BD-HazardVLM: Probing Vision-Language Models for Latent Defensive-Driving Hazards in Bangladesh Road Scenes

Abdullah Al Maruf[1]*, Md. Sajedul Islam[1], Tanmoy Mridha[1], and Irfan Hossain Bhuiyan[1]

Department of Computer Science and Engineering,
Rajshahi University of Engineering & Technology (RUET), Rajshahi, Bangladesh
{2103115,2103096,2103118,2103088}@student.ruet.ac.bd

Abstract. Vision-language models are increasingly used for driving-scene understanding, yet their defensive reasoning under weakly structured Global South traffic remains unclear. We introduce BD-HAZARD-VLM, a diagnostic benchmark of 500 Bangladesh road-scene images—375 from RSUD20K and 125 from TFP-BD—annotated for hazard presence, primary hazard category, risk timing, severity, defensive actions, visible evidence, and hallucination- and Western-road-bias traps. We evaluate Qwen2.5-VL-7B, InternVL3-8B, and MiniCPM-V-4.6 using a shared zero-shot task definition and structured output schema. Binary hazard-presence accuracy ranges from 88.0% to 89.2%, essentially matching the 89.0% majority-class baseline. Fine-grained reasoning is substantially weaker: category accuracy remains below 40%, risk-timing accuracy ranges from 1.6% to 12.4%, and exact defensive-action matching remains below 2%. Although InternVL3-8B achieves the highest severity accuracy at 61.2%, it has zero recall on the 58 high-severity cases. Qwen exhibits severe category mode collapse, while all models show weak latent-hazard anticipation and near-zero self-reported trap recognition. For the evaluated models, these findings show that coarse hazard recognition does not imply reliable fine-grained defensive-driving reasoning and motivate geographically diverse safety evaluation.

Keywords: Vision-language models · Defensive driving · Hazard reasoning · Bangladesh road scenes · Geographic robustness

1 Introduction

Vision-language models (VLMs) are increasingly used for interpretable driving-scene understanding, visual question answering, planning, and language-guided control [13,14,16,23]. Their ability to describe visible road users is promising, but safety-critical driving requires more than object recognition. A defensive driver must identify which element constitutes the primary hazard, estimate whether

* Corresponding author.

the risk is immediate or may emerge later, assess its severity, and select a suitable preventive action.

This distinction is particularly important for *latent hazards*. A pedestrian partly hidden behind a vehicle, a bus blocking the forward view, an unstable mixed-traffic interaction, or a roadside crowd may not represent an immediate collision threat in a single frame. Nevertheless, the scene may require slowing down, maintaining distance, avoiding an overtake, or preparing to brake before the danger becomes explicit. Existing driving-VLM benchmarks increasingly examine corner cases and model reliability [3, 21, 22], but fine-grained defensive reasoning under weakly structured traffic remains underexplored.

Bangladesh road scenes present a demanding test environment. They often contain heterogeneous vehicles, rickshaws and easybikes, CNG auto-rickshaws, motorcycles, informal pedestrian crossings, roadside markets, large-vehicle occlusion, damaged surfaces, waterlogging, and limited lane organization. South Asian datasets such as IDD and IDD-3D have established the importance of evaluating perception beyond well-structured Western road environments [4,17]. Bangladesh-specific datasets such as RSUD20K and TFP-BD provide geographically relevant imagery [8, 27], but their original annotations do not directly evaluate hazard timing, severity, or defensive action selection.

Geographic mismatch may also encourage unsupported assumptions. A VLM may infer zebra crossings, formal lane boundaries, sidewalks, traffic signals, or organized road behavior even when such evidence is absent. Prior studies demonstrate that multimodal models can hallucinate visual content and exhibit systematic geographic or contextual biases [6,7,9]. Such errors are especially concerning in driving, where a plausible but visually ungrounded explanation may conceal unsafe reasoning.

To study these limitations, we introduce BD-HazardVLM, a 500-image Bangladesh-centric benchmark with structured annotations for hazard presence, primary hazard category, risk timing, severity, defensive actions, visible evidence, hallucination traps, and Western-road-bias traps. We evaluate Qwen2.5-VL-7B, InternVL3-8B, and MiniCPM-V-4.6 using a shared zero-shot task definition and structured output schema.

