





MLLM-ISU: The First-Ever Comprehensive Benchmark for Multimodal Large Language Models based Intrusion Scene Understanding

Fujun Han¹, Peng Ye^{2,3}*

¹ School of Data Science, The Chinese University of Hong Kong, Shenzhen

² Shanghai AI Laboratory ³ MMLab, The Chinese University of Hong Kong

Introduction

Previous vision-based intrusion detection works mainly focus on improving the intrusion detection performance, without a comprehensive and in-depth understanding of the intrusion scene, as shown in Tab. 1. To fill this gap, we explore a novel task called Multimodal Large Language Models based Intrusion Scene Understanding (MLLM-ISU) and report a comprehensive benchmark for the task. To the best of our knowledge, the MLLM-ISU task is proposed for the first time.

Table 1: The comparison between previous promising intrusion detection and our MLLM-ISU task. • and • denote the low and high scene understanding capabilities of different intrusion models.

| Intrusion tasks | Task Type | Structure | Modal | Scene Understanding (†) |
|-----------------|--------------------------|-----------|--------------|----------------------------|
| PIDNet [28] | Detectable Single Task | Close | Single-Modal | • |
| MF-ID [11] | Detectable Single Task | Close | Single-Modal | • |
| MM-ID [12] | Detectable Single Task | Close | Single-Modal | • |
| Ada-iD [10] | Detectable Single Task | Close | Single-Modal | • |
| MLLM-ISU | Comprehension Multi Task | Open | Multi-Modal | • |

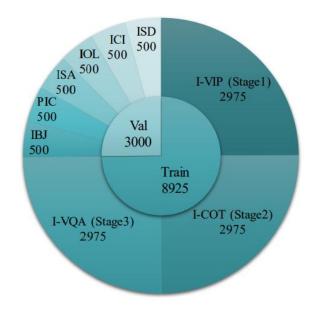
MLLM-ISU Dataset



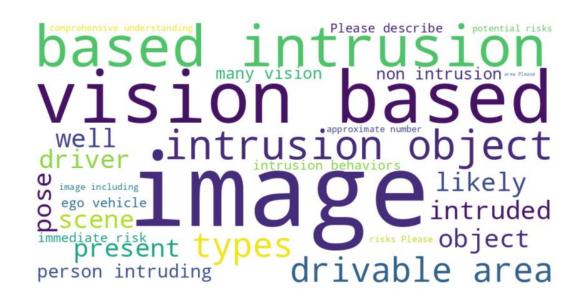


Data Collections

To accomplish the MLLM-ISU task, we first construct a new dataset, i.e., MLLM-ISU dataset.



Data statistics



The word Cloud of all questions

- · 3000 VQA Pairs for evaluation
- · 8925 training Pairs to help improve intrusion understanding performance for MLLMs





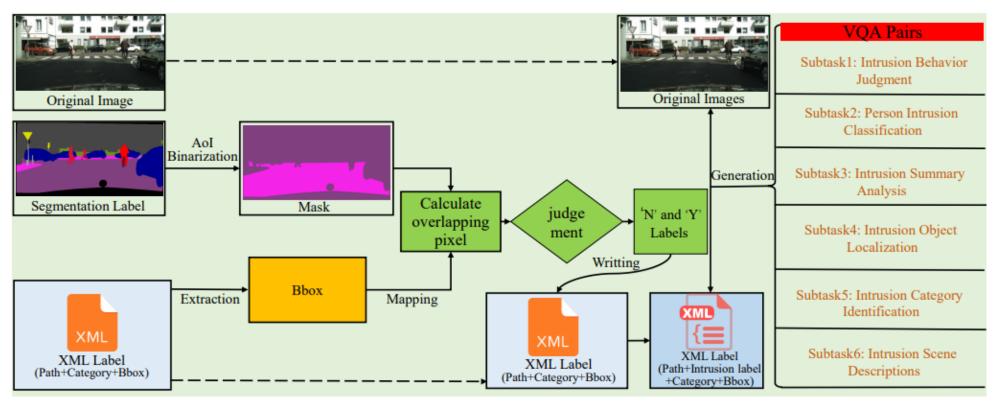
Question Definition (Six Subtasks)

