuver classification in § 3.1.

Models. We compare our framework with state-of-the-art (SOTA) models from video understanding and vision-language domains. From the DIP literature, we consider action anticipation models, including CNN-LSTM [10, 32], CEMFormer [20], and M^{2}MVT [45]. Additionally, we evaluate several general-purpose SOTA MLLMs in the zero-short setting, including InternVL [7], Mini-CPM-V 2.6 [50], LLaVA Next [17], Video-LLaMA [53], ShareGPT4 [6], Tarsier [42], and LongVU [34], as well as a driving-specific MLLM Dolphins [21].

Metrics. To evaluate DIP performance, we follow standard protocols from previous works [10, 20, 45], using accuracy and F_1-score as the primary evaluation metrics.

Implementation Details. To extract scene context $\mathcal{C}_{sc}$ (§ 3.1.1), we utilize Video-LLaMA [53] to obtain video-level captions ($\mathcal{C}_v$), and LLaVA [18] to generate frame-wise captions ($\mathcal{C}_f$). For vehicle detection and trajectory estimation, used to model the surrounding traffic context $\mathcal{C}_{src}$ (§ 3.1.2), we adopt CenterTrack [54]. To derive motion context $\mathcal{C}_{mc}$

Model	Params	FineTune	Brain4Cars [13]		AIDE [49]		DAAD [45]	
			Acc.	F_1	Acc.	F_1	Acc.	F_1
Dolphins [21]	9B	✗	40.25	38.73	58.67	55.64	30.45	28.99
InternVL2 [7]	8B	✗	34.57	27.54	60.24	60.55	9.39	23.84
Mini-CPM-V 2.6 [50]	8B	✗	35.82	31.36	57.34	59.06	22.14	12.86
LLaVA-NeXT [17]	7B	✗	10.86	2.16	37.34	69.37	24.83	9.87
Video-LLaMA [53]	7B	✗	38.60	23.71	67.79	62.07	22.14	9.20
ShareGPT4V [6]	7B	✗	20.87	19.89	55.79	55.72	20.80	18.28
Tarsier [42]	7B	✗	29.71	24.08	59.27	63.78	21.47	9.25
LongVU [34]	7B	✗	26.08	20.56	19.03	27.70	15.43	14.31
Qwen2.5-VL [4]	7B	✗	41.66	33.15	69.49	66.59	32.21	17.70
Video-LLaMA 3 [52]	7B	✗	25.73	23.91	57.90	59.37	26.17	21.01
Video-LLaMA [53]	7B	✓	44.00	41.67	49.34	37.47	36.48	52.08
Qwen2.5-VL [4]	7B	✓	48.77	39.90	71.38	70.49	39.74	40.92
Gebert et al. [10]	0.24B	✗	**72.89**	69.59	72.89	69.59	52.65	48.15
Rong et al. [32]	0.16B	✗	58.71	62.75	73.45	70.17	50.31	54.05
CEMFormer [20]	0.08B	✗	63.27	65.35	75.90	73.25	<u>58.87</u>	59.31
M^2MVT [45]	0.03B	✗	64.07	65.35	75.90	73.25	58.78	<u>59.91</u>
DriveXplain (Video-LLaMA)	15B	✗	<u>72.33</u>	71.34	<u>78.93</u>	53.14	52.52	44.31
DriveXplain (Qwen 2.5-VL)	15B	✗	64.49	52.96	52.22	51.83	42.28	46.96
Mobile-VideoGPT-ED [33]	0.5B	✓	60.31	56.97	68.93	66.38	50.30	46.00
Mobile-VideoGPT-ED [33]	1.5B	✓	61.85	60.09	71.49	71.92	52.10	47.33
Video-LLaMA-ED	7B	✓	71.24	<u>73.29</u>	**80.47**	**78.90**	57.66	54.03
Qwen2.5-VL-ED	7B	✓	72.28	**73.81**	77.80	<u>75.64</u>	**62.40**	**62.98**

Table 1. **DIP benchmark results.** Performance comparison of **Driving-specific VLM** , **General VLMs** , **Action Anticipation models** , and our framework (**DriveXplain** , **ED**). Accuracy (Acc.) and F_1(%) on Brain4Cars [13], AIDE [49], and DAAD [45] datasets. Finetune indicates whether the model was fine-tuned (✓) or evaluated in a zero-shot (✗) setting. **Bold** and <u>underline</u> is for best and second-best results.