Our results show that binary hazard-presence accuracy is misleading: a trivial majority-class predictor reaches 89.0%, essentially matching all evaluated models. Fine-grained reasoning is substantially harder. Hazard-category accuracy remains below 40%, risk-timing accuracy ranges from 1.6% to 12.4%, and exact defensive-action matching remains below 2%. Moreover, InternVL3-8B, despite achieving the highest aggregate severity accuracy, fails to identify any of the 58 high-severity cases. The models also exhibit category mode collapse and near-zero self-reported recognition of hallucination- and Western-bias-prone scenes.

Our main contributions are:

- We introduce BD-HazardVLM, a 500-image diagnostic benchmark for defensive-driving reasoning in Bangladesh mixed-traffic scenes.

- We define a structured taxonomy covering twelve locally relevant hazard categories, temporal risk, severity, defensive actions, visible evidence, and scene-level hallucination and Western-road-bias traps.
- We provide a controlled zero-shot evaluation of three open VLMs using the same 500 images, semantic task definition, output schema, and per-image statistical comparisons.
- We identify systematic failures in category selection, latent hazard anticipation, high-severity recognition, context-specific action recommendation, and self-reported trap awareness.

2 Related Work

2.1 Vision-Language Models for Autonomous Driving

Recent work has explored language-conditioned driving systems for scene understanding, explanation, planning, and control. DriveGPT4 combines multi-frame visual input with language generation and low-level control prediction, while LMDrive studies language-guided closed-loop driving with multi-view sensors and natural-language instructions [13, 23]. DriveLM organizes perception, prediction, and planning questions as a graph-structured reasoning task, and DriveVLM decomposes scene description, analysis, and hierarchical planning into language-mediated modules [14,16]. OmniDrive extends this direction toward 3D grounding and counterfactual reasoning for decision making and planning [19].

These systems demonstrate the potential of language models for interpretable driving, but most are designed around full-stack driving, trajectory prediction, question answering, or closed-loop control. Their benchmarks are also largely built from datasets such as nuScenes, CARLA, or BDD-X, whose visual and infrastructural assumptions differ from dense, weakly structured Bangladesh mixed traffic. In contrast, BD-HazardVLM isolates image-level defensive reasoning and evaluates whether a model can identify the primary hazard, distinguish latent from immediate risk, estimate severity, and recommend an appropriate defensive response.

2.2 Hazard, Corner-Case, and Reliability Evaluation

Driving corner cases have increasingly been used to test whether VLMs are visually grounded and safety-aware. CODA-LM evaluates large VLMs on self-driving corner cases through general perception, regional perception, and driving suggestion tasks [3]. DriveBench evaluates reliability under clean, corrupted, and text-only inputs and shows that plausible driving answers may arise from textual priors or general knowledge rather than visual evidence [21]. AutoTrust broadens evaluation to trustfulness, safety, robustness, privacy, and fairness in driving VLMs [22].

BD-HazardVLM is complementary to these efforts. It does not evaluate end-to-end control or generic trustworthiness across many attack types. Instead,

it targets a narrower defensive-driving question: whether a VLM can reason about locally relevant hazards before they become immediate. Its structured labels separate hazard identity, temporal risk, severity, and defensive action, enabling diagnosis of failures that may be hidden by a single overall score or by binary hazard detection. Recent benchmarks examine complementary safety-reasoning capabilities. DRAMA-X evaluates vulnerable-road-user intent, risk, and ego-action suggestions, while SCD-Bench studies safety cognition under ambiguous, misleading, or unsafe driving instructions [5, 25]. NuScenes-SpatialQA and RoadSceneBench instead emphasize spatial and relational road-scene reasoning [10, 15], whereas INSIGHT improves hazard localization and reasoning through supervised alignment [2]. In contrast, BD-HazardVLM evaluates zero-shot, image-level defensive reasoning in Bangladesh mixed traffic with explicit labels for hazard identity, timing, severity, and action.

2.3 South Asian and Bangladesh Road-Scene Datasets

Geographic coverage remains a central limitation of autonomous-driving data. The Indian Driving Dataset (IDD) introduced pixel-level annotations for unconstrained road environments, while IDD-3D added multi-camera and LiDAR data for 3D detection and tracking in unstructured Indian traffic [4, 17]. These datasets established the need to evaluate perception under heterogeneous vehicle types, weak lane discipline, dense traffic, and road layouts that differ from those in widely used Western benchmarks.