- Subtask1: Intrusion Behavior Judgment (IBJ). The subtask of IBJ is used to test whether the MLLMs can correctly make intrusion judgments. It is a binary classification and can measure the model's ability to determine intrusion events.
- Subtask2: Person Intrusion Classification (PIC). The subtask of PIC is designed to test the model's intrusion judgments for individual categories. It is used to measure the model's ability to categorize intrusion events at a fine-grained level.
- · Subtask3: Intrusion Summary Analysis (ISA). ISA is used to test the ability to summarize and analyze in the intrusion scene understanding task, i.e., how many intrusion and non-intrusion behaviors are in the image.
- · Subtask4: Intrusion Object Localization (IOL). IOL is designed to measure the localization capabilities of MLLMs. This subtask is used to determine the object location for different intrusion behaviors and to identify the most dangerous intrusion objects.
- · Subtask5: Intrusion Category Identification (ICI). This subtask is used to test the ability to identify the intrusion category. Ask the MLLMs which types have intruded into the divine area. about the type of intrusion detection behavior.
- · Subtask6: Intrusion Scene Descriptions (ISD). This is an open-level question that aims to assess the ability to describe intrusion detection scenes.

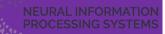
MLLM-ISU Dataset



VQA-Data Generation



The pipeline of efficient data generation strategy. Step 1: Use the segmentation label to get the AoI mask and read the Bbox of the XML document. Step 2: Calculate the overlapping pixels and give the Intrusion/Non-intrusion labels. Step 3: Write the intrusion labels for the original XML to get the new XML. Step 4: Combine the original image and new XML to generate final VQA Pairs.



Detailed cases

Intrusion Behavior Judgment



Is there any vision-based intrusion object in the image?

A: No

B: Yes

Person Intrusion Classification



Is there any person intruding into the drivable area in the image?

A: Yes

B: No

Intrusion Summary Analysis



How many vision-based intrusion or non-intrusion behaviors are present in the image?

A: 3 and 10

B: 4 and 9

C: 6 and 7

D: 7 and 10

Intrusion Object Localization



Which intrusion object is most likely to pose an immediate risk to the driver of the ego vehicle?

A: Bicycle

B: Bus

C: Motorcycle

D: Person

Intrusion Category Identification



Which types of objects have intruded into the drivable area?

A: Bus and truck

B: Motorcycle

C: Bicycle, car, person, and rider

D: Bus, train, and truck

Intrusion Scene Descriptions



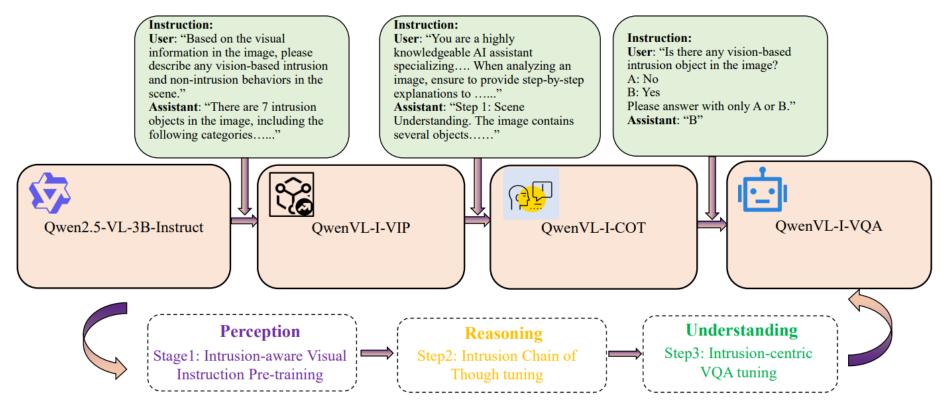
Please describe the comprehensive understanding of the scene in this image, including the types and approximate number of vision-based intrusion objects, as well as their potential risks.

The image contains 7 intrusion objects, consisting of 1 bicycle, 1 car, 5 persons. These objects have entered the restricted road region and pose potential risk to the ego vehicle.

The detailed cases of the proposed VQA evaluation pairs in the MLLM-ISU dataset

Three-stages Post-Training Framework

Training Framework



The workflow of the three-stage post-training. We take the Qwen2.5-VL-3B-Instruct as an example. For MLLMs, we first propose an Intrusion-aware Visual Instruction Pre-training strategy to obtain the initial coarse-grained perception capabilities. Then, an Intrusion Chain of Thought Tuning method is designed to enhance the ability to reason. Finally, we introduce an Intrusion-centric VQA Tuning to further enhance the capability of understanding fine-grained structures.

