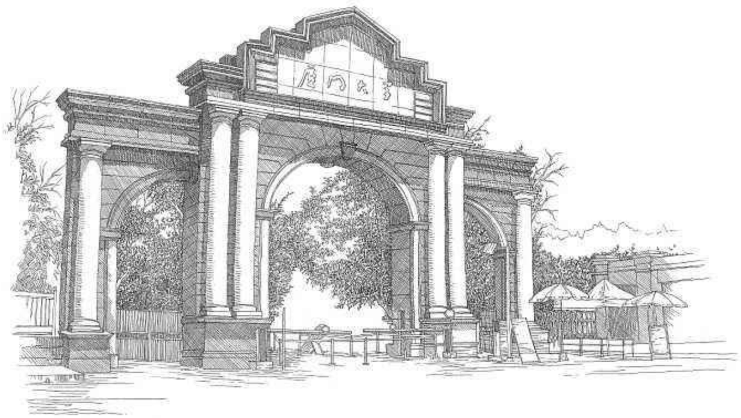
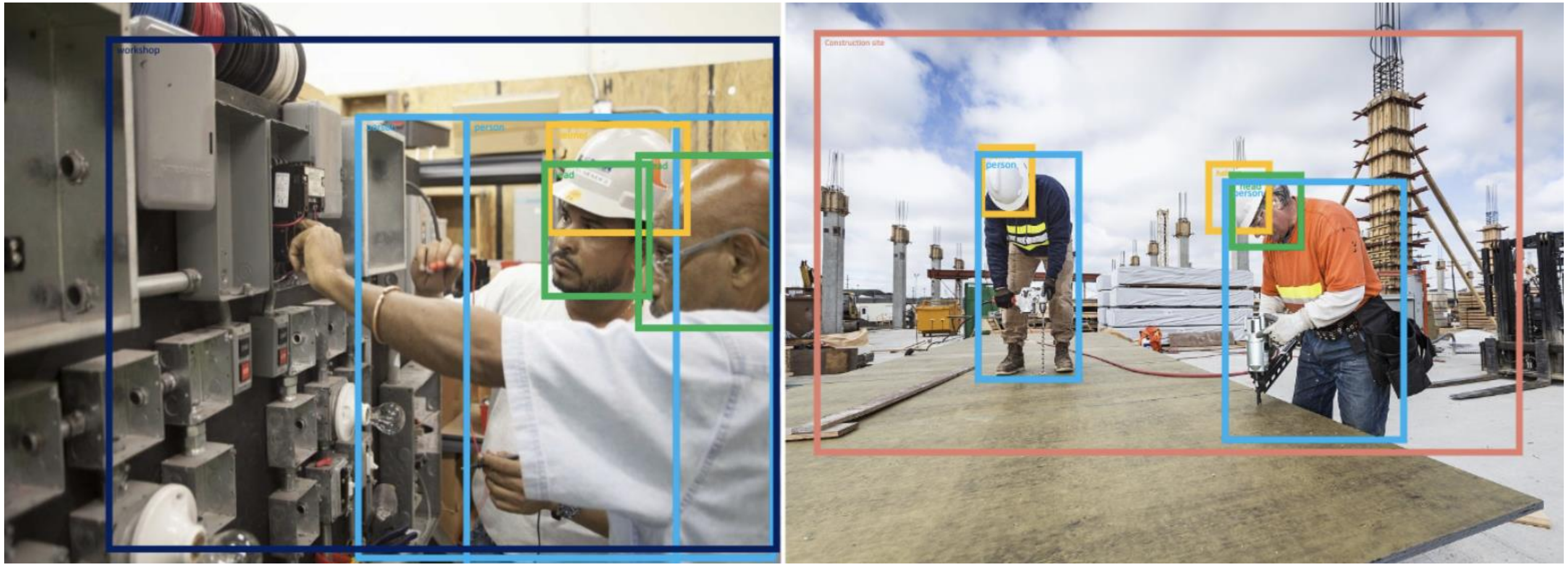
From Objects to Events: Unlocking Complex Visual Understanding in Object Detectors via LLM-guided Symbolic Reasoning

Yuhui Zeng[1*], Haoxiang Wu[1*], Wenjie Nie[1], Guangyao Chen[2†], Xiawu Zheng[1†],
Yunhang Shen[3], Jun Peng[1], Yonghong Tian[2], Rongrong Ji[1]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[2] Peking University [3]Tencent Youtu Lab



Motivation

- **Current Status of Object Detection：**
  - □Contemporary detectors localize and categorize entities with high precision, directly addressing “where is the object?”