For Bangladesh, RSUD20K provides more than 20,000 driving-perspective images with object-detection annotations for locally relevant road users and vehicles [27]. TFP-BD contains 23,678 images from urban Dhaka traffic videos, including irregular pedestrian movement, different road configurations, and varied environmental conditions [8]. Both datasets are valuable sources of geographically relevant visual data, but their original annotations target object detection, traffic flow, or pedestrian analysis rather than defensive-driving reasoning. BD-HazardVLM builds on their imagery by adding image-level hazard categories, risk timing, severity, recommended actions, and scene-level hallucination and geographic-bias traps.

2.4 Hallucination and Geographic Bias in VLMs

Large VLMs can generate objects or attributes that are unsupported by the image. POPE introduced systematic object-hallucination probing, while HallusionBench evaluates entangled language hallucination, visual illusion, and logical inconsistency [6, 9]. Bias benchmarks further show that multimodal models can exhibit systematic response disparities across social and geographic contexts [20]. Recent geographic evaluation also reports weaker VLM performance for less wealthy and sparsely represented regions, together with systematic location over-prediction [7].

Table 1: Composition of the BD-HazardVLM benchmark.

Source dataset	Number of images	Proportion
RSUD20K	375	75.0%
TFP-BD	125	25.0%
Total	500	100.0%

In driving scenes, such failures may appear as unsupported assumptions about lane markings, zebra crossings, sidewalks, traffic signals, or other infrastructure that is common in highly represented training regions but absent from the image. BD-HazardVLM therefore annotates *hallucination traps* and *Western-road-bias traps*. These labels describe scene susceptibility, not verified hallucinations in a model's complete response. Accordingly, our evaluation measures whether a model recognizes that unsupported reasoning is possible; direct hallucination measurement would require separate evidence-based human adjudication. Unlike NOPE and DO-Bench, which directly evaluate unsupported object existence claims and diagnose perceptual versus prior-driven failures [11,18], our trap fields measure whether a model self-reports that a scene is susceptible to unsupported reasoning.

3 The BD-HazardVLM Benchmark

3.1 Benchmark Scope and Image Sources

BD-HazardVLM is an image-level diagnostic benchmark for evaluating defensive driving hazards in Bangladesh road scenes. Table 1 summarizes the benchmark composition by source dataset. It contains 500 images, comprising 375 images from RSUD20K [27] and 125 images from TFP-BD [8]. The initial 100-image pilot subset is included within the final 500-image benchmark.

The source datasets provide complementary visual conditions. RSUD20K contributes road-level scenes containing pedestrians, rickshaws, easybikes, CNG auto-rickshaws, motorcycles, buses, trucks, private vehicles, and informal roadside activity. TFP-BD contributes dense mixed-traffic scenes and broader traffic-flow views. Each benchmark image is assigned a unique identifier from `BDHV_0001` to `BDHV_0500`. All annotations and model predictions are matched using this identifier rather than source-specific file paths.

The benchmark is designed for image-level reasoning rather than object detection, segmentation, or trajectory prediction. It therefore provides scene-level labels describing the most important defensive-driving hazard and the corresponding safe response.

3.2 Scenario-Driven Image Selection

Images were selected to maximize scenario diversity rather than to estimate the natural prevalence of road hazards. The selection process covered dense

mixed traffic, informal pedestrian crossings, rickshaw/easybike interactions, unpredictable CNG stopping, motorcycle weaving, large-vehicle occlusion, damaged road surfaces, waterlogging, roadside markets, animal- or cart-related obstacles, uncontrolled intersections, low visibility, and scenes without an obvious major hazard.

The sampling process also considered road type, traffic density, weather, time of day, visibility, vehicle composition, pedestrian activity, and the presence of visually competing hazards. This diagnostic selection strategy intentionally emphasizes difficult cases and should not be interpreted as a representative estimate of the frequency of hazards across all Bangladesh roads.

3.3 Annotation Taxonomy and Schema

Each image is annotated using a structured image-level schema. The core evaluation fields are:

- `hazard_present`: whether a major defensive-driving hazard is present;
- `primary_hazard_category`: the most important hazard affecting the ego driver's decision;
- `risk_timing`: whether the risk is `immediate`, `latent`, `both`, or `none`;
- `severity`: the estimated consequence level, categorized as `low`, `medium`, or `high`;
- `defensive_actions`: one or more recommended actions selected from a predefined vocabulary;
- `evidence`: a brief description of the visible cues supporting the annotation;
- `hallucination_trap`: whether the scene may encourage unsupported visual claims;
- `western_bias_trap`: whether the scene may encourage unsupported assumptions based on Western road infrastructure.