Experiment and Results

Experiment settings

Benchmark Evaluation Models. To evaluate and report the comprehensive performance of our proposed MLLM-ISU task, we conduct sufficient experiments and comparisons on some dominant MLLMs, i.e., Proprietary and Open-source MLLMs. For proprietary MLLMs, because the original model is not available, we adopt accessible APIs to test, e.g., Gemini-1.5-pro-latest [30], GPT-40 [19], and Claude-3.7-sonnet [3]. For Open-source MLLMs, we choose the Qwen-series models (i.e., Qwen2-VL-Instruct [34], Qwen2.5-VL-Instruct-3B [4], Qwen2.5-VL-Instruct-7B [4]), LLava-series models (i.e., LLava-1.5-7b-hf [22], LLava-1.5-13b-hf [22]), InternVL-series (i.e., InternVL2.5-1B [5], InternVL2.5-2B [5], InternVL2.5-8B [5]), and the latest InterVL3-2B [5], InterVL3-8B [5], Gemma-series (Gemma3-4B-it [29] and Gemma3-12B-it [29]), DeepSeek-VL2-tiny [35], DeepSeek-VL2-small [35], Kimi-VL-A3B-Instruct [31], Kimi-VL-A3B-Thinking [31], MiniCPM-V2.6 [37]. Unless specified, all experiments will be evaluated in a zero-shot manner.

Evaluation Metrics. In proposed MLLM-ISU task, *low-level understanding tasks* (L-U-T) and high-level understanding tasks (H-U-T) are designed to as multiple-choice. Therefore, the evaluation metrics of these subtasks are set to Accuracy, i.e., the proportion of options made correctly. For open-level understanding tasks (O-U-T), we use the BLEU-4 (BiLingual Evaluation Understudy) [26] metrics to evaluate the performance. More metrics and results can be found in **Appendix C**.

Experiment and Results



Main Results and Findings

Table 2: Comprehensive performance evaluation of the 20 dominant MLLMs on the proposed task. The **bold** and <u>underlined</u> results denote the best and second-best performance.

