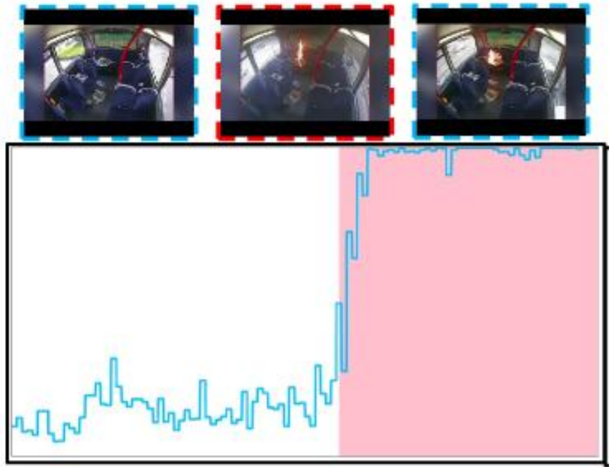
A Unified Reasoning Framework for Holistic Zero-Shot Video Anomaly Analysis

Dongheng Lin[1,2] ,Mengxue Qu[1] ,Kunyang Han[1] ,Jianbo Jiao[2] ,Xiaojie Jin[1] ,Yunchao Wei[1]

[1]Institute of Information Science, Beijing Jiaotong University, [2]The MIx Group, University of Birmingham

Challenges

- Most traditional VAD works stop at frame-wise **temporal** detection—little insight on **where/what/why** an event is abnormal.
- Recently video anomaly localization/understanding works emerged, but remain **data-dependent** and **task-specific**.



When?

Where?

The anomaly exists and its specific name is Arson. The anomaly event involves a newspaper on a seat suddenly catching fire and emitting smoke, which is an unusual and suspicious occurrence that unfolds from start to end without any apparent explanation or precedent. The basis for judging this anomaly is the unexpected and unexplained ignition of the newspaper, which defies the normal laws of physics and daily experiences, suggesting a possible intentional human intervention.

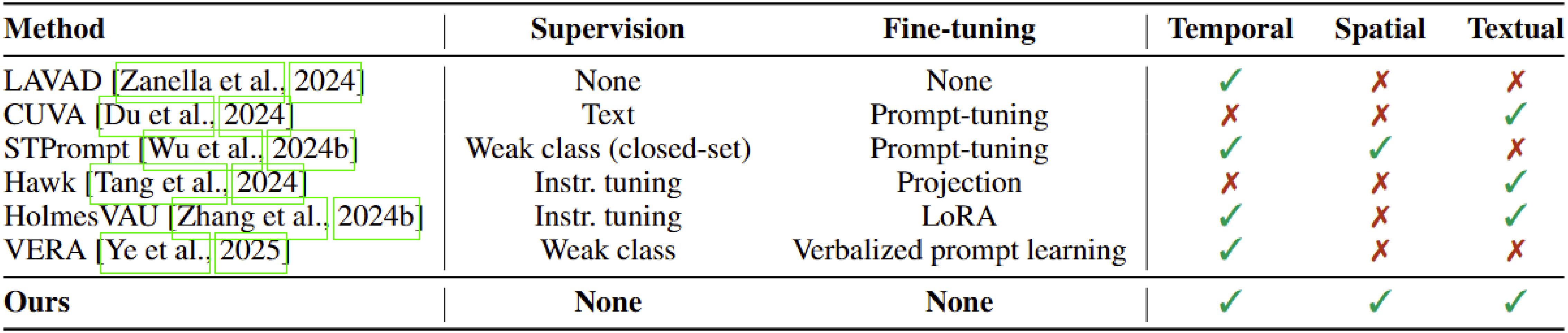
What & Why?

Examples from:

- Wu, Peng, et al. "Weakly supervised video anomaly detection and localization with spatio-temporal prompts." *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.
- Zhang, Huaxin, et al. "Holmes-vau: Towards long-term video anomaly understanding at any granularity." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Challenges

- Most traditional VAD works stop at frame-wise **temporal** detection—little insight on **where/what/why** an event is abnormal.
- Recently video anomaly localization/understanding works emerged, but remain **data-dependent** and **task-specific**.