(§ 3.1.3), we employ HybridNets [41] for road and lane segmentation ($s_{t,i}$), and compute dense optical flow ($d_{t,i}$) using the Farneback algorithm provided by `OpenCV`. For all datasets, video frames are uniformly sampled at 1FPS to ensure consistent and descriptive scene representations. The language model responsible for predicting maneuvers and generating (maneuver, explanation) pairs is LLaMA-3.1 (8B) [9]. We use Qwen2.5-VL [4] for as the judge $\mathcal{J}$ in explanation ordering (§ 3.2.2). All components in our pipeline, including VLMs and LLMs, operate in a zero-shot setting, except during the Explanation Distillation stage, where fine-tuning is conducted. Further details on the distillation procedure and hyperparameters for Video-LLaMA [53] and Qwen2.5-VL [48] are provided in the supplementary.

4.2. Comparisons with State-of-the-Arts

Zero-shot MLLMs. First, we evaluate a range of MLLMs, including both general-purpose and driving-specific models, as summarized in Table 1. In particular, we include MLLMs such as InternVL2 [7], Mini-CPM-V 2.6 [50], LLaVA-NeXT [17], Video-LLaMA [53], ShareGPT4 [6], Tarsier [42], LongVU [34] and the driving-specific model Dolphins [21] to assess their zero-shot performance on DIP. All models are evaluated using their default inference settings without task-specific fine-tuning. While many models demonstrate inconsistent maneuver classification performance (i.e., answering *what*), Qwen2.5-VL [4] shows reasonable performance over other models across Brain4Cars [13], AIDE [49], and DAAD [45] datasets. This can be attributed to their accurate object grounding and the use of dynamic FPS sampling.

Effectiveness of DriveXplain. Second, we evaluate our proposed DriveXplain by comparing it against state-of-the-art MLLMs and DIP methods. Notably, DriveXplain operates in a zero-shot setting and does not require any training data. We summarize these results in Table 1. Our DriveXplain significantly outperforms both general-purpose MLLMs [6, 7, 17, 34, 42, 50, 53] and existing action anticipation-based DIP models [10, 20, 32, 45] both in zero-shot and finetuned settings. Specifically, using Qwen2.5-VL [4], DriveXplain achieves accuracy improvements of 16% on Brain4Cars [13], 19% on AIDE [49], and 3% on DAAD [45] compared to general-purpose finetuned models. Similarly, with finetuned Video-LLaMA [53], it surpasses these models by 28%, 37%, and 16% on the respective datasets. While its performance is comparable to specialized DIP models [10, 20, 32, 45] for maneuver classification, DriveXplain does not require any training data and operates in zero-shot settings.

The strong performance gains of our DriveXplain framework stem from its effective use of driving-specific contextual information. By combining visual and geometric cues into high-level MLLM-generated descriptions, our model is better equipped to recognize complex patterns for accurate intent prediction. In contrast, existing MLLMs often default to predicting the '*forward*' class due to their limited ability to incorporate low-level geometric and directional information, which is essential for reliable driver intention prediction.

Effectiveness of ED. Third, we evaluate the impact of

OFM	RLM	$\mathcal{C}_{src}$	Acc. (%)	F_1 (%)
✗	✗	✗	39.49 -	24.55 -
✗	✓	✓	48.00 ↑8.51%	57.87 ↑33.32%
✓	✗	✓	56.36 ↑16.87%	54.27 ↑41.42%
✓	✓	✗	65.45 ↑25.96%	65.97 ↑168.7%
✓	✓	✓	**72.33** ↑32.84%	**71.34** ↑46.79%

Table 2. **Component-level ablation.** Significance of modules (OFM, RLM, and $\mathcal{C}_{src}$) for our framework on Brain4Cars [13] dataset. Percentage gains are shown relative to the first row.

ED by comparing MLLMs fine-tuned with (maneuver, explanation) pairs against zero-shot baselines. Specifically, we distill both the what (maneuver) and why (explanation) from our DriveXplain into best performing smaller models such as Video-LLaMA and Qwen2.5-VL, referred to as Video-LLaMA-ED and Qwen2.5-VL-ED in Table 1. Models trained only on maneuver classes consistently underperform, highlighting the limitations of conventional supervision, especially with the limited size of DIP datasets. In contrast, ED improves performance across all benchmarks by providing richer supervision.

Our fine-tuned models achieve state-of-the-art results, surpassing prior work by 4.57% on AIDE [49] and 3.53% on DAAD [45]. They also perform competitively on Brain4Cars [13] and outperform DriveXplain by 1.95% on AIDE and 9.88% on DAAD. To further assess the effectiveness of our knowledge distillation, we evaluate the recent lightweight Mobile-VideoGPT [33]. The distilled variants, Mobile-VideoGPT-ED (0.5B and 1.5B), consistently underperform, whereas the 7B model achieves clearly superior results across all datasets. This demonstrates that while smaller models retain part of the teacher's capability, a substantial performance gap remains, underscoring the importance of model capacity in effective distillation and aligning with scaling laws.