| Model | Source | Release | L-U-T (Easy) | | H-U-T (Difficult) | | | O-U-T (Open) | Average |
|----------------------------|-------------|---------|-----------------|-------|----------------------|-------|-------|-----------------|---------|
| | | | IBJ | PIC | ISA | IOL | ICI | ISD | |
| Human | - | - | 98.00 | 72.00 | 44.00 | 77.00 | 98.00 | 39.36 | 71.39 |
| Proprietary MLLMs | | | | | | | | | |
| GPT-4o [19] | OpenAI | 2024-08 | 41.28 | 68.27 | 29.38 | 53.23 | 83.80 | 6.74 | 47.12 |
| Gemini-1.5-pro-latest [30] | Google | 2024-04 | 7.40 | 65.40 | <u>37.00</u> | 68.40 | 68.80 | 6.04 | 42.17 |
| Claude-3.7-sonnet [3] | Anthropic | 2025-02 | 56.14 | 56.40 | 20.28 | 53.71 | 52.00 | 9.09 | 41.27 |
| Open-source MLLMs | | | | | | | | | |
| LLaVa1.5-7B-hf [22] | UW-M&Micro | 2023-10 | 94.00 | 41.80 | 22.60 | 26.60 | 33.80 | 12.12 | 38.49 |
| LLaVa1.5-13B-hf [22] | UW-M&Micro | 2023-10 | 71.00 | 62.20 | 23.60 | 30.00 | 51.80 | 12.33 | 41.82 |
| Qwen2-VL-7B-Instruct [34] | Alibaba | 2024-06 | 56.20 | 65.60 | 33.80 | 52.20 | 57.80 | 5.93 | 45.26 |
| MiniCPM-V2.6 [37] | OpenBMB | 2024-08 | 7.00 | 64.80 | 32.60 | 38.00 | 81.00 | 11.65 | 39.18 |
| InternVL2.5-1B [5] | OpenGVLab | 2024-12 | 22.80 | 59.60 | 30.60 | 27.60 | 48.00 | 12.87 | 33.58 |
| InternVL2.5-2B [5] | OpenGVLab | 2024-12 | 7.60 | 60.60 | 30.40 | 44.80 | 61.60 | 12.53 | 36.26 |
| InternVL2.5-8B [5] | OpenGVLab | 2024-12 | 28.40 | 64.60 | 28.80 | 56.80 | 80.00 | 15.27 | 45.65 |
| DeepSeek-VL2-tiny [35] | DeepSeek | 2024-12 | 64.60 | 65.40 | 32.40 | 41.80 | 73.80 | 18.79 | 49.47 |
| DeepSeek-VL2-small [35] | DeepSeek | 2024-12 | 40.60 | 64.00 | 32.80 | 74.40 | 81.00 | 8.12 | 50.15 |
| Qwen2.5-VL-3B-Instruct [4] | Alibaba | 2025-01 | 43.00 | 61.20 | 30.20 | 79.80 | 84.20 | 6.36 | 50.79 |
| Qwen2.5-VL-7B-Instruct [4] | Alibaba | 2025-01 | 16.00 | 61.40 | 24.00 | 64.80 | 81.40 | 4.68 | 42.05 |
| Gemma3-4B-it [29] | Google | 2025-03 | 52.60 | 64.20 | 25.20 | 50.40 | 54.60 | 5.93 | 42.16 |
| Gemma3-12B-it [29] | Google | 2025-03 | 58.20 | 54.80 | 26.00 | 65.40 | 81.00 | 5.73 | 48.52 |
| Kimi-VL-A3B-Instruct [31] | Moonshot AI | 2025-04 | 8.60 | 74.20 | 23.40 | 47.80 | 87.20 | 11.65 | 42.14 |
| Kimi-VL-A3B-Thinking [31] | Moonshot AI | 2025-04 | 45.40 | 41.00 | 21.40 | 25.20 | 23.40 | 5.22 | 26.97 |
| InternVL3-2B [5] | OpenGVLab | 2025-04 | 24.40 | 63.60 | 35.20 | 41.40 | 85.80 | 6.49 | 42.82 |
| InternVL3-8B [5] | OpenGVLab | 2025-04 | 12.40 | 70.60 | 34.20 | 36.40 | 63.00 | 6.58 | 37.20 |

- The best overall performance is Qwen2.5-VL-3B-Instruct, and its performance can reach 50.79%
- Current MLLMs are not capable of intrusion scene understanding
- The best MLLMs still have a wide gap with human performance.

Validation Experiments and Analyses



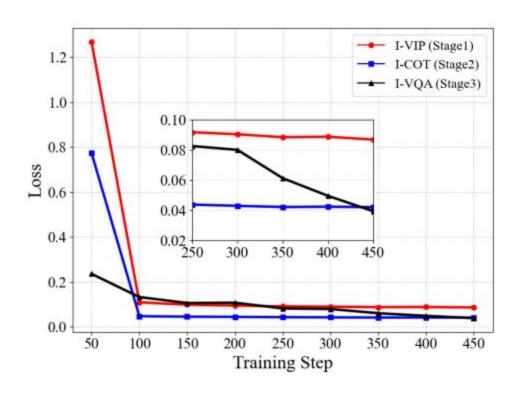
Table 3: The performance of the proposed three post-training stages on different MLLMs. I-VIP, I-COT, and I-VQA denote the proposed three different strategies in the training stages, respectively.

| Model+Method | IBJ | PIC | ISA | IOL | ICI | ISD | Average |
|--|-------|-------|-------|-------|-------|-------|---------|
| 3B Open-source MLLMs, Epoch=2 | | | | | | | |
| Qwen2.5-VL-3B-Instruct | 43.00 | 61.20 | 30.20 | 79.80 | 84.20 | 6.36 | 50.79 |
| Qwen2.5-VL-3B-Instruct+I-VIP | 49.00 | 61.60 | 28.00 | 81.00 | 86.60 | 6.03 | 52.04 |
| Qwen2.5-VL-3B-Instruct+I-VIP+I-COT | 52.20 | 58.60 | 25.00 | 83.60 | 89.60 | 24.00 | 55.50 |
| Qwen2.5-VL-3B-Instruct+I-VIP+I-COT+I-VQA | 94.60 | 67.20 | 36.80 | 88.80 | 97.20 | 48.31 | 72.15 |
| 7B Open-source MLLMs, Epoch=5 | | | | | | | |
| Qwen2.5-VL-7B-Instruct | 16.00 | 61.40 | 24.00 | 64.80 | 81.40 | 4.68 | 42.05 |
| Qwen2.5-VL-7B-Instruct+I-VIP | 85.60 | 63.40 | 23.20 | 81.40 | 91.60 | 30.51 | 62.62 |
| Qwen2.5-VL-7B-Instruct+I-VIP+I-COT | 95.40 | 66.00 | 24.60 | 82.00 | 95.20 | 54.37 | 69.60 |
| Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA | 95.80 | 78.40 | 53.60 | 92.60 | 99.40 | 50.44 | 78.37 |