The taxonomy contains twelve mutually exclusive primary hazard categories: informal pedestrian crossing; rickshaw or easybike hazard; unpredictable CNG stopping; bus or truck occlusion; motorcycle wrong-way movement or weaving; road-surface damage; waterlogging; roadside market or crowd; animal- or cart-related obstacle; uncontrolled intersection; low visibility; and control scenes without a major hazard. The machine-readable labels are provided in the released annotation guidelines.

The defensive-action vocabulary covers slowing down, stopping or yielding, preparing to brake, maintaining distance, avoiding overtaking, monitoring occlusion, changing lanes carefully, proceeding cautiously, and taking no action. Multiple actions are evaluated as an unordered set.

Risk-timing and severity labels are image-level defensive assessments based on visible scene evidence. They are not physical estimates of time-to-collision or dynamic collision probability, which would require motion and ego-state information unavailable from a single frame.

Table 2: Selected statistics of the 500-image benchmark.

Statistic	Count or prevalence
Hazard-positive images	445 (89.0%)
Control/no-major-hazard images	55 (11.0%)
Latent timing	404 (80.8%)
Immediate timing	41 (8.2%)
No-risk timing	55 (11.0%)
Both timing	0 (0.0%)
Hallucination-trap positive	324 (64.8%)
Western-bias-trap positive	285 (57.0%)

3.4 Annotation and Quality Control

Four undergraduate annotators participated in the annotation process, including three Computer Science and Engineering students and one Physics student. The 500 images were divided into four disjoint subsets of 125 images, with each annotator labeling one subset using the same annotation guideline and predefined taxonomy.

Two members of the annotation team—one CSE undergraduate and one Physics undergraduate—served as reviewers and examined all 500 annotations for label consistency, taxonomy compliance, evidence quality, and formatting errors. Ambiguous or inconsistent cases were revisited jointly with the original annotator and resolved through discussion and consensus. When consensus remained difficult, the project lead made the final decision based on visible image evidence and the predefined annotation rules.

This protocol provides complete centralized review of the benchmark. However, the initial annotations were not produced independently by multiple annotators for every image. We therefore do not report a formal inter-annotator agreement statistic and discuss this limitation in Section 7. Consequently, scores for subjective attributes such as risk timing, severity, and defensive-action sets should be interpreted relative to the adjudicated BD-HazardVLM labels rather than as estimates against a unique human-consensus ground truth.

3.5 Dataset Characteristics

Table 2 summarizes several important benchmark characteristics. A total of 445 images contain a major hazard, while 55 are annotated as control scenes. Risk timing is strongly dominated by latent hazards: 404 images are labeled `latent`, 41 are labeled `immediate`, and 55 are labeled `none`. No ground-truth image is labeled `both`.

The resulting class imbalance is intentional because the benchmark focuses on difficult hazard reasoning rather than balanced binary classification. Consequently, raw hazard-presence accuracy is treated as a secondary metric, while category-level, timing, severity, and defensive-action performance form the primary evaluation.

Table 3: Model-specific inference configurations.

Model	Precision	Image processing	Max tokens
Qwen2.5-VL-7B	4-bit NF4	200k–401k pixels	180
InternVL3-8B	8-bit	448^2, up to 6 tiles	256
MiniCPM-V-4.6	FP16	4x, up to 9 slices	190

3.6 Release and Licensing

The original source images will not be redistributed. Instead, the public release will contain the image manifest, source-image identifiers, annotations, taxonomy and annotation guidelines, evaluation prompts, model predictions, parsing utilities, and evaluation scripts. Users will obtain the original images from their respective dataset repositories and reconstruct the benchmark using the released manifest.

The annotation, documentation, and model-output layers are intended for release under the Creative Commons Attribution-NonCommercial 4.0 International license, while the accompanying evaluation and preparation code is intended for release under the MIT License. The release will retain source-dataset attribution and licensing information.

4 Evaluation Protocol

4.1 Models and Inference Setup

We evaluate three open vision-language models: Qwen2.5-VL-7B [1], InternVL3-8B [26], and MiniCPM-V-4.6 [12, 24]. Each model processes the same final set of 500 images in a zero-shot setting without task-specific training or fine-tuning. All models use greedy decoding with sampling disabled to reduce stochastic variation.