Additionally, we also measure the inference latency of different models on Brain4Cars [13]. Among them, Qwen2.5-VL-ED (7B) achieves the lowest latency at 329.77 $\pm$ 55 ms per video, while InternVL2, LLaVA-NeXT, and Video-LLaMA3 incur substantially higher latencies of 634 $\pm$ 95 ms, 639 $\pm$ 85 ms, and 524 $\pm$ 70 ms, respectively.

4.3. Ablation Experiments

We conduct ablation studies to evaluate three key design choices in our framework: (i) the influence of contextual cues such as optical flow and lane masks (Table 2), (ii) the effect of varying VLM and LLM configurations within DriveXplain (Table 3), and (iii) the ability of VLMs to independently generate explanations (Table 5).

Effects of Contextual Cues. To evaluate key components such as OFM, RLM, and surrounding context ($\mathcal{C}_{src}$) we perform ablations as shown in Table 2. Using only video- and frame-level captions results in frequent prediction of *forward* maneuver, indicating limited temporal and directional understanding. Adding optical flow significantly improves the prediction of *turn-related maneuvers* by capturing coarse directional motion across frames with surrounding context ($\mathcal{C}_{src}$) by 8.51% compared to the baseline. However, the model still struggles with finer maneuvers such as *lane changes*, where motion cues are subtle. Inclusion of lane masks helps address this by providing structural layout information, enhancing the model's ability to detect lateral movement with surrounding context ($\mathcal{C}_{src}$) by 8.36%. Furthermore, incorporating optical flow masks (OFM) along with road and lane masks (RLM) outperforms the previous two configurations by 9.09%, though it still does not achieve the highest performance. The addition of surrounding context provides high-level reasoning capabilities, enabling more accurate inference of maneuvers such as *U-turn* and *slowdown*, which require broader scene understanding. The full framework, integrating VLM-generated descriptions with OFM, RLM, and $\mathcal{C}_{src}$, achieves best performance, showning a gain of 32.84% improvement on Brain4Cars [13] dataset. Thus highlighting the complementary nature of these components.

Model	Params	Accuracy
VLM-wise comparison with LLaMA 3.1 [9] fixed		
LLaVA-Next Video [17]	7B	56.88
ShareGPT4V [6]	7B	51.44
Video-LLaMA [53]	7B	**72.33**
LLM-wise comparison with Video-LLaMA [53] fixed		
Qwen2.5 [48]	7B	64.85
LLaMA-3.1 [9]	8B	**72.33**

Table 3. **Proposed framework performance with different VLMs and LLMs.** Accuracy is reported for each model.

Performance comparison of large models in our framework. We evaluate the influence of model architecture by analyzing the VLM and LLM components independently (see Table 3). In the VLM comparison, Video-LLaMA [53] shows superior performance, surpassing LLaVA-Next Video [17] by 15% and ShareGPT4V [6] by 20% in accuracy. For the LLM comparison, with Video-LLaMA fixed as the visual encoder, LLaMA-3.1 [9] achieves 8% higher accuracy than Qwen2.5 [48]. These results suggest that the VLM contributes more significantly to downstream performance. However, the best accuracy is obtained when both VLM and LLM are strong and compatible, as shown by the combination of Video-LLaMA and LLaMA-3.1. This highlights the importance of temporal video comprehension in the visual module and strong reasoning capabilities in the language model for effective multimodal learning.

Can other MLLMs serve as Judge? To assess the impact of the judge model on data quality for distillation, we conduct an ablation study as shown in Table 4. The results show that Qwen2.5-VL [4] serves as a much stronger judge compared to LLaVA-NeXT [17] and Video-LLaMA3 [52], both of which exhibit notable performance drops. This highlights the critical role of the judge model, with