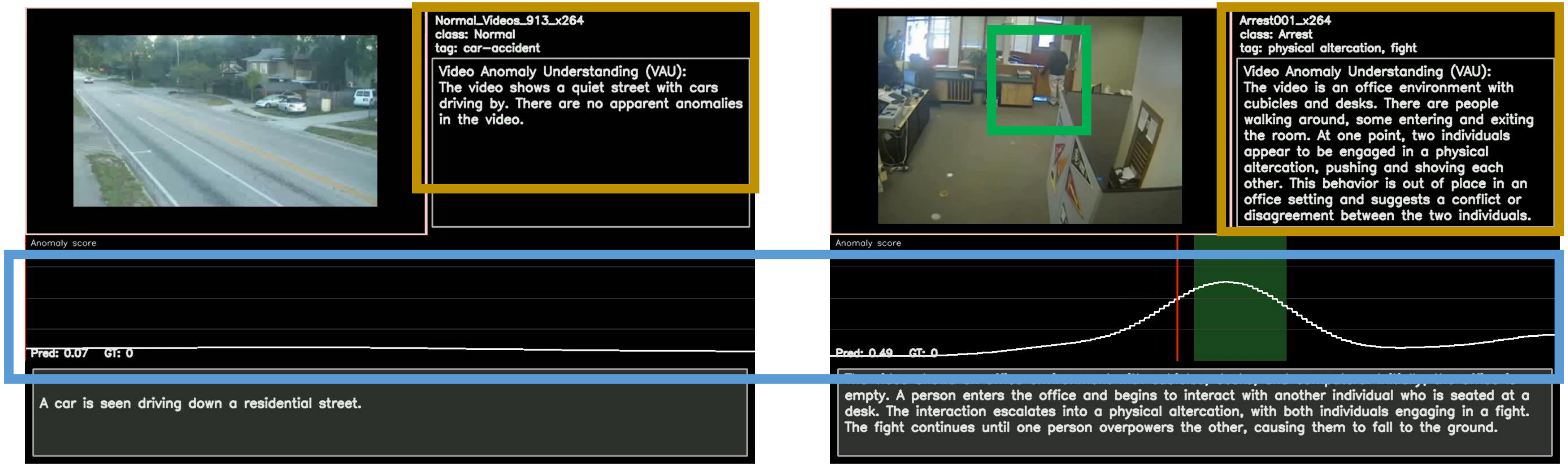


Method	Supervision	Fine-tuning	Temporal	Spatial	Textual
LAVAD [Zanella et al., 2024]	None	None	✓	✗	✗
CUVA [Du et al., 2024]	Text	Prompt-tuning	✗	✗	✓
STPrompt [Wu et al., 2024b]	Weak class (closed-set)	Prompt-tuning	✓	✓	✗
Hawk [Tang et al., 2024]	Instr. tuning	Projection	✗	✗	✓
HolmesVAU [Zhang et al., 2024b]	Instr. tuning	LoRA	✓	✗	✓
VERA [Ye et al., 2025]	Weak class	Verbalized prompt learning	✓	✗	✗
Ours	**None**	**None**	✓	✓	✓

Examples from:

- Wu, Peng, et al. "Weakly supervised video anomaly detection and localization with spatio-temporal prompts." *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024.
- Zhang, Huaxin, et al. "Holmes-vau: Towards long-term video anomaly understanding at any granularity." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Our Contribution



Normal_Videos_913_x264
class: Normal
tag: car–accident

Video Anomaly Understanding (VAU):
The video shows a quiet street with cars driving by. There are no apparent anomalies in the video.

A car is seen driving down a residential street.

Arrest001_x264
class: Arrest
tag: physical altercation, fight

Video Anomaly Understanding (VAU):
The video is an office environment with cubicles and desks. There are people walking around, some entering and exiting the room. At one point, two individuals appear to be engaged in a physical altercation, pushing and shoving each other. This behavior is out of place in an office setting and suggests a conflict or disagreement between the two individuals.

The video shows an office environment with cubicles, desks, and computers. Initially, the office is empty. A person enters the office and begins to interact with another individual who is seated at a desk. The interaction escalates into a physical altercation, with both individuals engaging in a fight. The fight continues until one person overpowers the other, causing them to fall to the ground.

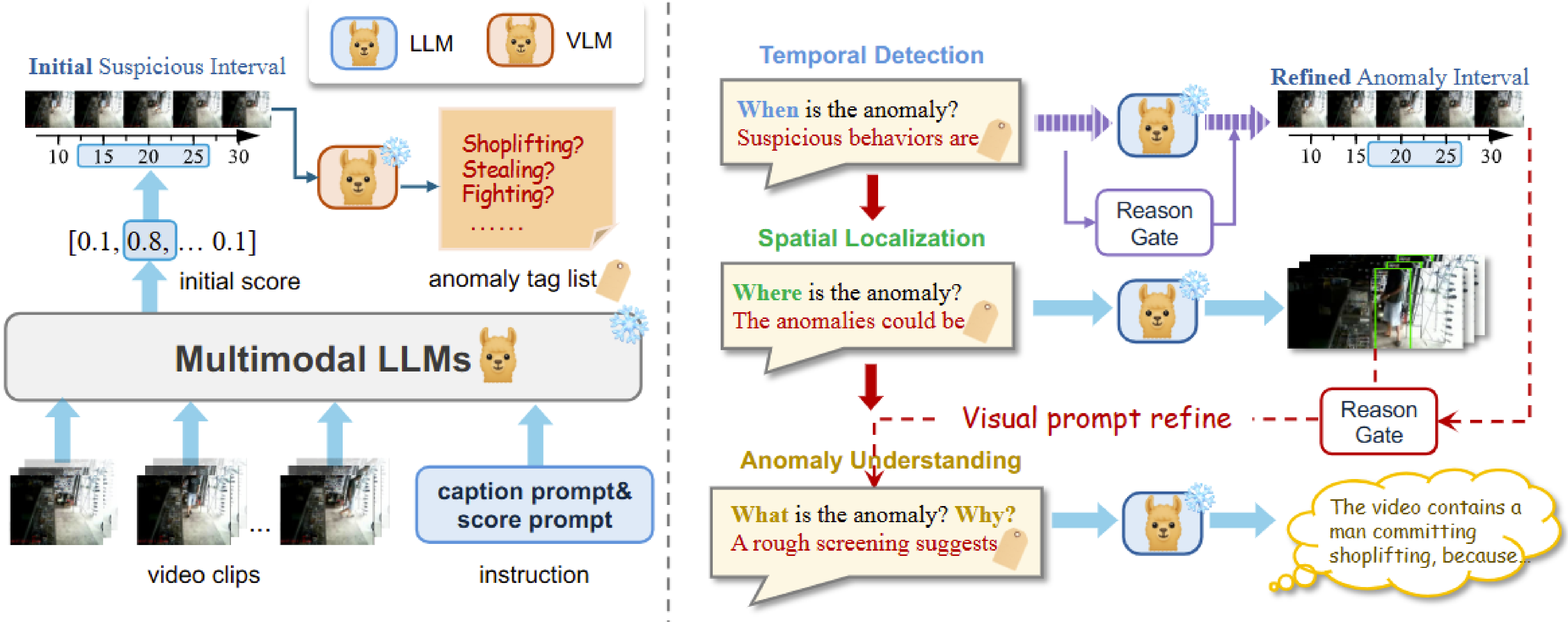
A **training-free** framework that unifies three parts of video anomalies:

1) Temporal detection → 2) Spatial localization → 3) Textual explanation in a single **zero-shot** pipeline.