Model-specific inference settings are summarized in Table 3. Qwen2.5-VL-7B is loaded using 4-bit NF4 quantization with double quantization and FP16 computation. Its visual input budget ranges from 200,704 to 401,408 pixels. InternVL3-8B uses 8-bit quantization and dynamic aspect-ratio-preserving 448×448 image tiling, with at most six tiles and an additional thumbnail. MiniCPM-V-4.6 is evaluated in FP16 using its native slice-based visual representation, with 4x downsampling and at most nine slices.

The models retain their native processors and image representations. Consequently, numerical precision and visual preprocessing are not identical across models. These differences reflect GPU-memory constraints and are disclosed when interpreting small performance differences.

4.2 Prompt and Output Schema

All models receive the same semantic task instruction, Bangladesh mixed-traffic context, hazard taxonomy, action vocabulary, and output schema. The prompt

requires the model to rely only on visible image evidence and explicitly warns against inventing lane markings, traffic signals, zebra crossings, sidewalks, stop signs, or other unsupported Western-road infrastructure.

Each response contains eight structured fields:

```
hazard_present, primary_hazard_category, risk_timing,
severity,
defensive_actions, evidence, possible_hallucination,
possible_western_road_bias
```

For control scenes, the required outputs are the control category, no-risk timing, low severity, and no action. Model outputs are parsed into the predefined label space. MiniCPM-V-4.6 is allowed one formatting-only retry when its initial response violates the schema; the retry provides no additional visual or semantic information. The evidence field is retained for qualitative auditing and is not scored as a separate quantitative task.

4.3 Evaluation Metrics

Hazard presence is evaluated using accuracy and compared with the majority-class baseline. Because 89.0% of the benchmark is hazard-positive, this binary task is treated as secondary.

For hazard category, risk timing, and severity, we report exact-match accuracy; for hazard category, we additionally report macro-F1. Defensive actions are treated as unordered sets and evaluated using exact set match, any-overlap rate, and mean Jaccard similarity.

The model-generated hallucination and Western-bias fields are evaluated against the corresponding human trap annotations using positive-flag rate and trap recall. These quantities measure whether a model recognizes that a scene is susceptible to unsupported reasoning; they do not measure actual hallucination frequency.

4.4 Statistical Analysis

For category, timing, and severity, pairwise model differences are tested using exact McNemar tests on per-image correctness. Defensive-action Jaccard scores are compared using paired Wilcoxon signed-rank tests. Within each task, the three pairwise p-values are adjusted using the Holm procedure. Corrected values below 0.05 are considered statistically significant.

5 Results

5.1 Overall Performance

Table 4 summarizes the core diagnostic performance of the three models on all 500 benchmark images. No model dominates every task. Qwen2.5-VL-7B

Table 4: Core performance on the 500-image BD-HazardVLM benchmark. All values are percentages; higher is better.

Model	Cat. Acc.	Cat. F1	Timing Acc.	Severity Acc.	Action Jac.
Qwen2.5-VL-7B	**38.2**	9.9	9.2	58.8	22.0
InternVL3-8B	35.0	9.4	**12.4**	**61.2**	25.8
MiniCPM-V-4.6	27.6	**10.5**	1.6	49.8	**29.9**

achieves the highest hazard-category accuracy, InternVL3-8B obtains the highest risk-timing and severity accuracy, and MiniCPM-V-4.6 achieves the highest defensive-action Jaccard score.

Hazard-presence baseline. Raw hazard-presence accuracy ranges from 88.0% to 89.2%. However, 445 of the 500 ground-truth images are hazard-positive. A trivial majority-class classifier that always predicts hazard-present therefore achieves 89.0% accuracy, essentially matching all three models. Consequently, binary hazard presence does not meaningfully discriminate model capability on this benchmark. The fine-grained category, timing, severity, and action tasks form the primary evaluation.

Statistical comparisons. Using the Holm-adjusted tests defined in Section 4, the category difference between Qwen2.5-VL-7B and InternVL3-8B is not significant ($p = 0.101$), whereas both outperform MiniCPM-V-4.6 ($p < 0.001$). For timing, InternVL3-8B outperforms Qwen2.5-VL-7B ($p = 0.005$), and both outperform MiniCPM-V-4.6 ($p < 0.001$). Severity differences are significant between InternVL3-8B and Qwen2.5-VL-7B ($p = 0.036$) and between either model and MiniCPM-V-4.6 ($p < 0.001$). All action-Jaccard pairwise differences are significant ($p < 0.001$).