“Discrete object recognizers” lack modeling of inter-object relations and contextual semantics, failing to answer depicted actions.

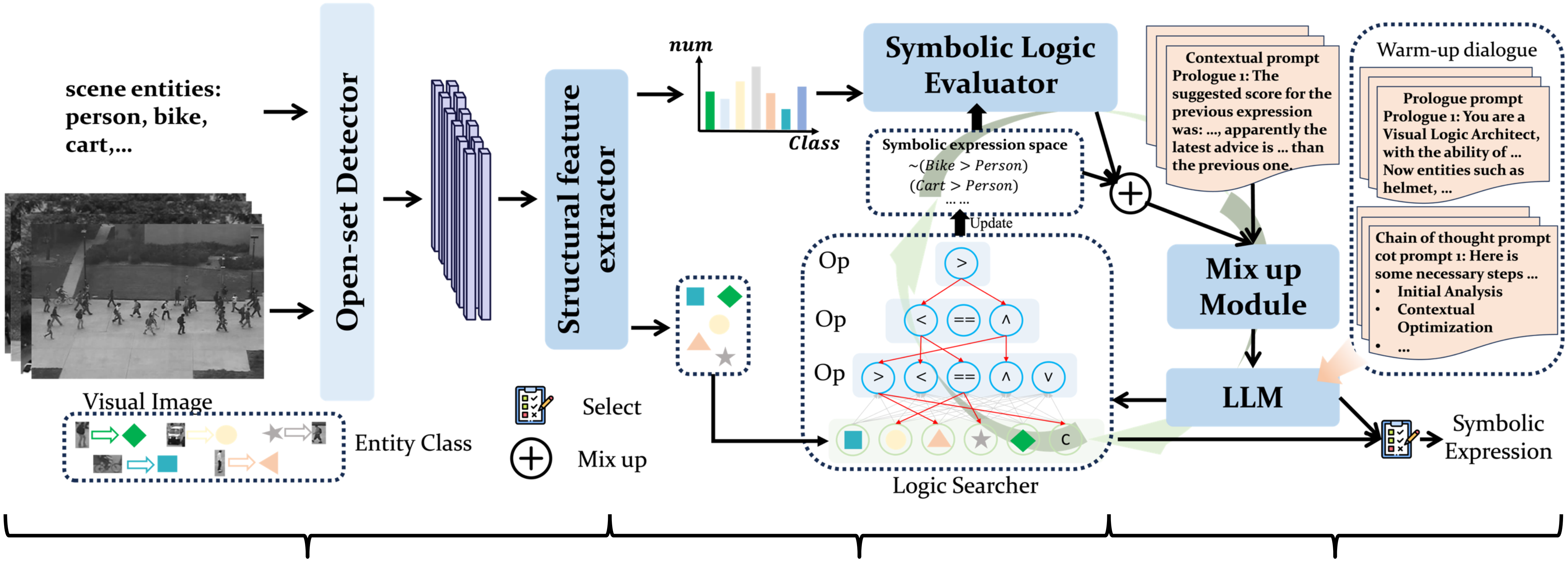
Motivation

- **Solutions for "Black-box" Models**
  - **Dedicated Event Recognition Systems**
    - **Higher precision**
    - **Requires specialized training, annotation cost, limited generalization**
  - **Multimodal Large Models**
    - **Issues in accuracy, interpretability, and verifiability**
    - **Stability and hallucination problems**

- **SymbolicDet: Detector + Symbolic Search**
  - **Task decoupling, transparent and readable conclusions**
    - **Detector + evolutionary search with LLM for reasoning**
    - **Decisions derived from logical expressions**
  - **Training-free, modular**
    - **Resource-efficient, highly portable**
  - **Significant performance gains**
    - **Substantial room for improvement in search efficiency and pattern quality**

Method

SymbolicDet Framework



Contextual prompt
Prologue 1: The suggested score for the previous expression was: ..., apparently the latest advice is ... than the previous one.

Prologue prompt
Prologue 1: You are a Visual Logic Architect, with the ability of ... Now entities such as helmet, ...

