ROAD: Rule-Grounded Context-Aware Open-Vocabulary Driver Anomaly Detection

Shuaibo Li[1,*] **Yingjin Li**[1,*] **Wei Ma**[2] **Fuxiang Zhai**[1]
Sixiang Chen[1] **Zhijie Qiu**[1] **Zhaolu Kang**[3] **Lei Zhu**[1,4,†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Beijing University of Technology [3]Peking University
[4]The Hong Kong University of Science and Technology

Abstract

Driver anomaly detection (DAD) is essential for safe driving, yet recognizing a driver action alone cannot determine its abnormality, as identical behaviors may warrant different judgments under varying safety rules and driving contexts. Existing approaches predominantly focus on in-cabin action recognition, with limited consideration of explicit safety constraints and unseen driver behaviors. To address these limitations, we introduce **ROAD**, a Rule-Grounded, Context-Aware, and Open-Vocabulary framework that unifies behavioral understanding and rule-conditioned anomaly reasoning. ROAD incorporates three complementary components: Context-Aware Driver Evidence Encoding integrates in-cabin observations with surrounding road conditions; Parallel Safety Rule Encoding independently represents safety constraints and incorporates rule-specific semantics into visual evidence; and Open-Vocabulary Behavior Encoding leverages pretrained multimodal large language model (MLLM) knowledge to model behaviors beyond predefined action categories. These representations are jointly integrated to produce structured anomaly judgments with supporting evidence and explanations. To progressively develop these capabilities, we adopt a three-stage curriculum that advances from multi-source behavior learning to context–rule alignment and joint recognition and anomaly learning. We further introduce **ROAD-Bench**, a rule-grounded benchmark spanning six safety-rule categories and diverse driving scenarios. ROAD-Bench provides context-dependent anomaly annotations with supporting visual evidence and explanations, refined through multi-agent consensus and human verification. Extensive experiments across three complementary evaluation settings demonstrate ROAD's effectiveness in rule-grounded and context-aware anomaly detection, driver-action recognition, and open-vocabulary generalization.

1 Introduction

Safe driving requires sustained driver attention and timely responses to changing traffic conditions. Distraction, drowsiness, and other unsafe driver behaviors can undermine these capabilities, increasing the risk of serious accidents [20, 23]. These risks highlight the importance of reliable driver monitoring. Driver Anomaly Detection (DAD) aims to identify potentially hazardous driver behaviors and states to support driving safety assessment [14]. In real-world driving, however, the same behavior can have different safety implications depending on the surrounding traffic conditions and applicable safety rules [20, 2]. Moreover, real-world driver behaviors may extend beyond predefined action categories, further complicating reliable anomaly detection [32, 29].

[*]Equal contribution.
[†]Lei Zhu is the corresponding author of this work.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Early driver behavior analysis methods relied on conventional deep learning architectures, such as 3D CNNs and LSTMs, to capture spatial and temporal patterns of driver activities [14, 30]. Recent advances in pretrained Vision Foundation Models (VFMs) have further expanded the capabilities of driver behavior understanding by introducing rich visual and semantic knowledge acquired from large-scale pretraining [24, 13]. DriveCLIP [8] leverages pretrained vision-language representations to recognize distracted driving behaviors, demonstrating the potential of transferring general visual knowledge to driver monitoring tasks. These advances have significantly improved the accuracy of driver-action recognition.

However, existing DAD approaches still equate action recognition with driver anomaly detection [8]. This assumption overlooks a crucial distinction: using a phone while legally parked may be permissible, whereas the same action while navigating a busy intersection may violate traffic rules and pose serious risks [20]. Similarly, turning one's head to check a blind spot before changing lanes is a necessary safety precaution, but the same movement toward a passenger in heavy traffic may indicate dangerous inattention—the same action, different verdict. Such judgments require explicit safety rules and contextual evidence extending beyond in-cabin observations to vehicle states and surrounding traffic [34, 18]. Furthermore, real-world driver behaviors are too diverse to be fully captured by predefined action categories [32, 29]. We argue that reliable DAD must possess three fundamental capabilities: (1) rule-grounded reasoning to evaluate behaviors against applicable rules, conditions, and exceptions; (2) context-aware assessment to integrate in-cabin and external driving cues when determining abnormality and its degree; and (3) open-vocabulary behavior understanding to capture the semantics of diverse and previously unseen driver behaviors beyond predefined action categories.

To meet these requirements for reliable DAD, we introduce ROAD, a Rule-Grounded, Context-Aware, and Open-Vocabulary framework that jointly reasons over explicit safety rules, driving contexts, and behavioral semantics to assess driver abnormality. At the core of ROAD is a rule-conditioned evidence reasoning mechanism supported by three complementary components. Specifically, **Context-Aware Driver Evidence Encoding** integrates in-cabin observations with surrounding road conditions through road-guided modulation and cross-view fusion, enabling driver behaviors to be interpreted in their corresponding driving contexts. **Parallel Safety Rule Encoding** independently encodes safety rules to prevent cross-rule interference and incorporates rule-specific semantics into contextualized visual evidence, highlighting its relevance to different safety constraints. Meanwhile, **Open-Vocabulary Behavior Encoding** leverages pretrained vision-language knowledge [5] to capture behavioral semantics beyond predefined action categories. Finally, **Structured Anomaly Reasoning** jointly processes rule-relevant visual evidence, safety rule representations, and behavioral semantics to recognize driver actions, identify activated safety rules, assess the degree of abnormality, and generate supporting evidence and explanations.