Pipeline Overview

Key ideas:

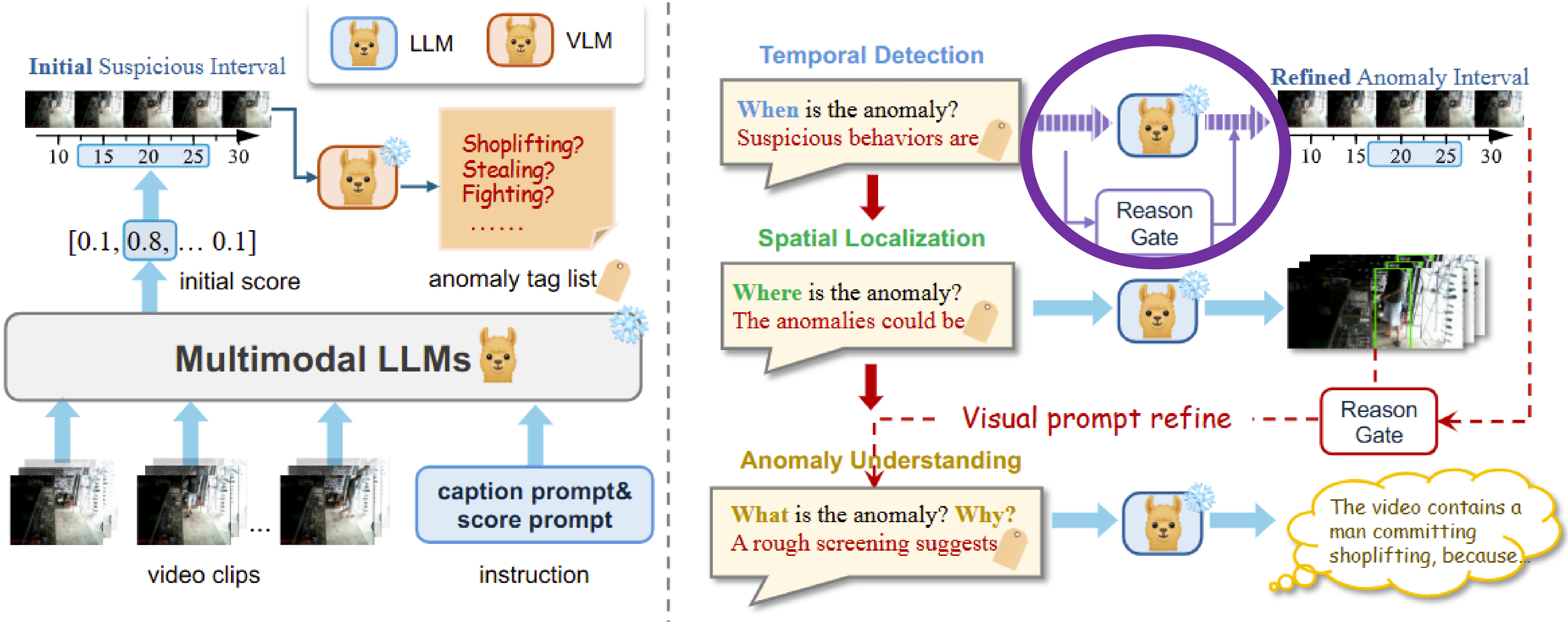
1. **Reasoning Efficiently for hard VAD tasks.**
2. **Chaining Separate tasks to reuse per-task priors.**



Pipeline Overview

Key ideas:

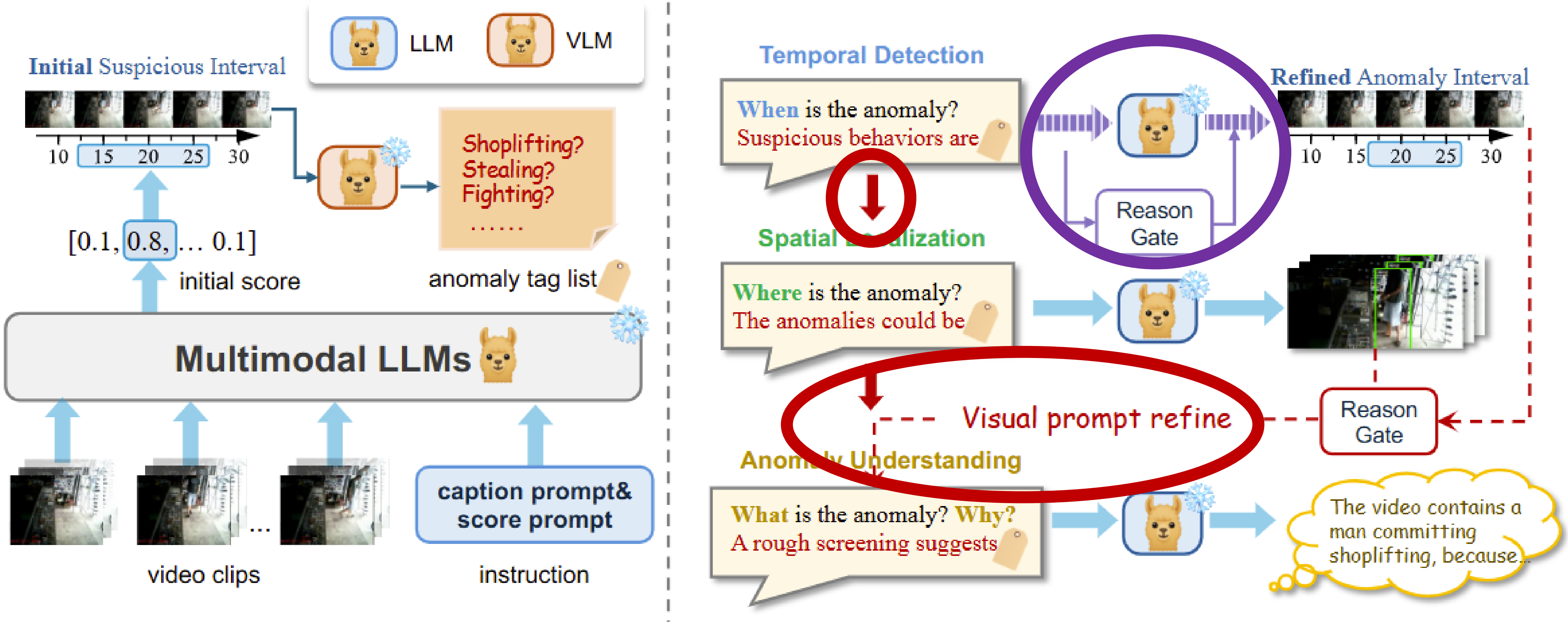
1. **Reasoning Efficiently for hard VAD tasks.**
2. **Chaining Separate tasks to reuse per-task priors.**



Pipeline Overview

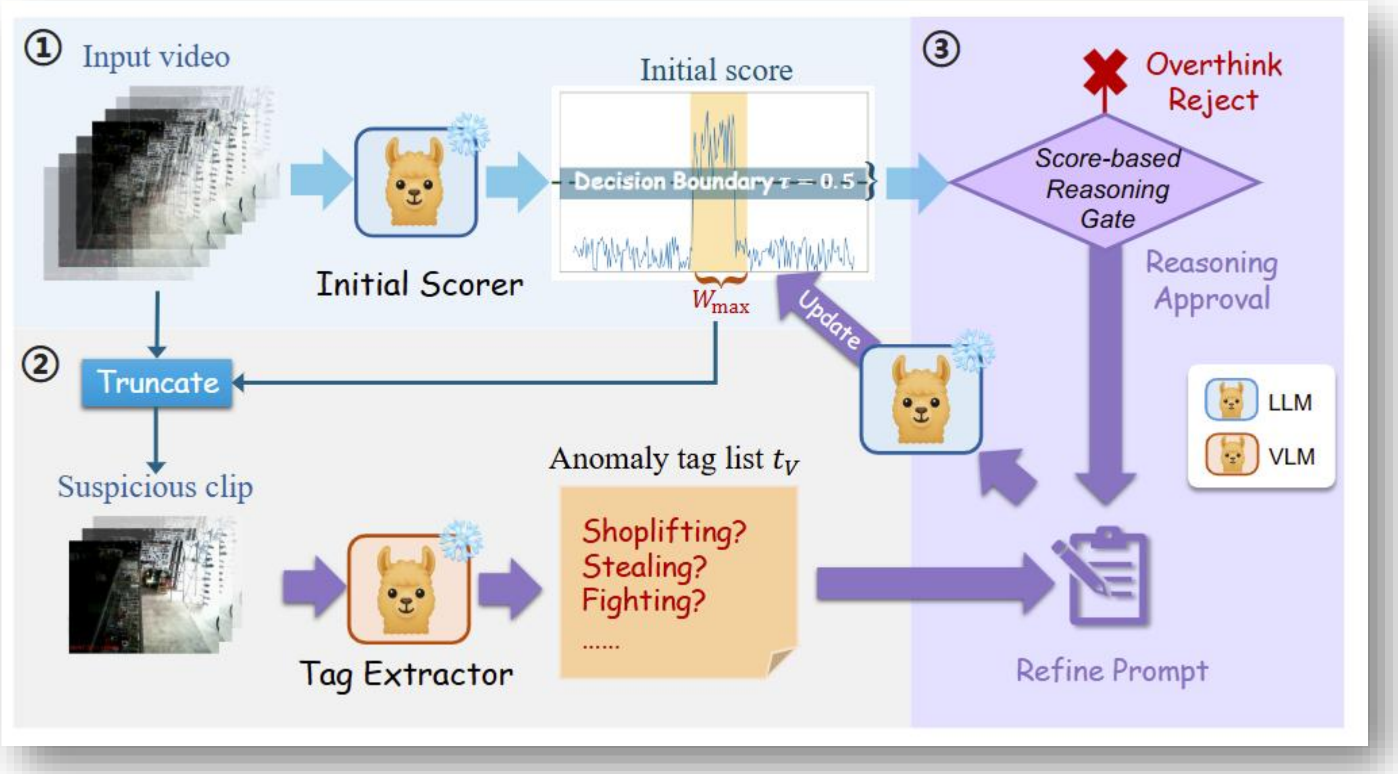
Key ideas:

1. **Reasoning Efficiently for hard VAD tasks.**
2. **Chaining Separate tasks to reuse per-task priors.**