Chain of thought prompt
cot prompt 1: Here is some necessary steps ...
- Initial Analysis
- Contextual Optimization
- ...

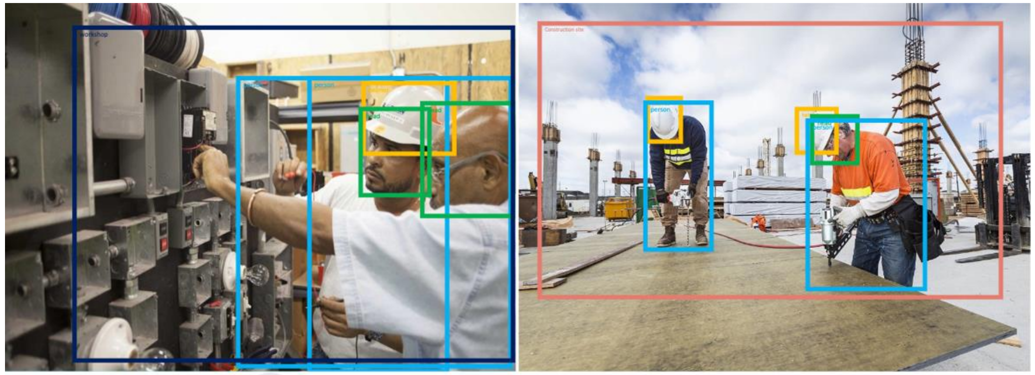
Scene Object Detection

Symbolic Search

Logical Reasoning

Method

Symbolic Logic Search

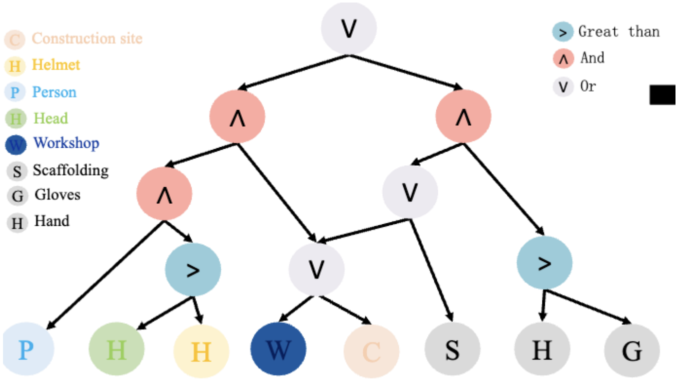


evaluating

searching

reasoning

Symbolic Regressor



Symbolic Pattern: (Head > Helmet) ∧ Person ∧ (Workshop ∨ Construction_Site) ∨ (Hand > Gloves) ∧ (Workshop ∨ Construction_Site ∨ Scaffolding)

Pattern Explain: This expression describes a workplace safety visual pattern requiring either a person wearing a helmet or hands wearing gloves in a workshop or construction site environment.

LLM

Prompt space

Experiment

Benchmark

- **Public Dataset**
  - **ERA Dataset**
  - **UCSD Ped2 Dataset**
  - **USED Dataset**

ERA DATASET
A Dataset and Deep Learning Benchmark for Event Recognition in Aerial Videos

- **Self-constructed Dataset**
  - **Multi-Event Dataset**
  - **Helmet-Mac Dataset**