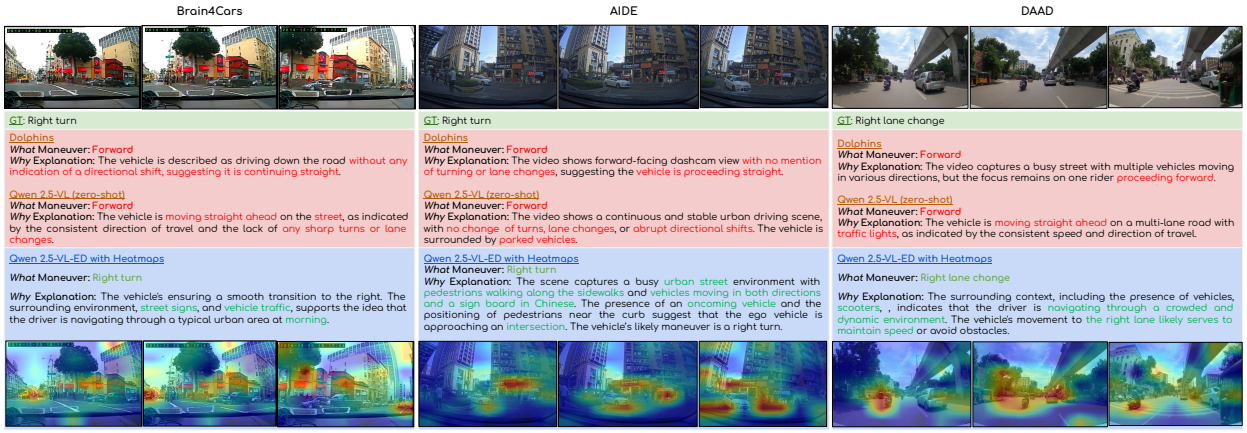

Figure 4. **Qualitative comparison of proposed framework, zero-shot Qwen2.5-VL [4], Dolphins across Brain4cars [13], AIDE [49], and DAAD [45] datasets.** We show maneuver prediction (*what*) and explanation (*why*), with attention heatmaps highlighting key regions.

Judge Model	Params	Finetuning Accuracy	
		Video-LLaMA-ED	Qwen2.5-VL-ED
LLaVA-NeXT [17]	7B	61.47	64.55
Video-LLaMA3 [52]	7B	68.99	70.37
Ours (Qwen2.5-VL [4])	**7B**	**72.28**	**71.24**

Table 4. **Performance comparison of ED on Brain4Cars [13] using different MLLMs as judge models for filtering.**

Model	Brain4Cars	AIDE	DAAD
Qwen2.5-VL [4]	12.31	22.89	24.83
Dolphins [21]	13.31	31.82	23.76
DriveXplain (Video-LLaMA)	**52.89**	**44.31**	<u>40.93</u>
DriveXplain (Qwen2.5-VL)	<u>43.11</u>	<u>42.14</u>	**44.29**

Table 5. **DriveXplain results on zero-shot VLMs.** Maneuver prediction accuracy is reported in %. **Bold** and <u>underline</u> indicate best and second-best results.

Qwen2.5-VL [4] establishing a more reliable benchmark.

Can MLLMs directly generate explanations? Our analysis reveals that existing MLLMs, domain-specific [21] or best generic model [4], they are inadequate for producing high-quality, causally grounded explanations in driving scenarios. When tasked with generating maneuver intent and corresponding explanations, these models tend to produce generic and temporally shallow descriptions that lack the fine-grained motion and contextual semantics critical necessary for reliable decision-level understanding. As shown in Table 5, this leads to degraded performance across all DIP datasets. These findings suggest that off-the-shelf MLLMs, without additional reasoning or structured inputs, are not well-suited for generating actionable explanations in downstream tasks. Examples are provided in supplementary.

4.4. Qualitative Analysis

As shown in Figure 4, our framework comprising of Explanation Distillation consistently produces more accurate and semantically grounded maneuver predictions across all datasets. On structured datasets such as Brain4Cars [13] and AIDE [49], driving-specific Dolphins and generic Qwen2.5-VL models often mispredict the maneuver as a *forward* maneuver when the ground truth is a *right turn*. Their generated explanations lack spatial awareness, due to the absence of directional cues. In contrast, our framework correctly predicts a right turn (answering *what*) and supports it with explanations (answering *why*) grounded in scene semantics such as the presence of *street signs*, *traffic flow*, and *surrounding urban environment* structures. The corresponding attention heatmaps localize relevant contextual regions such as *intersection entry points* and *curb directions*, further reinforcing model focus. A similar observation holds for the DAAD [45] dataset, further reinforcing the efficacy of our proposed approach.

5. Conclusion

We present a novel approach to DIP that extends beyond traditional action anticipation by jointly predicting a driver's maneuver (*what*) and providing natural language explanations (*why*) for that action. This dual prediction improves behavior understanding, supports reasoning, and enhances decision-making in safety-critical autonomous driving scenarios. To enable this, we propose DriveXplain, a zero-shot MLLM-based framework that integrates multiple visual cues such as scene context, surrounding environment, motion dynamics (optical flow), and road semantics to generate maneuver–explanation pairs. These pairs are distilled into a compact model that learns both intention and explanation in an end-to-end manner. This improves both reasoning ability and inference efficiency. Our method extensive quantitative and qualitative evaluations across structured and unstructured driving datasets show consistent performance improvements, underscoring the benefit of explanation-driven supervision for intent prediction. We hope our work advances the development of safer, more transparent, and explainable driver assistance systems.