Existing driver-behavior benchmarks predominantly focus on action labels [19, 17, 12], providing limited supervision for determining whether a behavior is anomalous under specific safety rules and driving contexts. To bridge this gap, we introduce **ROAD-Bench**, a rule-grounded driver anomaly detection benchmark constructed from AIDE [34] and organized around six categories of driving safety rules. Rather than assigning fixed abnormality labels to driver actions, ROAD-Bench provides context-dependent safety annotations grounded in driving evidence, activated safety rules, and their associated conditions and exceptions. To ensure reliable annotations and meaningful explanations, we employ a multi-agent consensus annotation pipeline supplemented by human verification. Building upon ROAD-Bench, we develop a three-stage curriculum training strategy that progressively enhances ROAD's overall performance. Specifically, Stage 1, Multi-Source Behavior Learning, employs prompt-based supervision across six diverse driver-behavior datasets, exposing the model to a broad spectrum of driver activities and enriching its behavioral semantic representations. Stage 2, Context–Rule Alignment, trains ROAD to align contextual visual evidence with applicable safety rules, using driver-action and rule-activation annotations from ROAD-Bench as supervision. Finally, Stage 3, Joint Recognition and Anomaly Learning, jointly optimizes structured anomaly prediction on ROAD-Bench and action classification on five auxiliary datasets, enabling the model to integrate diverse behavioral semantics, contextual evidence, and explicit safety constraints when determining abnormality and its degree. Furthermore, to comprehensively evaluate ROAD's effectiveness and generalization capabilities, we design three complementary experimental settings.

Our main contributions are summarized as follows:

- We propose ROAD, a rule-grounded, context-aware, and open-vocabulary DAD framework that integrates safety rules, driving contexts, and behavioral semantics to enable reliable anomaly judgments beyond action recognition.
- We introduce ROAD-Bench, a rule-grounded benchmark constructed from AIDE, covering six categories of safety rules with structured annotations for context-dependent anomaly assessment.
- We develop a three-stage training strategy that progressively enhances open-vocabulary behavioral understanding, context–rule alignment, and joint anomaly reasoning. We further design three complementary evaluation settings to assess ROAD's effectiveness and generalization.

2 Related Work

Driver Anomaly Detection. Existing driver anomaly detection (DAD) methods predominantly formulate the task as distracted driving activity recognition. Early approaches relied on conventional deep learning architectures, employing CNNs to extract appearance and posture cues from in-cabin observations [6]. Subsequent methods introduced spatiotemporal modeling, including 3D convolutions and recurrent networks, to capture motion patterns and temporal dependencies across video frames [30]. Beyond visual appearance, pose-guided approaches integrate 2D skeletal information with spatiotemporal features through Transformer-based fusion [3]. Recent advances have shifted toward leveraging pretrained Vision Foundation Models (VFMs) for enhanced semantic understanding and generalization. DriveCLIP exploits vision-language alignment for zero-shot and fine-tuned driver-action recognition, demonstrating the potential of pretrained representations for driver monitoring [8]. Despite these advances, action-centric methods often infer abnormality from in-cabin actions, overlooking how safety rules and road conditions can change the judgment of the same behavior. Moreover, predefined action categories cannot capture the full diversity of real-world driving behaviors. We therefore pursue rule-grounded anomaly reasoning that integrates in-cabin and road-context evidence while extending to previously unseen actions.

Open-Vocabulary Video Anomaly Detection. Recent advances in video anomaly detection (VAD) have increasingly leveraged pretrained vision-language models to move beyond predefined anomaly categories. VadCLIP exploits vision-language alignment for weakly supervised anomaly detection [33], while OVVAD combines category-agnostic detection with open-vocabulary classification to recognize novel anomalies [32]. Language-driven approaches further extend semantic understanding: LAVAD integrates video captioning with LLM reasoning for training-free detection [36], LAVIDA leverages MLLM-derived semantic representations for zero-shot generalization [5], and AnyAnomaly supports customizable anomaly definitions through visual question answering [2]. However, open-vocabulary recognition alone is insufficient for driver anomaly detection, where safety judgments must be grounded in applicable rules and driving contexts. This gap highlights the need for a unified framework that grounds open-vocabulary behavioral understanding in explicit safety constraints and contextual evidence, a direction we pursue in ROAD.

3 Problem Formulation and ROAD-Bench

3.1 Problem Formulation

We formulate Driver Anomaly Detection (DAD) as the task of assessing driver behaviors under surrounding driving conditions and applicable safety rules. Given synchronized in-cabin and road-context videos x^{in} and x^{out}, a collection of N safety Rule Cards $\mathcal{R} = \{r_i\}_{i=1}^{N}$, and a task instruction q, the objective is to learn a mapping:

$$(a, c, s, e, t) = \mathcal{F}\big(x^{\text{in}}, x^{\text{out}}, \mathcal{R}, q\big), \tag{1}$$

where the outputs comprise five structured fields. Specifically, a denotes *Driver Action*, identifying the observed driver behavior. c denotes *Activated Rules*, specifying which safety rules are activated by the behavior and driving context. s denotes *Anomaly Score*, quantifying the degree of abnormality, with a threshold of 0.5 used to distinguish anomalous from non-anomalous behaviors. Furthermore, e denotes *Supporting Evidence*, describing the visual and contextual evidence that supports the activation or non-activation of individual safety rules. Finally, t denotes *Explanation*, providing the rationale for the recognized driver action and explaining the assigned anomaly score in relation to the activated rules, driving conditions, and applicable exceptions.

