Appendix for MLLM-ISU

A More VQA evaluation pairs cases in the MLLM-ISU dataset

In this subsection, we provide more VQA Paris cases. In every subtask, we report two different cases, as shown in Fig. 11. The VQA pairs contain six subtasks and are designed to evaluate the comprehensive and in-depth understanding capability of the current MLLMs. Intrusion Behavior Judgment and Person Intrusion Classification are the choice questions on binary classification. These two subtasks are used to test the capability of basic understanding and are relatively easy. Intrusion Summary Analysis, Intrusion Object Localization, and Intrusion Category Identification are used to test the capability of deeper levels of understanding and are relatively difficult. Intrusion Scene Descriptions is an open subtask and is designed to test the capability for open scene understanding. Our subtask is diverse and rich. The design VQA pairs can meet the requirements of the MLLM-ISU task and provide the foundation for the task.

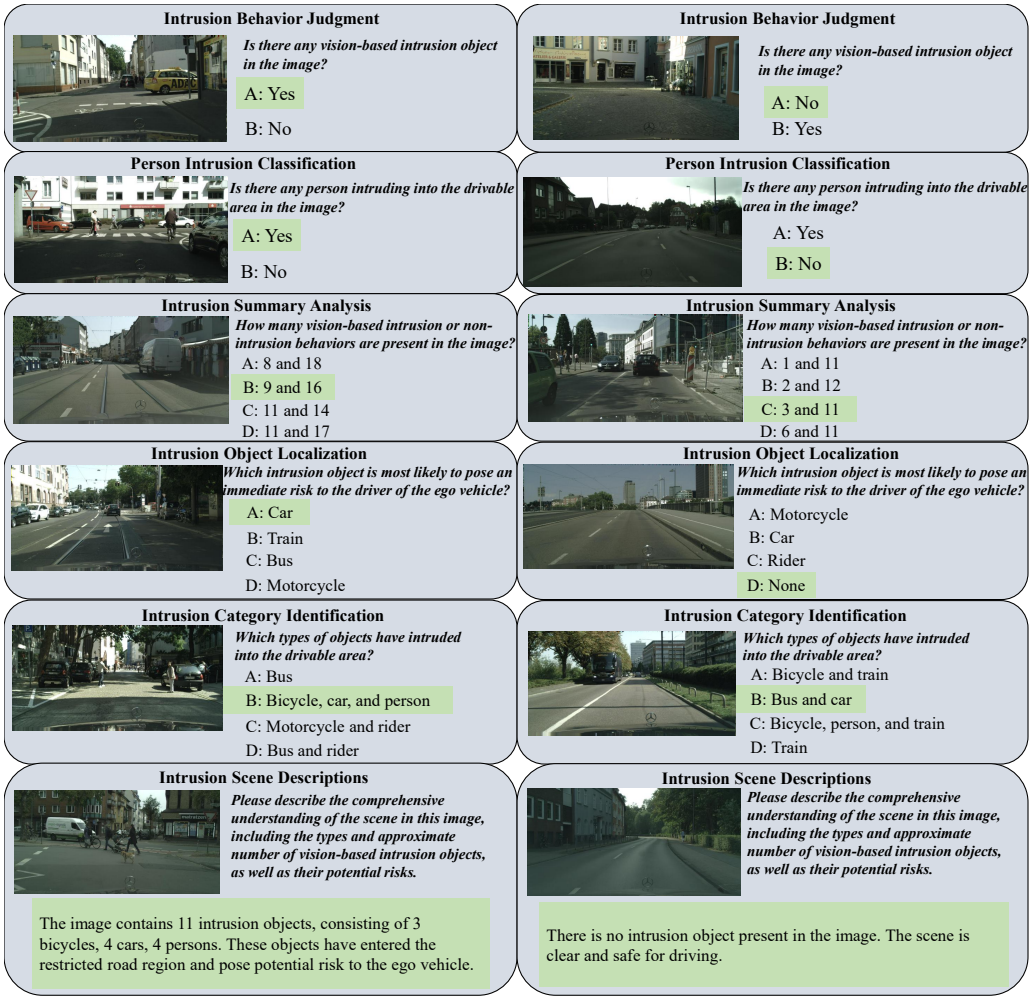


Figure 11: More VQA evaluation pairs cases in the MLLM-ISU dataset.