5.2 Hazard-Category Recognition

Hazard-category accuracy remains below 40% for all evaluated models. Qwen2.5-VL-7B obtains 38.2%, followed by InternVL3-8B at 35.0% and MiniCPM-V-4.6 at 27.6%. Their corresponding category macro-F1 scores are 9.9%, 9.4%, and 10.5%, respectively. Thus, MiniCPM-V-4.6 has the lowest exact accuracy but the highest macro-F1, reflecting broader—although still weak—coverage across the hazard taxonomy.

Qwen2.5-VL-7B exhibits substantial category mode collapse, assigning the rickshaw/easybike hazard category to 426 of 500 images (85.2%), although this category appears in only 205 ground-truth images (41.0%). Locally salient vehicle classes therefore often dominate the model's interpretation of the complete scene, including cases where the primary hazard is pedestrian behavior, large-vehicle occlusion, road damage, or an uncontrolled intersection.

5.3 Risk-Timing and Severity Estimation

Risk timing is the weakest evaluated capability. InternVL3-8B obtains 12.4% accuracy, Qwen2.5-VL-7B obtains 9.2%, and MiniCPM-V-4.6 obtains only 1.6%. The ground truth contains 404 latent, 41 immediate, and 55 no-risk (`none`) cases, with no image labeled `both`. In contrast, MiniCPM-V-4.6 predicts `both` for 283 images and `immediate` for 217 images, while never predicting `latent` or `none`. It is therefore correct on only eight images.

A raw-output audit confirms that the 1.6% MiniCPM-V-4.6 timing score is not caused by a label-extraction mismatch: every final raw `risk_timing` value exactly matches its parsed value. In 215 initial responses, the model produced the semantically equivalent phrase "both immediate and latent" rather than the permitted token `both`. The predefined single formatting-repair retry subsequently returned a schema-valid output. The final distribution therefore reflects genuine prediction collapse rather than parser loss.

Aggregate severity accuracy is higher, but it conceals safety-critical class failures. InternVL3-8B achieves the highest aggregate severity accuracy at 61.2%, yet its recall on the 58 high-severity cases is 0.0%. High-severity recall is only 5.2% for Qwen2.5-VL-7B and 22.4% for MiniCPM-V-4.6. Both InternVL3-8B and Qwen2.5-VL-7B predominantly collapse toward medium severity. MiniCPM-V-4.6 recognizes more high-severity cases but also overpredicts that class.

5.4 Defensive-Action Recommendation

Exact action-set matching is 1.6% for Qwen2.5-VL-7B, 0.4% for InternVL3-8B, and 1.0% for MiniCPM-V-4.6. Because exact matching requires the complete predicted action set to equal the reference set, partial-overlap measures are more informative. The any-overlap rate is 66.6%, 70.6%, and 73.2%, while mean Jaccard similarity is 22.0%, 25.8%, and 29.9%, respectively.

The models frequently recommend broadly safe actions such as `slow_down` and `prepare_to_brake`. However, they struggle to select context-specific responses such as `monitor_occlusion`, `avoid_overtaking`, `keep_distance`, or `stop_or_yield`. General safety language therefore does not consistently translate into precise defensive behavior.

5.5 Hallucination- and Western-Road-Bias Trap Recognition

Human annotations mark 64.8% of images as hallucination-prone and 57.0% as Western-road-bias-prone, yet all models almost always produce negative self-reported flags. Table 5 quantifies this mismatch.

For hallucination traps, recall is 0.0%, 0.3%, and 0.0% for Qwen2.5-VL-7B, InternVL3-8B, and MiniCPM-V-4.6, respectively. For Western-road-bias traps, recall is 1.1%, 1.8%, and 0.0%, respectively.

These recall values differ from the unconditional flag rates in Table 5 because recall is conditioned only on ground-truth-positive images, whereas the flag rates

Table 5: Ground-truth trap prevalence and unconditional model positive-flag rates, expressed as percentages. Each model flag rate is computed over all 500 images and may therefore include false positives on trap-negative images.

Trap	GT prev.	Qwen	InternVL	MiniCPM
Hallucination	64.8	0.0	0.4	0.6
Western-road bias	57.0	3.8	2.8	0.2

include false positives across all 500 images. For example, Qwen2.5-VL-7B produces 19 Western-bias-positive flags, corresponding to a 3.8% unconditional flag rate. Only three of these flags occur among the 285 ground-truth-positive images, yielding 1.1% recall. Similarly, MiniCPM-V-4.6 produces three hallucination-trap flags, but all three are false positives, resulting in a nonzero 0.6% flag rate and 0.0% recall.