We can find that, in multiple subtasks, the performance of MLLMs is improved, which verifies the effectiveness of the proposed three-stage post-training strategy. We provide a new paradigm for training and improving performance for different tasks.

Validation Experiments and Analyses

What role does each stage play for performance in the MLLM-ISU task?



Each of the three stages continuously learns the capability of coarse-grained understanding, reasoning understanding, and fine-grained understanding, respectively. These trends are reflected in the design of a three-stage progressive training strategy, and each stage plays a distinct and indispensable role.

Validation Experiments and Analyses

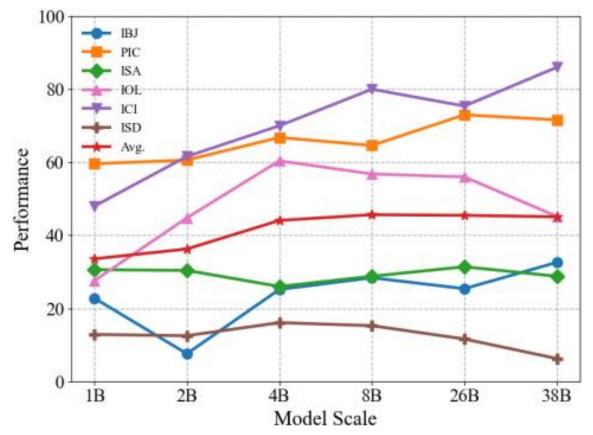
Generalization verification

Table 4: The results of generalization experiments.

| $Normal \rightarrow Foggy$ | | | | | | | | | |
|--------------------------------|---|---|--------------------------------|---|---|--|--|--|--|
| Model | Train stages | Avg. | Model | Train stages | Avg. | | | | |
| Qwen2.5-VL -3B-Instruct [4] | w/ stage1 w/ stage1&2 w/ stage1&2&3 | 50.88 51.87 55.15 72.17 | Qwen2.5-VL -7B-Instruct [4] | w/ stage1 w/ stage1&2 w/ stage1&2&3 | 44.79 61.27 69.10 77.66 | | | | |

We further go to verify the effectiveness of the three-stage training framework by the generalizability experiments. We can find that, 1) as the different stages of strategy are added, the understanding capability increases. 2) Compared with trained results in normal weather and the same setting, the results do not change much at each stage, even slightly above it, which indicates that the model has good generalization.

To what extent does model scale influence understanding performance?



We explore the relationship between the understanding performance of MLLMs and model scale. We use the InternVL2.5 series model to conduct the experiments due to its rich model scale. We can find that as the model size continues to increase, the average comprehension (red line) increases, and the best performance can be reached when the model scale is 8B. In the IOL subtask, the best performance is the 4B model, not the 38B model. These findings indicate that we need to strike a balance between model scale and subtask performance to achieve better capability in intrusion scene understanding.

Is a long chain of thought really good?



VQA

Is there any vision-based intrusion object in the image?

A: Yes

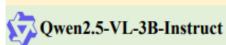
B: No

GT (A: Yes)



Qwen2.5-VL-7B-Instruct

"predict": "To determine if there is a vision-based intrusion object in the image, we need to identify any object that is not typically found in the scene or is out of place. In this image, the scene appears to be a typical urban street with parked cars, a person on a bicycle, and buildings. There are no objects that stand out as being out of place or unusual in this context. Given this analysis, the answer is: B: No"