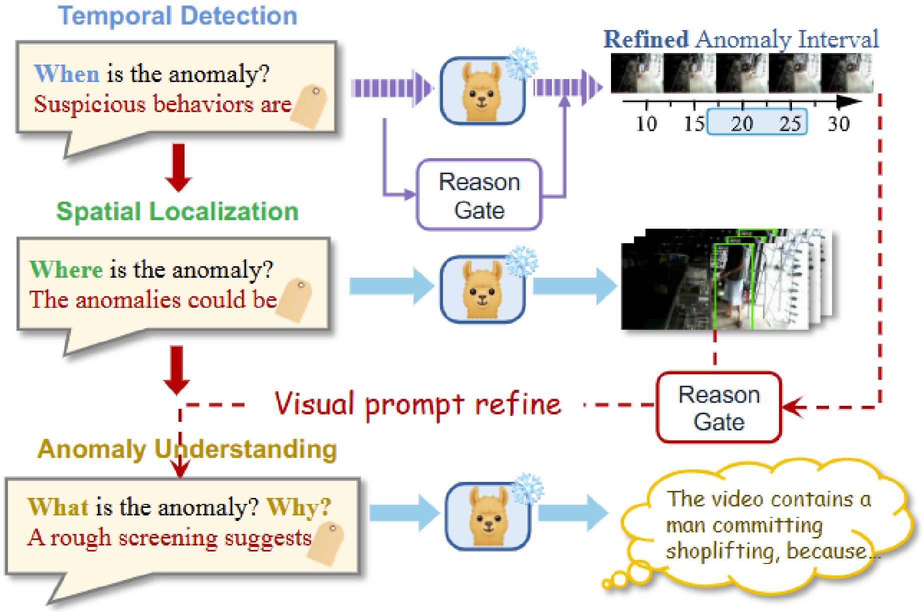
Intra-Task Reasoning (IntraTR)

- **IntraTR** refines temporal anomaly scores by extracting and leveraging anomaly priors from the most suspicious video segment.



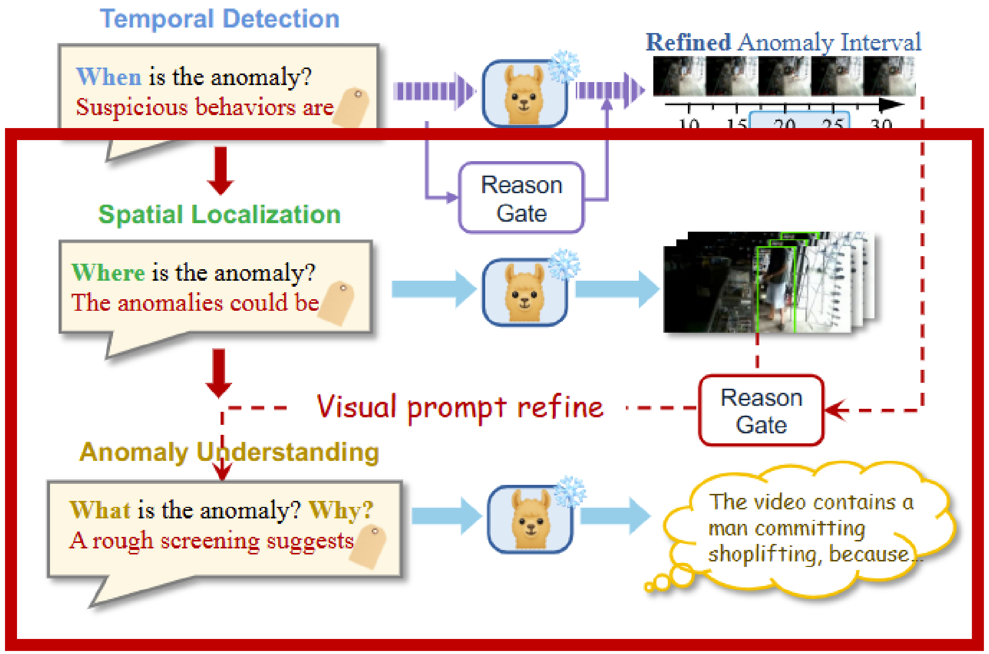
Inter-Task Chaining (InterTC)

- **InterTC** links temporal detection with spatial localization and textual explanation, enabling holistic anomaly analysis with clearer thinking/reasoning steps.



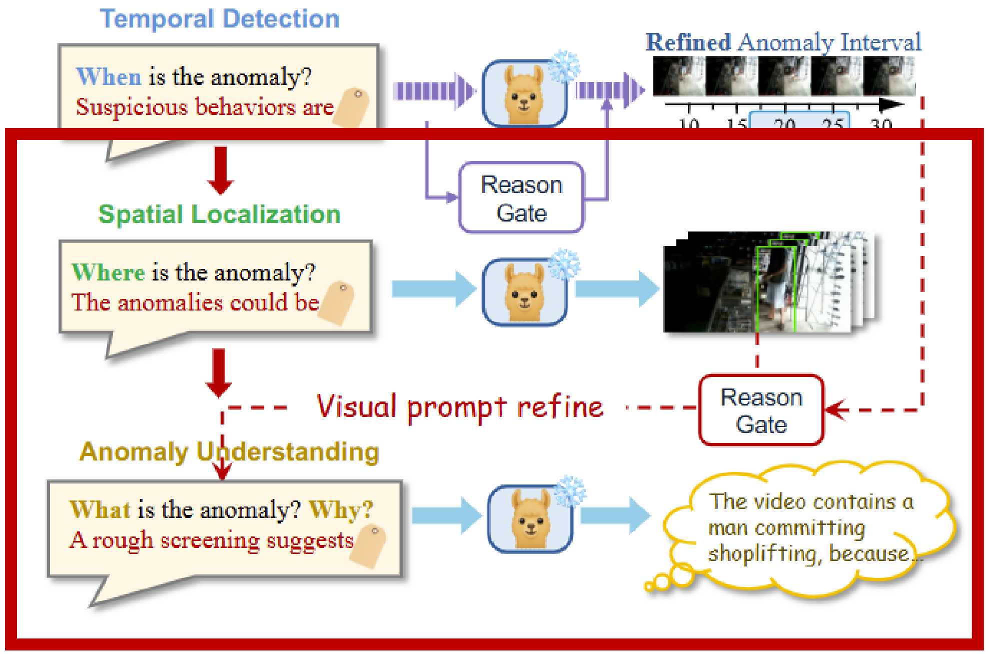
Inter-Task Chaining (InterTC)