These results measure self-reported trap recognition rather than actual hallucination frequency. A positive or negative self-reported flag does not establish whether the complete textual response contains an unsupported claim. Measuring actual hallucinations would require separate human adjudication of the raw model responses using a predefined rubric.

6 Qualitative Error Analysis

The four cases in Fig. 1 illustrate recurrent failure patterns. In the control scene in Fig. 1(a), all models activate a hazard response despite the absence of a major defensive-driving threat. In Fig. 1(b), the pedestrian hazard is partially recognized, but its latent timing is converted into an immediate or combined risk.

Figure 1(c) shows high-severity underestimation: the models focus on a visually salient rickshaw rather than prioritizing the vulnerable pedestrian. In Fig. 1(d), MiniCPM-V-4.6 identifies the correct bus/truck-occlusion category but still predicts incorrect timing and severity. These cases show that correct object or category recognition does not guarantee coherent defensive-driving reasoning across category, timing, severity, and action fields.

7 Limitations and Ethical Considerations

Scope and annotation. BD-HazardVLM is a focused 500-image diagnostic benchmark rather than a representative sample of all Bangladesh roads. Images were selected for scenario diversity, producing substantial class imbalance, particularly for hazard presence and risk timing. The benchmark also uses individual frames and therefore cannot directly evaluate motion, trajectories, or time-to-collision. Risk-timing and severity labels are thus image-level assessments from visible evidence; unobserved ego speed, acceleration, and object motion can change the appropriate dynamic interpretation of the same frame. Each

(a) GT: Control, none, low. All models predict a hazard.

(b) GT: Pedestrian, latent, medium. Latent risk becomes immediate or both.

(c) GT: Pedestrian, immediate, high. Models focus on a rickshaw and predict medium.

(d) GT: Bus/truck occlusion, latent, medium. Correct category still yields incorrect timing.

Fig. 1: Representative failure cases from BD-HazardVLM. The examples illustrate false hazard activation, weak latent-risk anticipation, high-severity underestimation, and inconsistent reasoning across output fields.

image received one initial annotation and centralized review, but not independent duplicate annotation. Because both reviewers were also annotators, each reviewed their own 125-image subset. We therefore do not report formal inter-annotator agreement, and the benchmark does not yet establish an independent human-agreement ceiling for the more subjective attributes.

Taxonomy and model comparability. The predefined taxonomy covers recurring Bangladesh mixed-traffic hazards but simplifies scenes containing multiple interacting risks by assigning one primary category. Moreover, although all models process the same 500 images and receive the same semantic task, they retain different native processors and numerical-precision settings because of GPU-memory constraints. Small performance differences should therefore be interpreted with care. Accordingly, the reported results characterize these three evaluated open models under the disclosed inference configurations and should not be interpreted as establishing equivalent failure rates for larger or proprietary VLMs.

Hallucination interpretation. The hallucination- and Western-bias-related outputs are model-generated self-reports. Their comparison with human trap labels measures whether a model recognizes that unsupported reasoning is possible; it does not measure the actual hallucination rate of the complete response. Direct measurement would require separate evidence-based human adjudication.

Release and responsible use. The original images will not be redistributed. We will release source identifiers, the image manifest, annotations, taxonomy documentation, prompts, model outputs, and evaluation utilities, while requiring users to obtain images from the original repositories and follow their licenses. The annotations are intended for research on multimodal reasoning and driving safety, not identification, surveillance, or safety-critical deployment. BD-HazardVLM is not a validated vehicle-safety component.

8 Conclusion

We introduced BD-HazardVLM, a 500-image Bangladesh-centric benchmark for image-level defensive-driving reasoning across hazard category, risk timing, severity, defensive actions, and reasoning traps.