B Detailed prompt template of Intrusion-aware Visual Instruction Pre-training

In this subsection, we introduce the detailed prompt template for our Intrusion-aware Visual Instruction Pre-training strategy, as shown in Fig. 12. We show two different prompt template cases, intrusion

(Upper) and non-intrusion (Lower) cases, respectively. We use the same system prompt: *<Based on the visual information in the image, please describe any vision-based intrusion and non-intrusion behaviors in the scene.>*. Then, we give the corresponding answers. In this way, we construct the training dataset and perform Lora supervised fine-tuning to get the initial perception capabilities for MLLMs.

<user> Based on the visual information in the image, please describe any vision-based intrusion and non-intrusion behaviors in the scene.

<assistant> There are 9 intrusion objects in the image, including the following categories: 'bicycle, person, rider'. Besides, there are also 10 non-intrusion objects, including 'bus, person'. The object posing the greatest potential risk is a 'person', The main reason is that the category 'person' is closest to the driver of the ego vehicle and may pose a potential risk for the driver. Therefore, caution is advised, and the driver may need to take some necessary action.

<user> Based on the visual information in the image, please describe any vision-based intrusion and non-intrusion behaviors in the scene.

<assistant> There is no vision intrusion object in the image. All 15 detected objects are non-intrusions, including the following categories: car, motorcycle, person. Therefore, the driver can safely drive.

Figure 12: Two different templates for Supervised Fine-Tuning in I-VIP (stage 1).

C More metrics and experimental results

In subsection **5.2**, for the subtask: Intrusion Scene Descriptions (ISD), we mainly adopt the **BLEU-4** to evaluate the performance. In this subsection, to calculate and provide a richer baseline, we also report some other metrics and performance for this subtask on some MLLMs, *i.e.*, Rouge-1, Rouge-2, Rouge-L, as shown in Tab. 8. Note that MLLM-ISU-CS and MLLM-ISU-BDD denote the benchmark datasets that are built based on Cityscape and BDD-100K datasets for our MLLM-ISU task, respectively.

Table 8: Some other metrics results for the ISD subtask.

Model	Source	Release	MLLM-ISU-CS			MLLM-ISU-BDD		
			Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
GPT-4o[19]	OpenAI	2024-08	15.75	1.80	7.50	14.78	1.62	7.03
LLaVa1.5-13B-hf[22]	UW-M&Micro	2023-10	19.95	3.33	12.29	20.64	3.39	12.32
MiniCPM-V2.6 [37]	OpenBMB	2024-08	19.14	2.21	11.42	19.86	2.48	12.07
InternVL2.5-2B[5]	OpenGVLab	2024-12	18.73	2.23	11.54	18.63	2.30	10.84
InternVL2.5-8B[5]	OpenGVLab	2024-12	21.49	2.34	13.40	20.10	1.96	12.07
DeepSeek-VL2-tiny[35]	DeepSeek	2024-12	22.66	4.31	16.49	23.02	4.43	16.08
DeepSeek-VL2-small[35]	DeepSeek	2024-12	20.41	2.32	8.60	19.09	1.65	8.98
Qwen2.5-VL-3B-Instruct[4]	Alibaba	2025-01	16.44	1.72	6.94	14.61	1.53	6.19
Qwen2.5-VL-7B-Instruct[4]	Alibaba	2025-01	14.74	1.64	5.07	13.75	1.57	4.86
Gemma3-4B-it[30]	Google	2025-03	14.56	1.92	5.31	13.82	1.74	5.07
Gemma3-12B-it[30]	Google	2025-03	13.77	1.45	5.40	13.73	1.53	5.23
Kimi-VL-A3B-Instruct[31]	Moonshot AI	2025-04	21.50	2.90	11.30	24.04	3.16	15.07
Kimi-VL-A3B-Thinking[31]	Moonshot AI	2025-04	13.98	1.61	5.41	13.43	1.72	5.46
InternVL3-2B[5]	OpenGVLab	2025-04	15.75	1.44	6.67	15.42	1.49	6.63
InternVL3-8B[5]	OpenGVLab	2025-04	16.53	1.64	6.67	15.42	1.65	6.21

D More training details

In this appendix, we present more details of the training for the proposed three-stage framework. For the cutoff_len parameters, we use 2048. For the learning_rate, we adopt the default setting, *i.e.*, 5e-5. All stages adopt the Lora as a supervised fine-tuning method.