USED: A Large Scale Social Event Detection Dataset

mm lab

Experiment

Experimental results of different detectors on multiple data

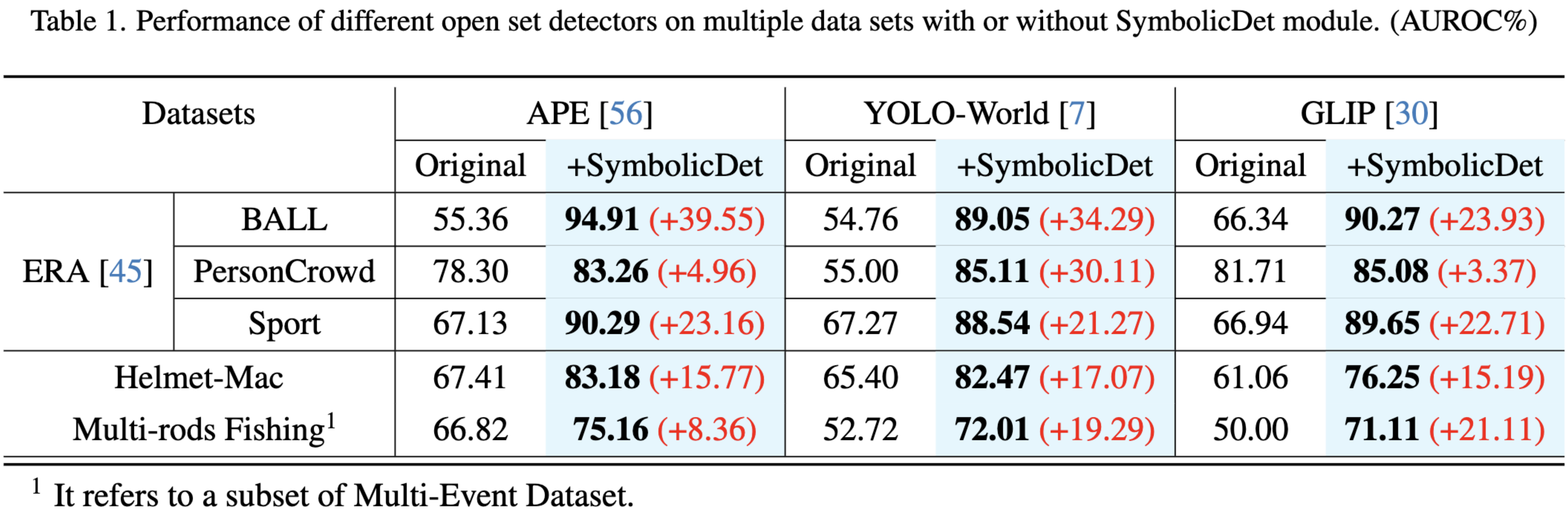


Table 1. Performance of different open set detectors on multiple data sets with or without SymbolicDet module. (AUROC%)

Datasets		APE [56]		YOLO-World [7]		GLIP [30]	
		Original	+SymbolicDet	Original	+SymbolicDet	Original	+SymbolicDet
ERA [45]	BALL	55.36	**94.91** (+39.55)	54.76	**89.05** (+34.29)	66.34	**90.27** (+23.93)
	PersonCrowd	78.30	**83.26** (+4.96)	55.00	**85.11** (+30.11)	81.71	**85.08** (+3.37)
	Sport	67.13	**90.29** (+23.16)	67.27	**88.54** (+21.27)	66.94	**89.65** (+22.71)
Helmet-Mac		67.41	**83.18** (+15.77)	65.40	**82.47** (+17.07)	61.06	**76.25** (+15.19)
Multi-rods Fishing[1]		66.82	**75.16** (+8.36)	52.72	**72.01** (+19.29)	50.00	**71.11** (+21.11)

[1] It refers to a subset of Multi-Event Dataset.