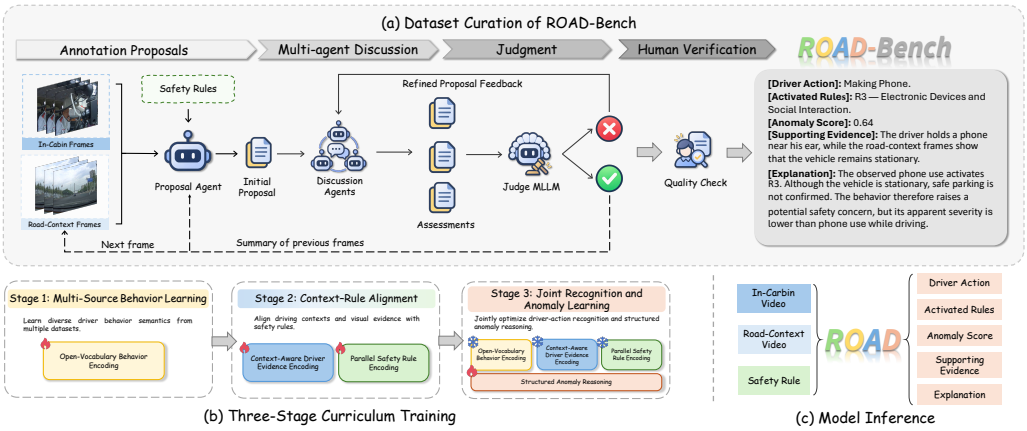

Figure 1: **Overview of ROAD-Bench and ROAD.** (a) ROAD-Bench construction through multi-agent annotation and human verification. (b) Three-stage curriculum training of ROAD. (c) Structured driver anomaly assessment during inference.

3.2 ROAD-Bench

To address the limitations of existing driver anomaly detection benchmarks in assessing behavioral safety under real-world driving conditions, we introduce **ROAD-Bench**, a rule-grounded benchmark for context-dependent driver anomaly assessment. Built upon synchronized in-cabin and road-context videos from AIDE [34], ROAD-Bench spans seven driver-action categories and incorporates six safety rules to evaluate driver behaviors across diverse driving scenarios. To establish the safety rule taxonomy, we systematically examine driving scenarios from AIDE alongside relevant traffic regulations and driver-safety guidelines. Based on this analysis, we formulate six Rule Cards that describe essential driving safety principles and distinguish appropriate from unsafe behaviors under different driving conditions. These rules provide a structured basis for context-dependent safety assessment, allowing the same driver action to receive different judgments depending on the surrounding circumstances. Following the official AIDE data split, ROAD-Bench comprises 2,898 synchronized video pairs, including 1,884 training, 405 validation, and 609 test samples. Each sample is annotated according to the five-field schema defined above, providing structured supervision for driver-action recognition, rule activation, and context-dependent anomaly assessment.

Multi-Agent Consensus Annotation To ensure the quality of safety annotations across diverse driving scenarios, we employ an iterative multi-agent annotation pipeline in which multiple MLLM agents reach consensus through proposal and discussion. As illustrated in Fig. 1(a), the pipeline consists of three main stages. Specifically, (1) Annotation Proposal. Given synchronized in-cabin and road-context videos together with the safety Rule Cards, an MLLM generates an initial safety annotation and explanation for the observed driving behavior. (2) Multi-Agent Discussion. The proposed annotation is sequentially reviewed by other MLLM agents, each of which provides a supporting or opposing assessment with its rationale. Each agent also assigns a five-level severity rating based on the observed driver behavior, driving context, and applicable rule conditions and exceptions. (3) Judgment. A judge MLLM reviews the initial proposal and subsequent discussions to determine whether a majority of agents agree on the proposed annotation and its rationale. If agreement is not reached, the judge revises the proposal and initiates another round of discussion. Otherwise, the annotation is finalized. For behaviors confirmed to violate applicable safety rules, the severity ratings are averaged and normalized to obtain a fine-grained anomaly score. Finally, after each batch of videos has been annotated, human verifiers examine a sampled subset to assess annotation quality and determine whether the batch requires re-annotation, thereby further enhancing the reliability of the resulting annotations.

4 Methodology

We introduce **ROAD**, a rule-grounded, context-aware framework for open-vocabulary driver anomaly detection. To assess driver behaviors under varying driving conditions and explicit safety constraints,