Table 9: The detail setting for the training experiments.

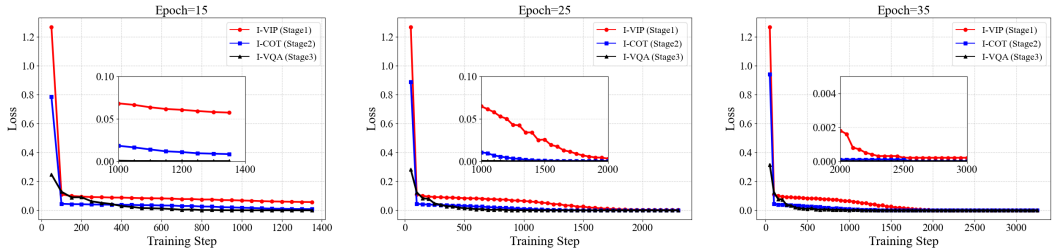
Setting	Value	Setting	Value
cutoff_len (stage1)	2048	preprocessing_num_workers	16
cutoff_len (stage2)	2048	per_device_train_batch_size	1
cutoff_len (stage3)	2048	per_device_eval_batch_size	1
gradient_accumulation_steps	8	learning_rate	5e-5
num_train_epochs (3B/7B)	2/5	finetuning_type	lora

E More metrics and results of three post-training stages

We further verify the effectiveness of the proposed three-Stage Post-Training Framework. Like the previous experiment, we use the Qwen2.5-VL-7B-Instruct to conduct the experiment in five different epochs, *i.e.*, Epoch=15, 25, 35, 45, 50, as shown in Tab. 10. We can find that as the different stages are added, the performance increases and reaches 78.19%, 78.39%, 77.97%, 78.25%, and 77.82%, respectively. Besides, in different subtasks, our framework can also give a performance gain, which denotes that the three different Supervised Fine-tuning strategies are effective, *i.e.*, Perception (Intrusion-aware Visual Instruction Pre-training)→Reasoning (Intrusion Chain of Thought Tuning)→Understanding (Intrusion-centric VQA Tuning). We also give the training loss in different epochs, as shown in Fig. 13. We can find that in different epochs, models can learn the different capabilities of the three stages. As the training step increases, the loss of the model changes less, especially after 1500 steps. Therefore, we believe it is important to choose appropriate training steps.

Table 10: More performance results 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.

Model+Method	IBJ	PIC	ISA	IOL	ICI	ISD	Average
7B Open-source MLLMs, Epoch=15							
Qwen2.5-VL-7B-Instruct [4]	16.00	61.40	24.00	64.80	81.40	4.68	42.05
Qwen2.5-VL-7B-Instruct+I-VIP	81.00	65.60	24.40	81.80	92.00	29.93	62.46
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT	95.40	68.40	22.80	81.60	96.60	34.97	66.63
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA	97.00	79.40	48.80	93.20	99.40	51.33	**78.19**
7B Open-source MLLMs, Epoch=25							
Qwen2.5-VL-7B-Instruct [4]	16.00	61.40	24.00	64.80	81.40	4.68	42.05
Qwen2.5-VL-7B-Instruct+I-VIP	80.80	65.80	27.40	83.00	94.00	29.84	63.47
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT	92.60	72.40	29.40	84.00	95.40	39.35	68.86
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA	95.20	77.80	54.00	93.40	99.20	50.73	**78.39**
7B Open-source MLLMs, Epoch=35							
Qwen2.5-VL-7B-Instruct [4]	16.00	61.40	24.00	64.80	81.40	4.68	42.05
Qwen2.5-VL-7B-Instruct+I-VIP	87.60	67.40	29.80	82.40	94.20	29.79	65.20
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT	94.20	69.20	31.80	83.80	96.40	38.58	69.00
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA	96.40	77.60	52.00	92.00	99.20	50.61	**77.97**
7B Open-source MLLMs, Epoch=45							
Qwen2.5-VL-7B-Instruct [4]	16.00	61.40	24.00	64.80	81.40	4.68	42.05
Qwen2.5-VL-7B-Instruct+I-VIP	84.40	66.60	27.60	85.80	95.00	29.90	64.88
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT	94.00	66.60	27.60	84.60	96.80	35.89	67.58
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA	96.20	78.40	51.00	93.40	99.20	51.28	**78.25**
7B Open-source MLLMs, Epoch=50							
Qwen2.5-VL-7B-Instruct [4]	16.00	61.40	24.00	64.80	81.40	4.68	42.05
Qwen2.5-VL-7B-Instruct+I-VIP	84.80	66.40	27.60	81.80	94.00	29.94	64.09
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT	93.80	66.20	31.80	83.60	95.40	40.89	68.62
Qwen2.5-VL-7B-Instruct+I-VIP+I-COT+I-VQA	95.60	78.00	52.60	91.60	98.80	50.29	**77.82**