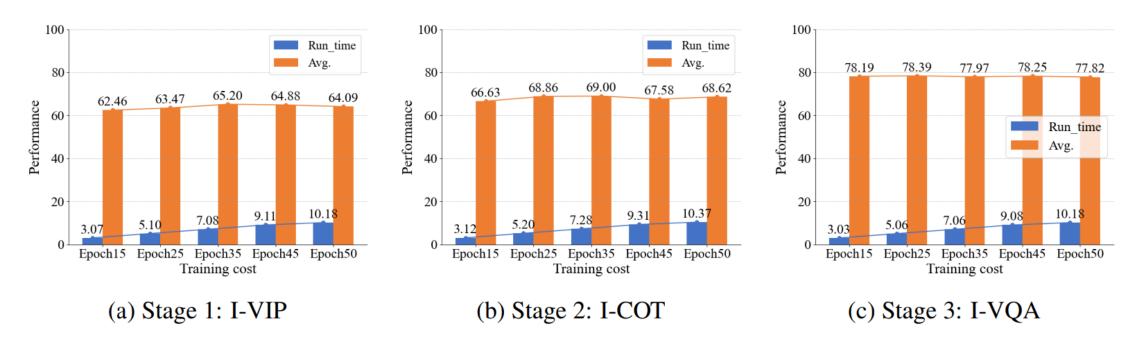


"predict": "A: Yes"



In previous experiments, we found that larger models do not perform as well as smaller models in the proposed MLLM-ISU task. To explore the reasons, we compare the processes of the two models in a simple binary classification VQA pair. We find that Qwen2.5-VL-7B-Instruct seems to think too much about the simple problems, which leads to poor or incorrect final results. The phenomenon suggests that long chain thinking is not always good.

How significant is the impact of training cost on performance



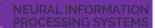
We explore the impact of training cost on performance. We can find that the average performance does not continuously improve as training costs increase. It reaches its optimum at a certain critical point. As the training cost increases, the model can lead to overfitting and poor generalization performance. Therefore, we need to balance the relationship between the training data and the cost of the model to achieve the best performance.

Processing costs and latency

Table 5: The experimental results of processing costs and latency on val (500 images).

| Model | Task | Runtime in val | Latency | Memory | Predicted_samples/s |
|-------|------|----------------|---------|----------|---------------------|
| 3B | IBJ | 5m34s | 0.668 | 9565MiB | 1.499 |
| 3B | PIC | 5m46s | 0.692 | 9565MiB | 1.446 |
| 7B | IBJ | 5m44s | 0.688 | 18441MiB | 1.453 |
| 7B | PIC | 5m47s | 0.694 | 18441MiB | 1.443 |

We test the runtime, latency, and efficiency on two different models without any inference acceleration strategy, i.e., 3B and 7B. The input size of the image is 1024×2048 , as shown in Tab. 5. We can find that our model has a low latency of 0.668 and a high predicted sample rate of 1.499.





Dataset diversity

Table 6: Comprehensive performance evaluation of the 18 dominant MLLMs on the MLLM-ISU-BDD. The **bold** and underlined results denote the best and second-best performance.