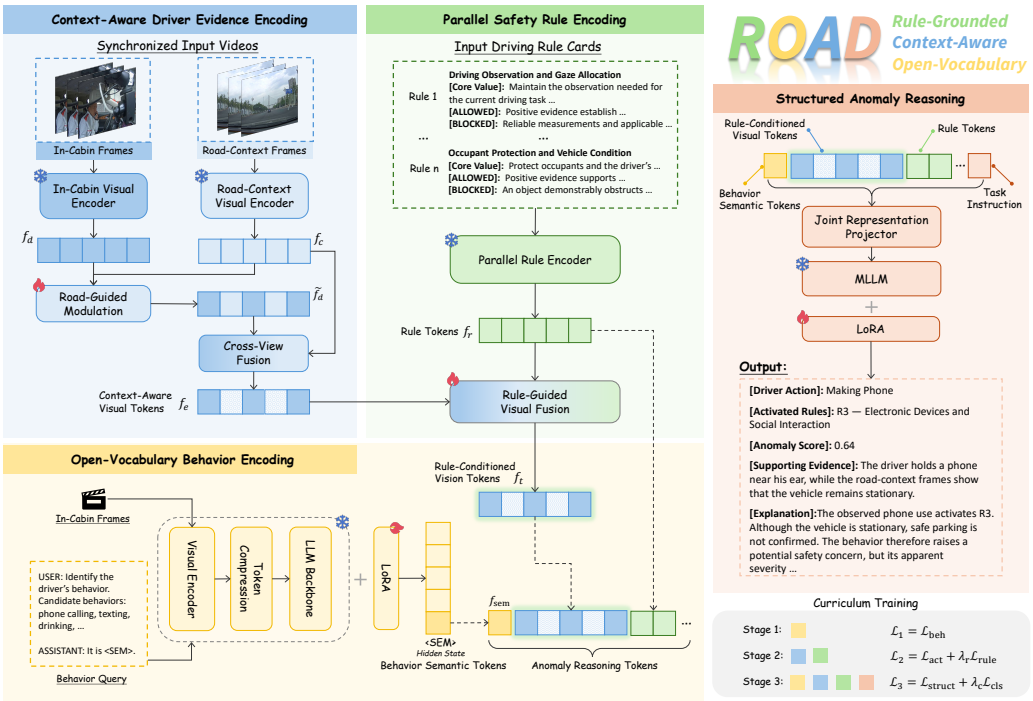

Figure 2: **Overview of the ROAD framework.** ROAD consists of three complementary components: (1) Context-Aware Driver Evidence Encoding integrates in-cabin observations and road conditions through road-guided modulation and cross-view fusion; (2) Parallel Safety Rule Encoding independently encodes safety rules and incorporates rule-specific semantics into contextual visual evidence; and (3) Open-Vocabulary Behavior Encoding leverages an MLLM to capture behavioral semantics beyond predefined action categories. Structured Anomaly Reasoning integrates these representations to generate structured anomaly assessments. The framework is trained through a three-stage curriculum.

ROAD integrates Context-Aware Driver Evidence Encoding to incorporate road-context information into driver representations, and Parallel Safety Rule Encoding to independently encode safety rules and incorporate rule-specific semantics into visual evidence. Open-Vocabulary Behavior Encoding further leverages a pretrained MLLM to capture behavioral semantics beyond predefined action categories. Building on these representations, Structured Anomaly Reasoning generates structured anomaly assessments. ROAD is further optimized through a three-stage curriculum training strategy. The overall architecture is illustrated in Fig. 2.

4.1 Context-Aware Driver Evidence Encoding

The safety implications of driver behaviors can vary with surrounding road conditions. To capture this dependency, we introduce a context-aware evidence encoding mechanism that combines road-guided modulation with cross-view fusion. Given synchronized in-cabin and road-context videos, the frozen In-Cabin Visual Encoder and Road-Context Visual Encoder extract driver features f_d and road-context features f_c, respectively. A trainable Road-Guided Modulation module then derives channel-wise modulation weights w_c from the mean-pooled road-context features to adaptively refine driver representations:

$$\tilde{f}_d = w_c \odot f_d, \tag{2}$$

where $\odot$ denotes channel-wise modulation. To further integrate information from both views, the Cross-View Fusion module summarizes the modulated driver features $\tilde{f}_d$ and road-context features f_c through average pooling. The resulting descriptors are jointly projected into a shared contextual representation f_{ctx}. This shared context is then incorporated into both feature streams, which are concatenated along the token dimension to obtain context-aware visual tokens:

$$f_e = \text{Concat}\left(\tilde{f}_d + f_{\text{ctx}},\ f_c + f_{\text{ctx}}\right). \tag{3}$$

This design preserves view-specific information while incorporating surrounding driving conditions, providing contextualized visual evidence for subsequent rule-grounded anomaly reasoning.

4.2 Parallel Safety Rule Encoding

Safety rules provide essential criteria for assessing whether driver behaviors are appropriate under specific driving conditions. However, jointly encoding multiple Rule Cards may introduce unwanted dependencies between different safety constraints and make their representations sensitive to input ordering. To address these challenges, we employ a parallel rule encoding mechanism that preserves rule-specific semantics while mitigating positional bias. Specifically, given a collection of driving Rule Cards $\mathcal{R} = \{r_i\}_{i=1}^{N}$, we organize each Rule Card as an independent segment delimited by a pair of special anchor tokens. We then apply rule-wise attention masking to prevent information exchange between different rules while preserving causal attention within each segment. Meanwhile, equivalent positional encoding based on RoPE [28] assigns consistent relative positions across rule segments, reducing biases introduced by rule ordering. A frozen Parallel Rule Encoder $\mathcal{E}_r$ processes all rules in parallel to produce Rule Tokens:

$$f_r = \mathcal{E}_r(\mathcal{R}). \tag{4}$$

To connect these independently encoded rules with visual observations, we further introduce a trainable Rule-Guided Visual Fusion module. Through attention-based interactions between Rule Tokens f_r and context-aware visual tokens f_e, the module integrates rule-specific semantics into the visual evidence. The resulting rule-conditioned visual tokens f_t integrate contextual visual evidence with rule-specific information for subsequent structured anomaly reasoning.