Experiment

Comparison of experimental results on multiple public data

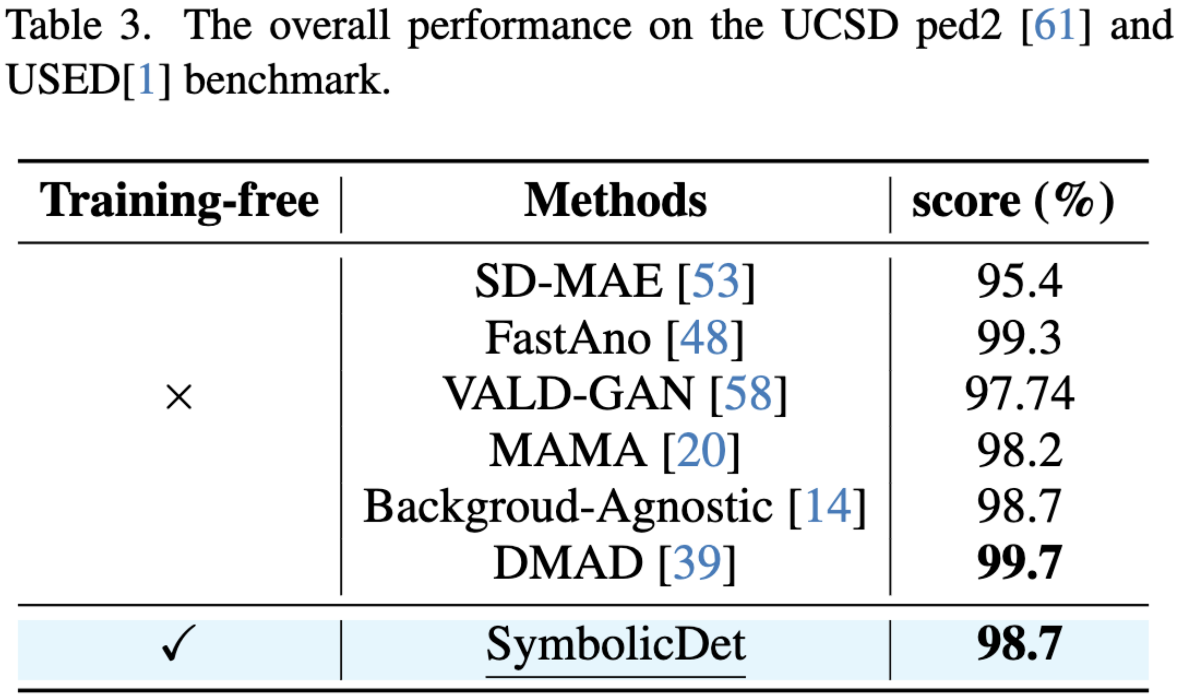
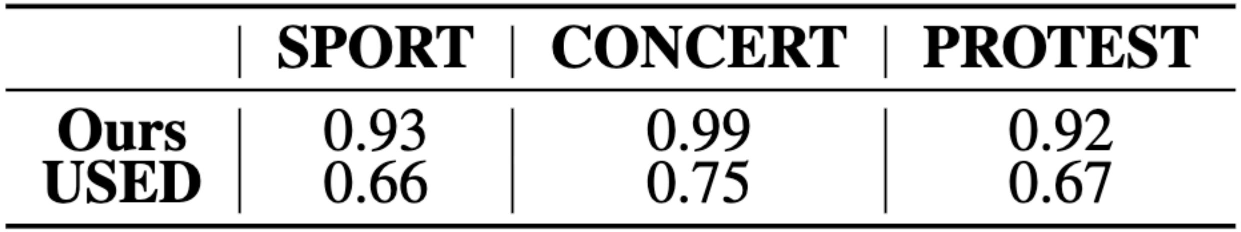


Table 3. The overall performance on the UCSD ped2 [61] and USED[1] benchmark.

Training-free	Methods	score (%)
×	SD-MAE [53]	95.4
	FastAno [48]	99.3
	VALD-GAN [58]	97.74
	MAMA [20]	98.2
	Backgroud-Agnostic [14]	98.7
	DMAD [39]	**99.7**
✓	SymbolicDet	**98.7**