| Model | Source | Release | IBJ | PIC | ISA | IOL | ICI | ISD | Average |
|----------------------------|-------------|---------|-------|-------|-------|-------|-------|-------|---------|
| Human | - | - | 90.00 | 96.67 | 38.89 | 85.56 | 96.67 | 35.61 | 73.90 |
| GPT-4o [19] | OpenAI | 2024-08 | 19.38 | 84.41 | 27.62 | 62.14 | 89.98 | 6.40 | 48.32 |
| LLaVa1.5-7B-hf [22] | UW-M&Micro | 2023-10 | 92.65 | 21.60 | 21.38 | 24.94 | 35.41 | 12.03 | 34.67 |
| LLaVa1.5-13B-hf [22] | UW-M&Micro | 2023-10 | 71.71 | 79.29 | 21.60 | 27.17 | 32.52 | 12.32 | 40.77 |
| Qwen2-VL-7B-Instruct [34] | Alibaba | 2024-06 | 25.84 | 86.19 | 36.53 | 50.56 | 59.91 | 6.07 | 44.18 |
| MiniCPM-V2.6 [37] | OpenBMB | 2024-08 | 5.79 | 83.96 | 30.29 | 33.85 | 77.95 | 12.44 | 40.71 |
| InternVL2.5-1B [5] | OpenGVLab | 2024-12 | 24.05 | 79.73 | 25.61 | 27.84 | 35.86 | 12.53 | 34.27 |
| InternVL2.5-2B [5] | OpenGVLab | 2024-12 | 5.35 | 83.96 | 30.73 | 45.66 | 43.88 | 11.69 | 36.88 |
| InternVL2.5-8B [5] | OpenGVLab | 2024-12 | 24.05 | 77.73 | 29.84 | 61.92 | 66.37 | 13.80 | 45.62 |
| DeepSeek-VL2-tiny [35] | DeepSeek | 2024-12 | 73.05 | 86.64 | 30.73 | 51.22 | 76.61 | 18.25 | 56.08 |
| DeepSeek-VL2-small [35] | DeepSeek | 2024-12 | 32.07 | 85.30 | 27.39 | 64.59 | 79.51 | 9.20 | 49.68 |
| Qwen2.5-VL-3B-Instruct [4] | Alibaba | 2025-01 | 35.41 | 77.73 | 29.18 | 89.76 | 80.62 | 5.79 | 53.08 |
| Qwen2.5-VL-7B-Instruct [4] | Alibaba | 2025-01 | 18.49 | 82.85 | 26.50 | 83.52 | 86.41 | 4.47 | 50.37 |
| Gemma3-4B-it [29] | Google | 2025-03 | 49.00 | 84.19 | 22.49 | 48.11 | 32.07 | 5.73 | 40.27 |
| Gemma3-12B-it [29] | Google | 2025-03 | 49.00 | 43.65 | 24.28 | 71.94 | 83.07 | 5.60 | 46.26 |
| Kimi-VL-A3B-Instruct [31] | Moonshot AI | 2025-04 | 16.93 | 86.86 | 26.50 | 36.08 | 88.42 | 17.21 | 45.33 |
| Kimi-VL-A3B-Thinking [31] | Moonshot AI | 2025-04 | 44.10 | 20.27 | 20.94 | 24.50 | 20.71 | 5.08 | 22.60 |
| InternVL3-2B [5] | OpenGVLab | 2025-04 | 18.04 | 75.72 | 33.18 | 42.32 | 77.28 | 6.57 | 42.19 |
| InternVL3-8B [5] | OpenGVLab | 2025-04 | 8.91 | 87.53 | 25.17 | 37.86 | 42.76 | 6.13 | 34.73 |

1) To verify the universality of the proposed pipeline of VQA-Data Generation and enhance the diversity of intrusion scene types in realworld environments, we create a new benchmark dataset based on the BDD-100K for the MLLM-ISU task, namely MLLM-ISU-BDD. 2) We can find that current MLLMs are not capable of intrusion scene understanding, especially in harsh and volatile scenarios. Besides, we can also find that even the best models still have a wide gap with human performance.

Discussions



Exploration of light task-specific strategy

Table 7: The quantitative results with the light taskspecific strategy. Avg. denotes the average of six sub-tasks in different datasets.

| Model | Method | Avg. w/ MLLM-ISU-CS | Avg. w/ MLLM-ISU-BDI | |
|---------------------------|-------------------|---------------------|----------------------|--|
| | Zero-shot | 42.16 | 40.27 | |
| Gemma3-4B-it [29] | w/ light strategy | 45.60 | 43.14 | |
| | Performance Gain | +3.44 | +2.87 | |
| | Zero-shot | 42.14 | 45.33 | |
| Kimi-VL-A3B-Instruct [31] | w/ light strategy | 44.38 | 47.90 | |
| | Performance Gain | +2.24 | +2.57 | |
| | Zero-shot | 42.82 | 42.19 | |
| InternVL3-2B [5] | w/ light strategy | 44.09 | 46.26 | |
| | Performance Gain | +1.27 | +4.07 | |

We can find that our strategy can improve the upper-bound performance, achieving average performance gains of 1.27 to 4.07, which proves the validity of our light task-specific strategy.

In this paper

- We explore a novel task called Multimodal Large Language Models based Intrusion
 Scene Understanding (MLLM-ISU) and report a comprehensive benchmark for the
 task. To the best of our knowledge, the MLLM-ISU task is proposed for the first time.
- We introduce an effective automatic visual question-answer generation pipeline, overcoming the limitation that existing datasets are not directly applicable to the MLLM-ISU task. Besides, an effective post-training framework with sequential training strategies is designed to improve the performance of the MLLM-ISU task.
- We conduct comprehensive experiments to evaluate the intrusion understanding capability of current MLLMs and verify the effectiveness of the proposed framework.