4.3 Open-Vocabulary Behavior Encoding

Real-world driving involves diverse driver activities that may not be fully captured by predefined action categories. To enhance the understanding of such behaviors, we leverage the semantic knowledge of a pretrained MLLM to construct open-vocabulary behavioral representations. Given an in-cabin video x^{in} and a behavior query p_s, a visual encoder first extracts visual tokens from the sampled frames. Since dense visual tokens often contain substantial background redundancy, we employ reverse-attention-based token compression [5] to reduce redundant information while preserving distinctive visual cues. The compressed visual tokens are then processed together with the behavior query by the MLLM. To obtain a compact behavioral representation, we introduce a special token `<SEM>` to aggregate semantic information from the MLLM. The final-layer hidden state corresponding to this token is extracted as:

$$f_{\text{sem}} = \Phi_{\text{MLLM}}(x^{\text{in}}, p_s). \tag{5}$$

The resulting behavior semantic tokens f_{sem} are subsequently integrated with rule-conditioned visual evidence to support structured anomaly reasoning.

4.4 Structured Anomaly Reasoning

We condition an MLLM on rule-conditioned visual tokens f_t, rule tokens f_r, behavior semantic tokens f_{sem}, and a task instruction q to generate structured anomaly assessments comprising driver actions, activated rules, anomaly scores, supporting evidence, and explanations. In parallel, a lightweight classification head provides auxiliary supervision for driver-action recognition. Inspired by prior work [37], we jointly optimize structured text generation and action classification, allowing discriminative action supervision to strengthen behavioral understanding while generative supervision grounds the shared representations in safety rules and driving context.

4.5 Three-Stage Curriculum Training

To progressively develop open-vocabulary behavioral understanding and rule-grounded anomaly reasoning, we adopt a three-stage curriculum training strategy.

Stage 1: Multi-Source Behavior Learning. To enhance ROAD's open-vocabulary behavior understanding, we first train the Open-Vocabulary Behavior Encoding branch on diverse driver-behavior datasets. Given an in-cabin visual sample and a behavior-recognition prompt, the MLLM learns

to generate the corresponding driver-action label. With the visual encoder frozen, cross-dataset behavioral supervision extends the model's semantic coverage beyond any single action taxonomy. The training objective is defined as

$$\mathcal{L}_1 = \mathcal{L}_{\text{beh}}, \tag{6}$$

where $\mathcal{L}_{\text{beh}}$ denotes the MLLM text generation loss supervised by ground-truth action labels.

Stage 2: Context–Rule Alignment. In this stage, we train ROAD to align contextual visual evidence with applicable safety rules, using driver-action and rule-activation annotations from ROAD-Bench. Given paired in-cabin and road-context videos together with safety Rule Cards, ROAD produces rule-conditioned visual tokens f_t and rule tokens f_r, which are jointly fed into a frozen MLLM. The model is supervised to generate the driver-action label and the activation status of each safety rule. With the visual, rule, and language backbones frozen, we optimize the trainable context-aware encoding and rule-guided visual fusion modules using the MLLM text generation loss:

$$\mathcal{L}_2 = \mathcal{L}_{\text{act}} + \lambda_r \mathcal{L}_{\text{rule}}, \tag{7}$$

where $\mathcal{L}_{\text{act}}$ and $\mathcal{L}_{\text{rule}}$ denote the text generation losses supervised by ground-truth action labels and multi-label rule-activation annotations, respectively. This stage establishes the alignment between contextual visual evidence and safety rules, providing a foundation for subsequent structured anomaly reasoning.

Stage 3: Joint Recognition and Anomaly Learning. Finally, we jointly optimize structured anomaly reasoning and driver-action recognition using ROAD-Bench and auxiliary driver-behavior datasets from Stage 1. For ROAD-Bench samples, the MLLM learns to generate structured anomaly assessments. For auxiliary samples, a lightweight MLP classification head operating on the dedicated classification token predicts driver-action categories using in-cabin observations alone. These two tasks are jointly optimized through:

$$\mathcal{L}_3 = \mathcal{L}_{\text{struct}} + \lambda_c \mathcal{L}_{\text{cls}}, \tag{8}$$

where $\mathcal{L}_{\text{struct}}$ denotes the MLLM text generation loss for structured anomaly assessments, and $\mathcal{L}_{\text{cls}}$ represents the auxiliary action classification loss. By combining generative anomaly supervision with discriminative action supervision, this stage encourages behavioral understanding and rule-grounded reasoning to mutually reinforce each other through shared representations.

5 Experiments

5.1 Experimental Setup

Datasets and Evaluation Settings. To comprehensively evaluate ROAD's capabilities, we conduct experiments under three complementary settings. (1) Setting 1: Rule-Grounded Driver Anomaly Detection. We conduct our primary evaluation on ROAD-Bench, which provides synchronized in-cabin and road-context videos with structured safety annotations. This setting assesses driver-action recognition, rule activation prediction, and context-dependent anomaly detection. For action recognition, we report top-1 accuracy (Act-Acc) and Macro-F1 (Act-F1) across seven action categories. Rule activation is evaluated as a multi-label classification task using exact-match accuracy (Rule-EMAcc), requiring all six rule labels to match, and Macro-F1 averaged across individual rules (Rule-F1). For anomaly detection, we report binary accuracy (Ano-Acc) with a threshold of 0.5 and the area under the precision–recall curve (Ano-AUPRC) computed from continuous anomaly scores. (2) Setting 2: Driver-Action Recognition. We evaluate ROAD on five widely used driver-behavior benchmarks, including SAM-DD [35], SynDD1 [25], DMD [23], AUC [1], and StateFarm [19]. These datasets cover diverse driver activities and recording conditions. Following established evaluation protocols [8], we report Act-Acc on each benchmark. (3) Setting 3: Open-Vocabulary Generalization. We further evaluate ROAD on 3MDAD [12] to assess its ability to recognize previously unseen driver behaviors beyond predefined action categories. We report Act-Acc and Act-F1 to measure open-vocabulary recognition performance.