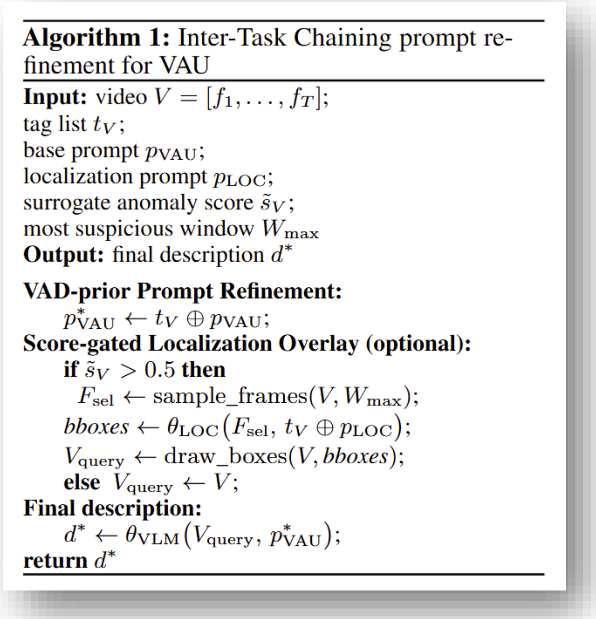
- **InterTC** links temporal detection with spatial localization and textual explanation, enabling holistic anomaly analysis with clearer thinking/reasoning steps.



Inter-Task Chaining (InterTC)

- **InterTC** links temporal detection with spatial localization and textual explanation, enabling holistic anomaly analysis with clearer thinking/reasoning steps.



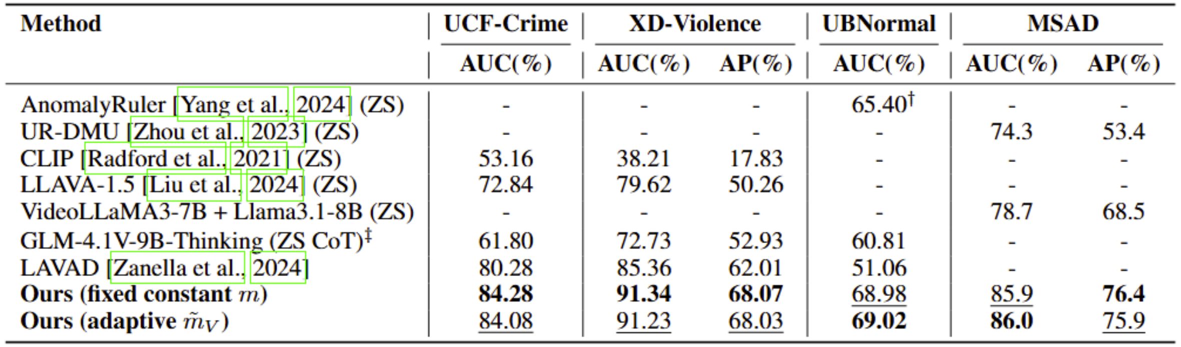


```
Algorithm 1: Inter-Task Chaining prompt refinement for VAU
Input: video $V = [f_1, \ldots, f_T]$;
tag list $t_V$;
base prompt $p_{\mathrm{VAU}}$;
localization prompt $p_{\mathrm{LOC}}$;
surrogate anomaly score $\tilde{s}_V$;
most suspicious window $W_{\max}$
Output: final description $d^*$

VAD-prior Prompt Refinement:
    $p^*_{\mathrm{VAU}} \leftarrow t_V \oplus p_{\mathrm{VAU}}$;
Score-gated Localization Overlay (optional):
    if $\tilde{s}_V > 0.5$ then
      $F_{\mathrm{sel}} \leftarrow$ sample_frames$(V, W_{\max})$;
    $bboxes \leftarrow \theta_{\mathrm{LOC}}\big(F_{\mathrm{sel}},\ t_V \oplus p_{\mathrm{LOC}}\big)$;
    $V_{\mathrm{query}} \leftarrow$ draw_boxes$(V, bboxes)$;
    else $V_{\mathrm{query}} \leftarrow V$;
Final description:
    $d^* \leftarrow \theta_{\mathrm{VLM}}\big(V_{\mathrm{query}},\ p^*_{\mathrm{VAU}}\big)$;
return $d^*$
```

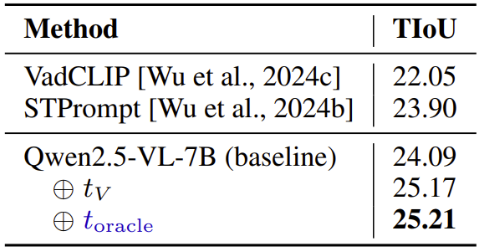
Results at a Glance

➢ Temporal Video Anomaly Detection (VAD)