(a) Epoch 15 (b) Epoch 25 (c) Epoch 35

Figure 13: The training loss in three different epochs.

F More generalization verification experiments

In this appendix, we further conduct more generalization verification experiments and report more experimental results for the three-stage training framework. Specifically, we use the three different foggy coefficients to conduct, *i.e.*, α=0.02, α=0.01, and α=0.005, respectively. The Qwen2.5-VL-3B-Instruct [4] and Qwen2.5-VL-7B-Instruct [4] are chosen as the baseline model, and the result is shown in Tab. 11. We can find that our three-stage training framework has strong generalization performance and shows promising performance on several different tasks.

Table 11: More generalization results of the proposed three post-training stages on different tasks. Note that our strategy is to increase them one by one. α denotes the foggy coefficients in Cityscape [7].

Normal→Foggy, $\alpha = 0.02$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -3B-Instruct [4]	-	47.00	61.20	31.00	78.00	81.60	6.51	50.88
	w/ stage1	51.60	61.60	28.60	78.80	83.80	6.81	51.87
	w/ stage1&2	55.80	57.60	25.00	81.60	85.60	25.30	55.15
	w/ stage1&2&3	94.40	67.80	37.40	87.40	97.60	48.41	**72.17**
Normal→Foggy, $\alpha = 0.01$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -3B-Instruct [4]	-	45.80	62.20	30.60	78.20	82.40	6.58	50.96
	w/ stage1	50.80	61.40	28.00	80.00	84.40	6.67	51.88
	w/ stage1&2	55.20	58.20	24.60	81.20	87.40	25.02	55.27
	w/ stage1&2&3	94.40	68.20	36.60	88.40	97.40	48.22	**72.20**
Normal→Foggy, $\alpha = 0.005$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -3B-Instruct [4]	-	45.00	60.80	30.40	78.40	82.40	6.54	50.59
	w/ stage1	50.80	61.80	27.80	81.00	84.00	6.73	52.02
	w/ stage1&2	55.00	58.00	24.60	83.80	88.00	25.42	55.80
	w/ stage1&2&3	94.40	67.60	38.20	88.00	97.60	48.34	**72.36**
Normal→Foggy, $\alpha = 0.02$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -7B-Instruct [4]	-	31.20	62.60	24.20	63.60	82.40	4.72	44.79
	w/ stage1	76.40	63.40	23.80	80.80	92.80	30.41	61.27
	w/ stage1&2	94.00	63.00	24.20	82.60	96.00	54.81	69.10
	w/ stage1&2&3	95.20	77.40	49.80	94.00	99.20	50.33	**77.66**
Normal→Foggy, $\alpha = 0.01$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -7B-Instruct [4]	-	26.60	63.00	24.40	61.00	82.00	4.75	43.63
	w/ stage1	81.00	63.60	23.80	80.60	92.00	30.41	61.90
	w/ stage1&2	94.40	64.20	24.80	81.60	96.00	55.24	69.37
	w/ stage1&2&3	95.60	79.20	51.20	94.20	99.20	50.34	**78.29**
Normal→Foggy, $\alpha = 0.005$								
Model	Train stages	IBJ	PIC	ISA	IOL	ICI	ISD	Avg.
Qwen2.5-VL -7B-Instruct [4]	-	22.20	62.40	24.60	62.00	81.00	4.72	42.82
	w/ stage1	85.40	63.60	24.00	79.60	91.40	30.67	62.45
	w/ stage1&2	94.60	65.40	24.60	82.20	96.00	55.79	69.77
	w/ stage1&2&3	96.20	79.60	52.80	93.60	99.20	50.87	**78.71**

G More model scale results on InternVL3-series models

In addition to InternVL2.5-series models, we also conduct model scale experiments in the latest InternVL3-series model, as shown in Fig. 14. We can find that, like the InternVL2.5-series model, the best average performance can be reached when the model scale is 9B, not the largest 38B model. We think this phenomenon has something to do with overthinking the model, where overthinking simple problems instead creates illusions. This is something we need to study further.