Baseline Methods. To provide a comprehensive evaluation, we compare ROAD against a diverse set of baseline methods across three experimental settings. For Setting 1, we compare ROAD with open-source MLLMs, including Qwen3-VL-8B and Qwen3-VL-32B [4], InternVL3.5-8B and InternVL3.5-38B [31], as well as Qwen3-VL-8B-SFT, a supervised fine-tuned variant of Qwen3-VL-8B. We also consider closed-source models, including GPT-5.4 [21], GPT-5.6 [22], Gemini 3.1

Table 1: **Performance comparison on ROAD-Bench.** Driver-action recognition is evaluated using top-1 accuracy (Act-Acc) and Macro-F1 (Act-F1). Rule activation prediction is measured by exact-match accuracy (Rule-EMAcc) and Macro-F1 (Rule-F1). Anomaly detection is evaluated using binary accuracy (Ano-Acc) and area under the precision–recall curve (Ano-AUPRC). All results are reported in percentages (%). The best and second-best results are marked in **bold** and underline, respectively.

Methods	Action Recognition		Rule Activation Prediction		Anomaly Detection	
	Act-Acc	Act-F1	Rule-EMAcc	Rule-F1	Ano-Acc	Ano-AUPRC
Qwen3-VL-8B	46.28	31.47	65.84	47.26	62.18	68.37
Qwen3-VL-32B	58.67	43.39	73.92	57.86	68.23	76.41
InternVL3.5-8B	51.74	35.79	69.37	51.64	64.27	70.48
InternVL3.5-38B	63.81	48.19	76.48	63.06	69.84	78.42
GPT-5.4	61.43	45.72	75.29	60.18	70.92	77.63
GPT-5.6	68.14	52.38	78.57	65.42	72.36	79.84
Gemini-3.1-Pro	65.47	55.16	77.26	67.93	73.64	80.57
Gemini-3.1-Flash-Lite	54.93	38.71	72.18	55.37	66.41	72.74
Qwen3-VL-8B-SFT	71.36	62.74	80.28	70.84	75.83	82.36
Ours	**79.84**	**71.58**	**86.91**	**78.46**	**84.37**	**87.92**

Table 2: **Performance comparison on driver-action recognition benchmarks.** We report top-1 accuracy (Act-Acc, %) on three video-based and two image-based datasets. Image-only methods process video frames independently and aggregate predictions through majority voting. The best and second-best results are marked in **bold** and underline, respectively.

Methods	Processing Type	Video-based			Image-based	
		SAM-DD	SynDD1	DMD	AUC	StateFarm
ResNet18	Image-only	95.91	47.55	85.95	95.54	82.68
VGG16	Image-only	96.86	12.56	56.15	96.18	–
ShuffleNet-v2	Image-only	95.56	–	84.80	94.38	–
MobileNet-v2	Image-only	94.78	36.10	93.90	94.74	80.04
MIFI	Video-only	96.98	80.43	93.20	–	–
DriveCLIP	Image & Video	97.86	81.85	98.44	96.58	83.15
Ours	Image & Video	**98.92**	**90.58**	**98.87**	**97.83**	**88.39**

Pro, and Gemini 3.1 Flash-Lite [7]. For Setting 2, following prior work [8], we compare ROAD with established image-based models, including ResNet18 [9], VGG16 [27], ShuffleNet-v2 [16], and MobileNet-v2 [26], alongside driver-action recognition methods MIFI [15] and DriveCLIP [8]. For Setting 3, we adopt the same MLLM baselines as Setting 1 and further include DriveCLIP [8] to compare recognition performance on seen and unseen driver-action categories.

Implementation Details. The In-Cabin Visual Encoder and Road-Context Visual Encoder are both based on OpenCLIP ViT-L/14 [11] and remain frozen during training. The Parallel Rule Encoder, Open-Vocabulary Behavior Encoding, and Structured Anomaly Reasoning are built upon Qwen3-VL-8B-Instruct [4]. The rule encoder remains frozen, while the MLLM branches are fine-tuned with LoRA [10], keeping the pretrained backbone weights frozen. Training follows a three-stage curriculum, with each stage conducted for 10 epochs. The learning rate is set to 1×10^{-4} for Stages 1 and 2. In Stage 3, the fusion and projection modules and the reasoning adapter are optimized with learning rates of 2×10^{-6} and 2×10^{-5}, respectively. We employ the AdamW optimizer with a weight decay of 0.01, a 10% linear warm-up followed by cosine learning rate decay, and gradient clipping at 1.0. The effective batch size is 24 for Stage 1 and 48 for Stages 2 and 3. The loss coefficients are set to $\lambda_r = 1.0$ and $\lambda_c = 0.5$. All experiments are conducted on four NVIDIA A100 GPUs.

Table 3: **Performance comparison on open-vocabulary driver-action recognition.** We report top-1 accuracy (Act-Acc, %) and Macro-F1 (Act-F1, %) on seen and unseen action categories of 3MDAD. The best and second-best results are marked in **bold** and underline, respectively.