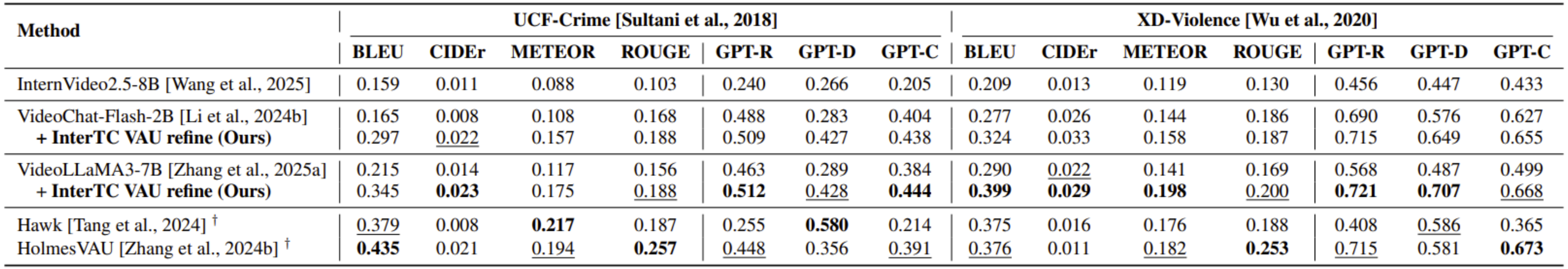
Method	UCF-Crime	XD-Violence		UBNormal	MSAD	
	AUC(%)	AUC(%)	AP(%)	AUC(%)	AUC(%)	AP(%)
AnomalyRuler [Yang et al., 2024] (ZS)	-	-	-	65.40†	-	-
UR-DMU [Zhou et al., 2023] (ZS)	-	-	-	-	74.3	53.4
CLIP [Radford et al., 2021] (ZS)	53.16	38.21	17.83	-	-	-
LLAVA-1.5 [Liu et al., 2024] (ZS)	72.84	79.62	50.26	-	-	-
VideoLLaMA3-7B + Llama3.1-8B (ZS)	-	-	-	-	78.7	68.5
GLM-4.1V-9B-Thinking (ZS CoT)‡	61.80	72.73	52.93	60.81	-	-
LAVAD [Zanella et al., 2024]	80.28	85.36	62.01	51.06	-	-
Ours (fixed constant m)	**84.28**	**91.34**	**68.07**	68.98	85.9	**76.4**
Ours (adaptive $\tilde{m}_V$)	84.08	91.23	68.03	**69.02**	**86.0**	75.9

➢ Spatial Video Anomaly Localization (VAL)



Method	TIoU
VadCLIP [Wu et al., 2024c]	22.05
STPrompt [Wu et al., 2024b]	23.90
Qwen2.5-VL-7B (baseline)	24.09
$\oplus\ t_V$	25.17
$\oplus\ t_{\text{oracle}}$	**25.21**

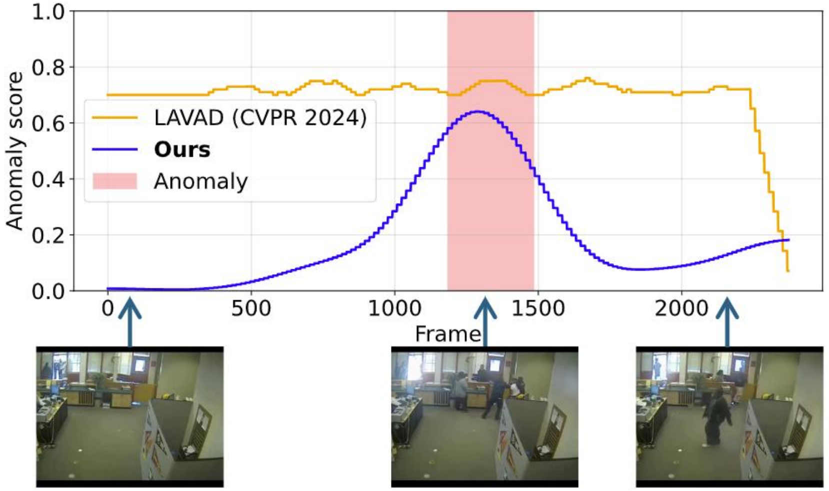
➢ Textual Video Anomaly Understanding (VAU)



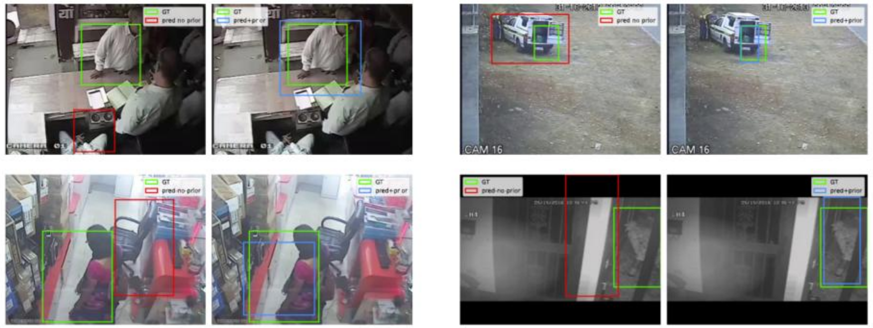
Method	UCF-Crime [Sultani et al., 2018]							XD-Violence [Wu et al., 2020]						
	BLEU	CIDEr	METEOR	ROUGE	GPT-R	GPT-D	GPT-C	BLEU	CIDEr	METEOR	ROUGE	GPT-R	GPT-D	GPT-C
InternVideo2.5-8B [Wang et al., 2025]	0.159	0.011	0.088	0.103	0.240	0.266	0.205	0.209	0.013	0.119	0.130	0.456	0.447	0.433
VideoChat-Flash-2B [Li et al., 2024b]	0.165	0.008	0.108	0.168	0.488	0.283	0.404	0.277	0.026	0.144	0.186	0.690	0.576	0.627
+ InterTC VAU refine (Ours)	0.297	0.022	0.157	0.188	0.509	0.427	0.438	0.324	0.033	0.158	0.187	0.715	0.649	0.655
VideoLLaMA3-7B [Zhang et al., 2025a]	0.215	0.014	0.117	0.156	0.463	0.289	0.384	0.290	0.022	0.141	0.169	0.568	0.487	0.499
+ InterTC VAU refine (Ours)	0.345	**0.023**	0.175	0.188	**0.512**	0.428	**0.444**	**0.399**	**0.029**	**0.198**	0.200	**0.721**	**0.707**	0.668
Hawk [Tang et al., 2024] †	0.379	0.008	**0.217**	0.187	0.255	**0.580**	0.214	0.375	0.016	0.176	0.188	0.408	0.586	0.365
HolmesVAU [Zhang et al., 2024b] †	**0.435**	0.021	0.194	**0.257**	0.448	0.356	0.391	0.376	0.011	0.182	**0.253**	0.715	0.581	**0.673**