Across three VLMs, category accuracy remains below 40%, timing accuracy ranges from 1.6% to 12.4%, and exact action-set matching remains below 2%. InternVL3-8B misses all 58 high-severity cases despite its leading aggregate severity accuracy, while Qwen2.5-VL-7B exhibits severe category collapse. All models show weak latent-risk anticipation and near-zero self-reported trap recognition. For these three evaluated open VLMs, the results show that coarse hazard recognition does not ensure reliable fine-grained defensive reasoning. Future work will add duplicate annotation, video evaluation, broader baselines, and subgroup analysis.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Chen, D., Zhang, Z., Cheng, L., Liu, Y., Yang, X.T.: Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and edge case evaluation. arXiv preprint arXiv:2502.00262 (2025)
3. Chen, K., Li, Y., Zhang, W., Liu, Y., Li, P., Gao, R., Hong, L., Tian, M., Zhao, X., Li, Z., Yeung, D.Y., Lu, H., Jia, X.: Automated evaluation of large vision-language models on self-driving corner cases. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7806–7815 (2025). https://doi.org/10.1109/WACV61041.2025.00759
4. Dokania, S., Hafez, A.H.A., Subramanian, A., Chandraker, M., Jawahar, C.V.: Idd-3d: Indian driving dataset for 3d unstructured road scenes. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4482–4491 (2023)
5. Godbole, M., Gao, X., Tu, Z.: Drama-x: A fine-grained intent prediction and risk reasoning benchmark for driving. arXiv preprint arXiv:2506.17590 (2025)
6. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language

models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14375–14385 (2024)
7. Huang, J., Huang, J.t., Liu, Z., Liu, X., Wang, W., Zhao, J.: Ai sees your location—but with a bias toward the wealthy world. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 18019–18039 (2025). https://doi.org/10.18653/v1/2025.emnlp-main.910
8. Islam, M.M., Das, A., Shams, K., Hasan, M.R., Hasan, K.F., Rashid, M.R.A., Chowdhury, A., Ali, M.S., Islam, M., Shahjalal, M., Masum, S.: Tfp-bd: An image dataset for traffic flow and pedestrian movement analysis on bangladeshi urban roads. Data in Brief **59**, 111398 (2025). https://doi.org/10.1016/j.dib.2025.111398
9. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.20
10. Liu, X., Wang, H., Wang, Y., Cai, J., Cao, Z., Yang, J., Lu, Z.: RoadSceneBench: A lightweight benchmark for mid-level road scene understanding. arXiv preprint arXiv:2511.22466 (2025)
11. Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., Fung, P.: Negative object presence evaluation (nope) to measure object hallucination in vision-language models. arXiv preprint arXiv:2310.05338 (2023)
12. OpenBMB: Minicpm-v 4.6 model card. https://huggingface.co/openbmb/MiniCPM-V-4.6 (2026), accessed July 2026
13. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: Lmdrive: Closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15120–15130 (2024)
14. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beisswenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: Computer Vision – ECCV 2024. pp. 256–274 (2024). https://doi.org/10.1007/978-3-031-72943-0_15
15. Tian, K., Mao, J., Zhang, Y., Jiang, J., Zhou, Y., Tu, Z.: Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. arXiv preprint arXiv:2504.03164 (2025)
16. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
17. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.V.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1743–1751 (2019). https://doi.org/10.1109/WACV.2019.00190
18. Wang, J., Chen, J., Xiao, M., Cheng, Y., Li, Y., Yin, Z.: Do-bench: An attributable benchmark for diagnosing object hallucination in vision-language models. arXiv preprint arXiv:2604.22822 (2026)
19. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22442–22452 (2025)
20. Wang, S., Cao, X., Zhang, J., Yuan, Z., Shan, S., Chen, X., Gao, W.: Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language models. arXiv preprint arXiv:2406.14194 (2024)

21. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6585–6597 (2025)
22. Xing, S., Hua, H., Gao, X., Zhu, S., Li, R., Tian, K., Li, X., Huang, H., Yang, T., Wang, Z., Zhou, Y., Yao, H., Tu, Z.: Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. arXiv preprint arXiv:2412.15206 (2024)
23. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE Robotics and Automation Letters **9**(10), 8186–8193 (2024). https://doi.org/10.1109/LRA.2024.3440097
24. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
25. Zhang, E., Gong, P., Dai, X., Huang, M., Lv, Y., Miao, Q.: Evaluation of safety cognition capability in vision-language models for autonomous driving. arXiv preprint arXiv:2503.06497 (2025)
26. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
27. Zunair, H., Khan, S., Hamza, A.B.: Rsud20k: A dataset for road scene understanding in autonomous driving. In: IEEE International Conference on Image Processing. pp. 708–714 (2024). https://doi.org/10.1109/ICIP51287.2024.10648203