Methods	Seen Categories		Unseen Categories	
	Act-Acc	Act-F1	Act-Acc	Act-F1
Qwen3-VL-8B	57.63	45.28	44.17	33.46
Qwen3-VL-32B	69.28	60.47	58.74	49.36
InternVL3.5-8B	61.14	49.82	47.93	37.18
InternVL3.5-38B	74.78	67.46	65.24	56.81
GPT-5.4	72.84	63.19	62.47	53.28
GPT-5.6	78.46	70.13	67.52	62.58
Gemini-3.1-Pro	76.42	72.86	<u>70.18</u>	<u>66.47</u>
Gemini-3.1-Flash-Lite	64.81	54.83	52.41	41.72
Qwen3-VL-8B-SFT	<u>82.59</u>	<u>76.53</u>	68.84	60.34
DriveCLIP	80.37	74.62	66.13	58.42
Ours	**85.06**	**78.67**	**78.36**	**73.79**

Table 4: **Ablation studies on ROAD-Bench.** Driver-action recognition is evaluated using top-1 accuracy (Act-Acc), rule activation prediction using exact-match accuracy (Rule-EMAcc), and anomaly detection using binary accuracy (Ano-Acc). All results are reported in percentages (%).

(a) Core Components

Configuration	Act-Acc	Rule-EMAcc	Ano-Acc
w/o Road-Guided Mod.	78.63	83.46	77.86
w/o Parallel Rule Enc.	79.27	81.64	80.74
w/o Behavior Semantics	75.18	84.28	81.92
Full ROAD	**79.84**	**86.91**	**84.37**

(b) Training Strategy

Configuration	Act-Acc	Rule-EMAcc	Ano-Acc
w/o Stage 1	73.52	85.93	82.76
w/o Stage 2	77.91	82.73	79.53
w/o Action CLS	76.74	85.17	83.49
Full ROAD	**79.84**	**86.91**	**84.37**

5.2 Performance Comparison

We conduct extensive comparisons with state-of-the-art models across three complementary settings: rule-grounded anomaly detection, driver-action recognition, and open-vocabulary generalization.

Driver Anomaly Detection on ROAD-Bench. Table 1 presents the performance comparison of ROAD against open-source, closed-source, and task-specific fine-tuned MLLMs on ROAD-Bench. All models are evaluated using the same synchronized in-cabin and road-context videos and safety Rule Cards. ROAD achieves the highest performance across all six evaluation metrics. Notably, compared with the strongest baseline, Qwen3-VL-8B-SFT, ROAD improves rule-activation exact-match accuracy by 6.63 percentage points and anomaly detection AUPRC by 5.56 percentage points. These gains stem from the complementary roles of Context-Aware Driver Evidence Encoding and Parallel Safety Rule Encoding. The former incorporates surrounding road conditions into driver representations, while the latter integrates rule-specific semantics into contextual visual evidence. Their integration strengthens rule activation prediction and enables more accurate anomaly assessment under varying driving conditions.

Driver-Action Recognition on Diverse Benchmarks. Table 2 presents the comparison of driver-action recognition accuracy across five established benchmarks. ROAD achieves the highest accuracy on all five datasets. Notably, compared with the strongest baseline, DriveCLIP, ROAD delivers improvements of 8.73 and 5.24 percentage points on SynDD1 and StateFarm, respectively. These gains stem from our Joint Recognition and Anomaly Learning strategy. Structured anomaly generation encourages the model to learn behavioral representations informed by driving contexts and safety rules, while auxiliary action classification strengthens the discrimination of driver actions. Their joint optimization enhances driver-action recognition across diverse behaviors and both video- and image-based benchmarks.

Open-Vocabulary Generalization on 3MDAD. Table 3 presents the comparison of driver-action recognition performance on seen and unseen categories of 3MDAD. ROAD achieves the highest Act-Acc and Act-F1 in both settings, with particularly substantial improvements on unseen categories.

Notably, compared with the strongest general-purpose MLLM, Gemini-3.1-Pro, ROAD improves unseen-category Act-Acc and Act-F1 by 8.18 and 7.32 percentage points, respectively. Furthermore, ROAD surpasses the task-specific Qwen3-VL-8B-SFT baseline by 9.52 percentage points in unseen-category accuracy. These gains demonstrate the effectiveness of Open-Vocabulary Behavior Encoding, which leverages pretrained MLLM semantic knowledge to capture diverse driver behaviors beyond predefined action categories. By learning transferable behavioral representations, ROAD achieves stronger generalization to previously unseen actions while maintaining competitive performance on seen categories.

5.3 Ablation Study

Analysis of Core Components. Table 4(a) evaluates the contributions of ROAD's core components. Removing Road-Guided Modulation reduces anomaly detection accuracy from 84.37% to 77.86%, highlighting the importance of incorporating surrounding road conditions into driver representations. Disabling Parallel Rule Encoding lowers rule-activation exact-match accuracy to 81.64%, demonstrating the benefit of independently encoding safety rules to mitigate cross-rule interference. Furthermore, removing behavioral semantic information decreases action recognition accuracy to 75.18%, underscoring the role of open-vocabulary representations in recognizing diverse driver behaviors.

Analysis of Curriculum Training. Table 4(b) examines the effectiveness of ROAD's training strategy. Removing Stage 1 reduces action recognition accuracy from 79.84% to 73.52%, demonstrating the importance of learning diverse driver behaviors through multi-source supervision. Without Stage 2, rule-activation exact-match accuracy and anomaly detection accuracy decrease to 82.73% and 79.53%, respectively, highlighting the contribution of Context–Rule Alignment to rule-grounded anomaly assessment. Furthermore, removing auxiliary action classification lowers action recognition accuracy to 76.74%. This result demonstrates the benefit of Joint Recognition and Anomaly Learning, where discriminative action supervision complements structured anomaly generation to strengthen driver behavior representations.