Acknowledgments

The project was supported by iHub-Data and Mobility at IIIT Hyderabad.

References

[1] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, 2023.

[2] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *WACV*, 2025.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv:2309.16609*, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and et al. Qwen2.5-vl technical report. *arXiv:2502.13923*, 2025.

[5] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *CVPR*, 2022.

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, and et al. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024.

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[8] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, and et al. A survey on multimodal large language models for autonomous driving. In *WACV (Workshops)*, 2024.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. The llama 3 herd of models. *CoRR*, 2024.

[10] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *IV*, 2019.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPSW*, 2014.

[12] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and et al. Gpt-4o system card. *CoRR*, 2024.

[13] 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.

[14] Ashesh Jain, Avi Singh, Hema Swetha Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *ICRA*, 2016.

[15] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. ADAPT: action-aware driving caption transformer. In *ICRA*, 2023.

[16] Nima Khairdoost, Mohsen Shirpour, Michael A. Bauer, and Steven S. Beauchemin. Real-time driver maneuver prediction using LSTM. *IEEE Trans. Intell. Veh.*, 2020.

[17] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, 2024.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[19] Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. VITED: video temporal evidence distillation. In *CVPR*, 2025.

[20] Yunsheng Ma, Wenqian Ye, Xu Cao, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, and Ziran Wang. Cemformer: Learning to predict driver intentions from in-cabin and external cameras via spatial-temporal transformers. In *ITSC*, 2023.

[21] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *ECCV (45)*, 2024.

[22] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with GPT. *CoRR*, 2023.

[23] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoît Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and et al. Lingoqa: Visual question answering for autonomous driving. In *ECCV*, 2024.

[24] Leandro Masello, German Castignani, Barry Sheehan, Finbarr Murphy, and Kevin McDonnell. On the road safety benefits of advanced driver assistance systems in different driving contexts. *TRIP*, 2022.

[25] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *ECCV*, 2024.

[26] Oluwatobi Olabiyi, Eric Martinson, Vijay Chintalapudi, and Rui Guo. Driver action prediction using deep (bidirectional) recurrent neural network. *CoRR*, 2017.

[27] Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good teachers explain: Explanation-enhanced knowledge distillation. In *ECCV*, 2024.

[28] Chirag Parikh, Deepti Rawat, Rakshitha R. T, Tathagata Ghosh, and Ravi Kiran Sarvadevabhatla. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. *CoRR*, 2025.

[29] SungYeon Park, Minjae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and DongKyu Kim. VLAAD: vision and language assistant for autonomous driving. In *WACV (Workshops)*, 2024.

[30] Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, and et al. Skywork R1V: pioneering multimodal reasoning with chain-of-thought. *arXiv:2504.05599*, 2025.

[31] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *AAAI*, 2024.

[32] Yao Rong, Zeynep Akata, and Enkelejda Kasneci. Driver intention anticipation based on in-cabin and driving scene monitoring. In *ITSC*, 2020.

[33] Abdelrahman Shaker, Muhammad Maaz, Chenhui Gou, Hamid Rezatofighi, Salman Khan, and Fahad Shahbaz Khan. Mobile-videogpt: Fast and accurate video understanding language model. *arXiv preprint arXiv:2503.21782*, 2025.

[34] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv:2410.17434*, 2024.

[35] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, 2024.

[36] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*, 2023.

[37] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *CoRL*, 2024.

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[39] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025.

[40] Koen Vellenga, H. Joe Steinhauer, Göran Falkman, and Tomas Björklund. Evaluation of video masked autoencoders' performance and uncertainty estimations for driver action and intention recognition. In *WACV*, 2024.

[41] Dat Vu, Bao Ngo, and Hung Phan. Hybridnets: End-to-end perception network. *CoRR*, 2022.

[42] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *CoRR*, 2024.

[43] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *CVPR*, 2025.

[44] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and et al. Perception-aware policy optimization for multimodal reasoning. *arXiv:2507.06448*, 2025.

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

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

[47] Ding Xinpeng, Han Jinahua, Xu Hang, Laing Xiaodan, Hang Xu, Zhang Wei, and Li Xiaomeng. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. 2024.

[48] 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.

[49] 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.

[50] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and et al. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, 2024.

[51] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *ICLR*, 2017.

[52] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv:2501.13106*, 2025.

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

[54] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020.

[55] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. In *ECCV*, 2024.