Results at a Glance

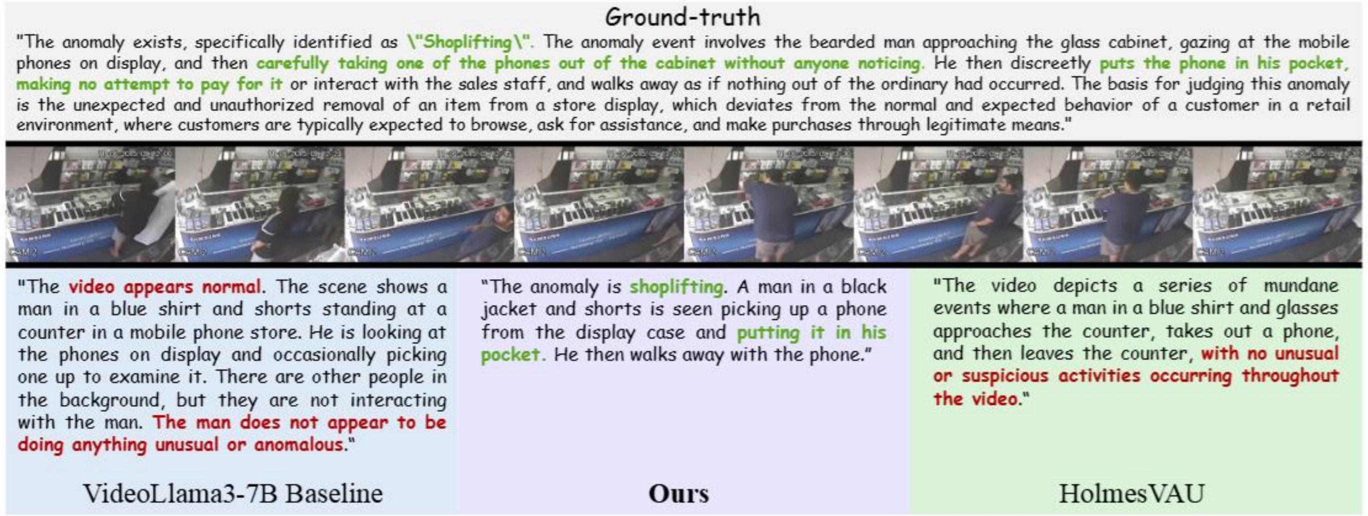
➢ Temporal Video Anomaly Detection (VAD)



➢ Spatial Video Anomaly Localization (VAL)



➢ Textual Video Anomaly Understanding (VAU)



Ground-truth

"The anomaly exists, specifically identified as **\"Shoplifting\"**. The anomaly event involves the bearded man approaching the glass cabinet, gazing at the mobile phones on display, and then **carefully taking one of the phones out of the cabinet without anyone noticing**. He then discreetly **puts the phone in his pocket, making no attempt to pay for it** or interact with the sales staff, and walks away as if nothing out of the ordinary had occurred. The basis for judging this anomaly is the unexpected and unauthorized removal of an item from a store display, which deviates from the normal and expected behavior of a customer in a retail environment, where customers are typically expected to browse, ask for assistance, and make purchases through legitimate means."

"The **video appears normal**. The scene shows a man in a blue shirt and shorts standing at a counter in a mobile phone store. He is looking at the phones on display and occasionally picking one up to examine it. There are other people in the background, but they are not interacting with the man. **The man does not appear to be doing anything unusual or anomalous.**"

VideoLlama3-7B Baseline

"The anomaly is **shoplifting**. A man in a black jacket and shorts is seen picking up a phone from the display case and **putting it in his pocket**. He then walks away with the phone."

Ours

"The video depicts a series of mundane events where a man in a blue shirt and glasses approaches the counter, takes out a phone, and then leaves the counter, **with no unusual or suspicious activities occurring throughout the video.**"

HolmesVAU

Conclusions

- In general, our work unified VAD→VAL→VAU in a single zero-shot chained pipeline by:
    - Reasoning when needed (**IntraTR**)
    - Reusing priors across tasks (**InterTC**)

- Limitations & Future work
    - Compute/latency from Heavy VLMs
    - Probably try to harness zero-shot capability of V-JEPA-style “world model”?

Thanks for listening!

NEURAL INFORMATION PROCESSING SYSTEMS

WEI Lab@ BJTU

MIX University of Birmingham

UNIVERSITY OF BIRMINGHAM