6 Conclusion

We present **ROAD**, a rule-grounded, context-aware, and open-vocabulary framework that advances driver anomaly detection beyond action recognition. By integrating explicit safety rules, in-cabin and road-context evidence, and open-vocabulary behavioral semantics, ROAD enables context-dependent anomaly judgments with supporting evidence and explanations. We further introduce **ROAD-Bench**, a rule-grounded benchmark for evaluating driver safety under diverse driving conditions, alongside a three-stage curriculum that progressively develops behavioral understanding, context–rule alignment, and structured anomaly reasoning. Evaluations across three complementary settings examine ROAD's effectiveness in rule-grounded anomaly detection, driver-action recognition, and open-vocabulary generalization. Our work provides a unified perspective on reliable driver anomaly detection, shifting the focus from recognizing driver actions to understanding their safety implications.

Limitations and Future Work. ROAD currently relies on six safety-rule categories and visual observations, which may not fully capture diverse driving regulations or conditions requiring additional vehicle-state information. Moreover, its evaluation is limited to existing driving benchmarks. Future work will expand rule and scenario coverage, incorporate additional driving information, and explore more robust anomaly assessment in real-world environments.

7 Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515140021), the Guangdong Science and Technology Department (No. 2024ZDZX2004), the Program of Guangdong Education Department, the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), the National Natural Science Foundation of China (Nos. 62672013 and 62176010), and the Beijing Natural Science Foundation (No. 4252029).

References

[1] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. Real-time distracted driver posture classification. 2017.

[2] Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. AnyAnomaly: Zero-shot customizable video anomaly detection with LVLM. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3026–3035, 2026.

[3] Erkut Akdag, Zeqi Zhu, Egor Bondarev, et al. Transformer-based fusion of 2d-pose and spatio-temporal embeddings for distracted driver action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2023.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

[6] Hesham M Eraqi, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of advanced transportation*, 2019(1):4125865, 2019.

[7] Google. Gemini API models. Google AI for Developers Documentation, 2025. URL `https://ai.google.dev/gemini-api/docs/models`.

[8] Md Zahid Hasan, Jiajing Chen, Jiyang Wang, Mohammed Shaiqur Rahman, Ameya Joshi, Senem Velipasalar, Chinmay Hegde, Anuj Sharma, and Soumik Sarkar. Vision-language models can identify distracted driver behavior from naturalistic videos. *IEEE Transactions on Intelligent Transportation Systems*, 25(9):11602–11616, 2024.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[11] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

[12] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3mdad. *Signal Processing: Image Communication*, 88:115960, 2020.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[14] Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 91–100, 2021.

[15] Jian Kuang, Wenjing Li, Fang Li, Jun Zhang, and Zhongcheng Wu. Mifi: Multi-camera feature integration for robust 3d distracted driver activity recognition. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):338–348, 2023.

[16] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[17] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019.

[18] Xianhui Meng, Yuchen Zhang, Zhijian Huang, Zheng Lu, Ziling Ji, Yaoyao Yin, Hongyuan Zhang, Guangfeng Jiang, Yandan Lin, Long Chen, et al. Is your VLM for autonomous driving safety-ready? a comprehensive benchmark for evaluating external and in-cabin risks. *arXiv preprint arXiv:2511.14592*, 2025.

[19] Anna Montoya, Dan Holman, SF_data_science, Taylor Smith, and Wendy Kan. State farm distracted driver detection. Kaggle Competition Dataset, 2016. URL `https://www.kaggle.com/competitions/state-farm-distracted-driver-detection`.

[20] National Highway Traffic Safety Administration. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Notice of Federal Guidelines, Federal Register, 78 FR 24818–24890, April 2013.

[21] OpenAI. GPT-5.4 model documentation. OpenAI API Documentation, 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.4`.

[22] OpenAI. GPT-5.6 Sol model documentation. OpenAI API Documentation, 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.6-sol`.

[23] Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *European Conference on Computer Vision*, pages 387–405. Springer, 2020.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[25] Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, and Shuo Wang. Synthetic distracted driving (syndd1) dataset for analyzing distracted behaviors and various gaze zones of a driver. *Data in brief*, 46:108793, 2023.

[26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[29] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37:139751–139785, 2024.

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[31] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[32] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024.

[33] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6074–6082, 2024.

[34] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20459–20470, 2023.

[35] Haohan Yang, Haochen Liu, Zhongxu Hu, Anh-Tu Nguyen, Thierry-Marie Guerra, and Chen Lv. Quantitative identification of driver distraction: A weakly supervised contrastive learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):2034–2045, 2023.

[36] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024.

[37] Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19638–19648. IEEE, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have explicitly stated our contributions and research scope in both the abstract and the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have explicitly discussed the limitations of our work in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical and methodological, and does not present theoretical theorems or formal proof claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have detailed the datasets and implementation settings in the experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We plan to make our code and data publicly available upon acceptance of the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have detailed the training and evaluation procedures in the Implementation Details section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report average performance across evaluation benchmarks.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have specified the hardware and computational frameworks used in our experiments in the Implementation Details section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the Introduction section, we discuss the importance of driving safety and the potential societal benefits of our research in improving road safety.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not entail such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided a detailed reference list.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have included detailed documentation in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We have detailed the use of LLMs in the Implementation Details section.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.