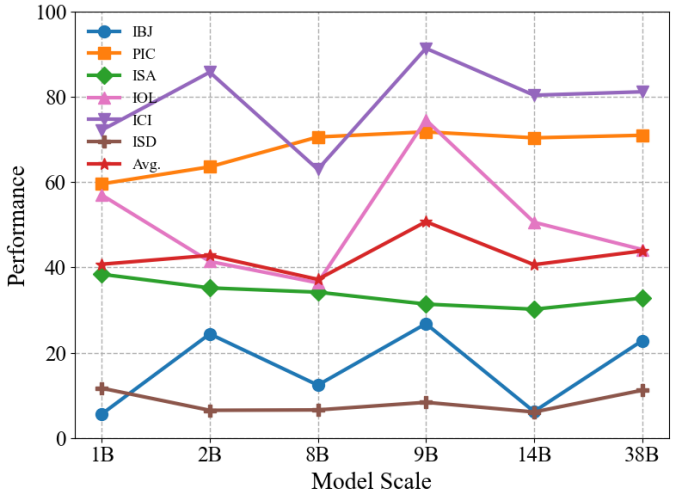


Figure 14: The model scale results on InternVL3-series

H The detailed information for proposed MLLM-ISU-BDD

To verify the universality of the proposed pipeline of VQA-Data Generation and enhance the diversity of intrusion scene types in real-world environments, we create a new benchmark dataset for the MLLM-ISU task, namely MLLM-ISU-BDD. The MLLM-ISU-BDD is built based on the BDD-100K datasets. The detailed method can refer to Fig. 3. Our new MLLM-ISU-BDD datasets contain rich intrusion scene types, *e.g.*, multiple different weather (Clear, Cloudy, Rainy, Foggy, Night), different geographic environment (City, Highway, Suburban/Rural), different period of time (Daytime, Dusk, Night), and Different transportation environments (Heavy Traffic, Empty Road). We clean the original dataset based on the proposed intrusion detection task features. Finally, our datasets contain 8892 training Pairs and 2694 VQA evaluation Pairs. Our extended benchmark explicitly includes nighttime, adverse weather, and non-urban roads, enabling more comprehensive evaluation of the intrusion scene understanding task in real-world environments.

I More discussion and interesting finding

Discussion on model version performance variations. In Tab. 2 and Tab. 6, we observed that newer versions of multimodal large models (MLLMs) do not always outperform their predecessors on our proposed intrusion scene understanding task, *e.g.*, in Tab. 2, InternVL3-8B is lower than InternVL2.5-8B, the interesting phenomenon also occurs in previous work [18]. We think this is reasonable. The main reason is that, during the model update process, priority is typically given to improving abstract reasoning, instruction-following, and general linguistic capabilities rather than low-level visual perception. Consequently, newer models may excel in complex reasoning and multi-turn understanding but show reduced sensitivity to small, partially occluded, or contextually subtle intrusion targets. Besides, stronger reasoning abilities typically imply longer chains of thought. However, in section 5.4, we find that longer reasoning isn't always effective for our tasks. These findings indicate that when applying general MLLMs to visual intrusion understanding, we need to explore the task-specific adaptation strategies to enhance their perception and recognition of fine-grained intrusion cues.

J Limitations

To the best of our knowledge, the MLLM-ISU is proposed for the first time and is the first attempt in the intrusion detection field. We believe our work will produce positive effects in several application areas, *e.g.*, autonomous driving, intelligent monitoring, and security. We believe designing more comprehensive benchmarks, *e.g.*, richer understanding tasks, and exploring more efficient improvement training or training-free strategies, *e.g.*, training-free reasoning method (Retrieving Augmented Generation), is a worthy research direction in the future.