	SPORT	CONCERT	PROTEST
Ours	0.93	0.99	0.92
USED	0.66	0.75	0.67

Experiment

Ablation experiment

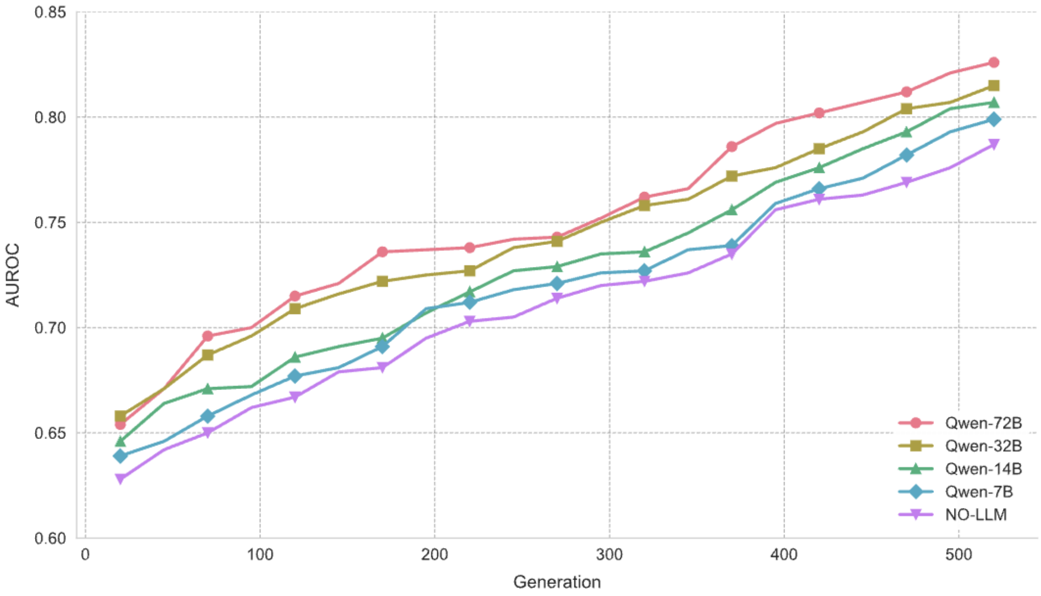


Figure 3. Performance on SymbolicDet with or without LLM.

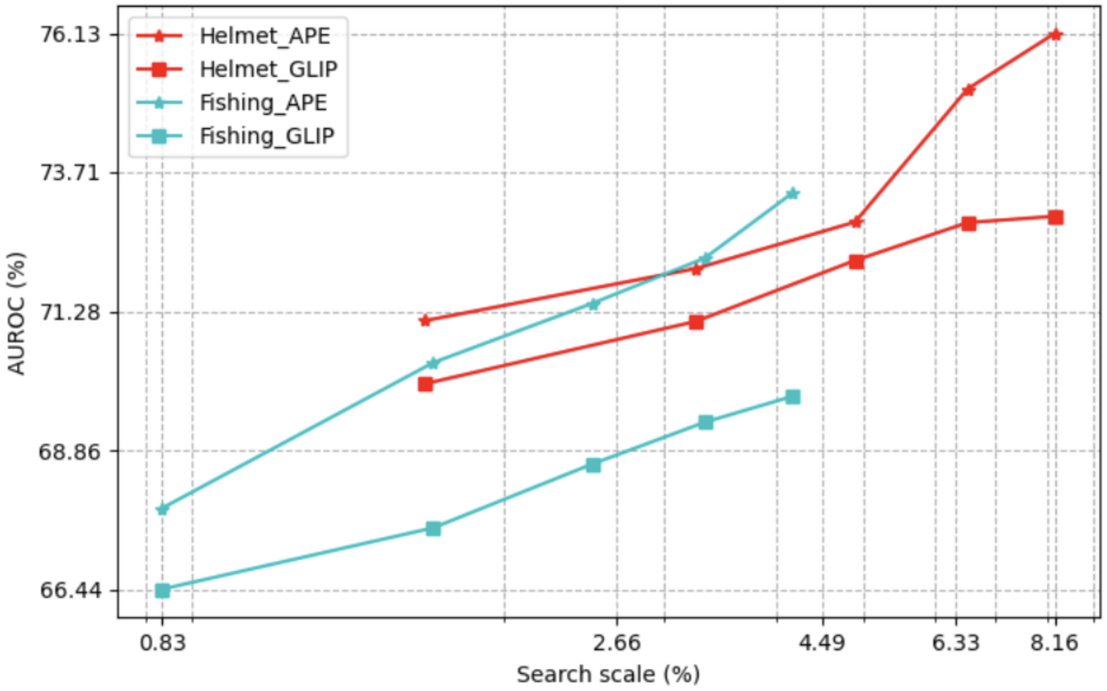


Figure 4. Performance on different search scales.

	wo llm	7B	14B	32B	72B
Run time(s/500it)	80.66	265.5	281.01	268	267
Cost time(s)[1]	69.66	45.92	44.91	30.35	26.07
Memory(MB)	293.65	218.56	218.34	218.68	218.82

Table 5. The computational overhead of symbolic search process.

[1] It refers to the time needed to achieve the same performance.

Conclusion

■ Summary & Outlook

□ Proposed an event-discovery framework for static-image scenes

- □ Extended the capability boundary of off-the-shelf detectors
- □ Provided interpretability for event discovery and a solution route
- □ Leveraged LLM symbolic reasoning within an efficient pipeline

□ Limitations

- □ Image-only: temporal cues required for richer events
- □ Discrete logic-based event definition: cannot describe evolution or degree
- □ Overall performance tightly bound to detector quality

□ Future Work

- □ Close the perception–reasoning feedback loop
- □ Scale to continuous video scenes
- □ Continuous event definition, e.g., via fuzzy logic

Thank you!