UAL-Bench: The First Comprehensive Unusual Activity Localization Benchmark

Hasnat Md Abdullah[1] Tian Liu[1] Kangda Wei[1] Shu Kong[2] Ruihong Huang[1]

[1]Texas A&M University [2]University of Macau

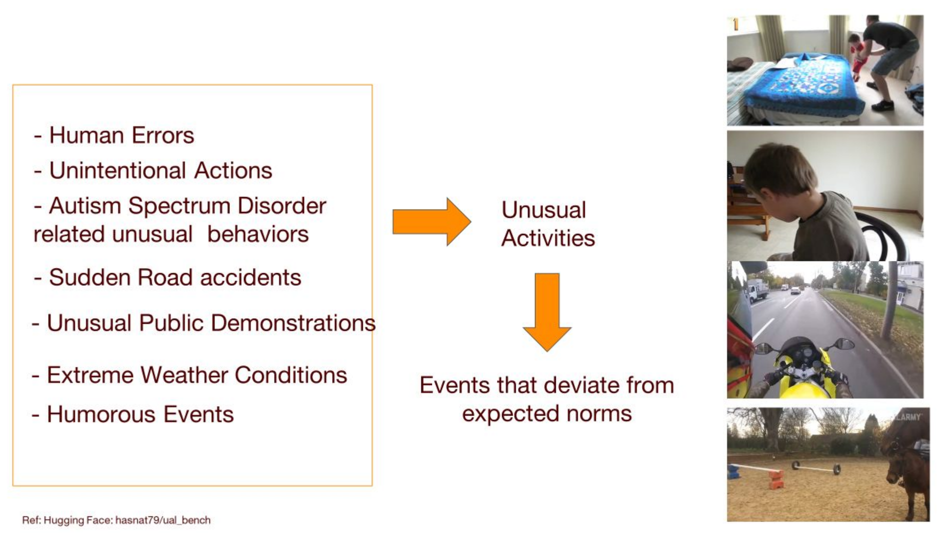
WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Introduction

Incident

"In 2018, a high-speed train collision in Turkey claimed 9 lives and injured over 80, all due to a single human error—an operator's split-second mistake in assigning the wrong track, as revealed by haunting surveillance footage."

Defining Unusual Activities



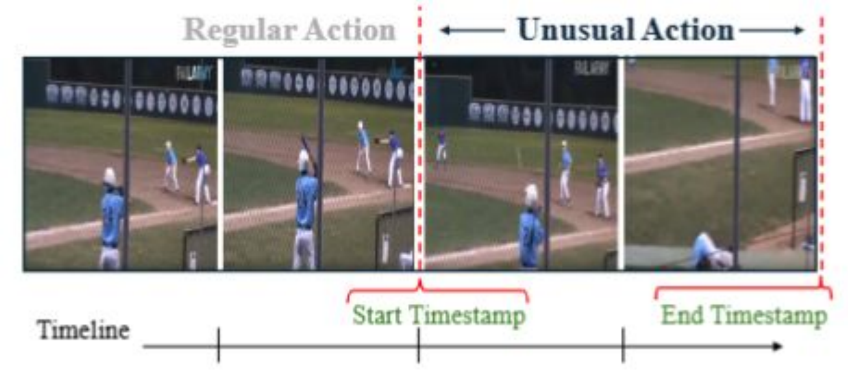
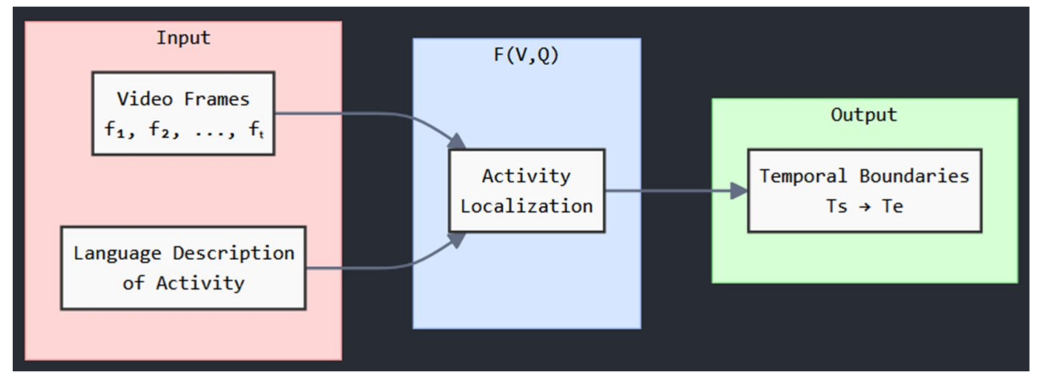


Figure 1. Example of an unusual activity in a baseball game scene. In the 3rd and 4th frames, the ball unexpectedly strikes a batter's head, causing him to fall on the ground. This event is classified as an unusual action. This video is from UAG-OOPS dataset and the file name is *Bats & Balls Fail Compilation _ By FailArmy 20160.mp4*.

Defining Unusual Activity Localization Task [1]



Motivation

Challenges

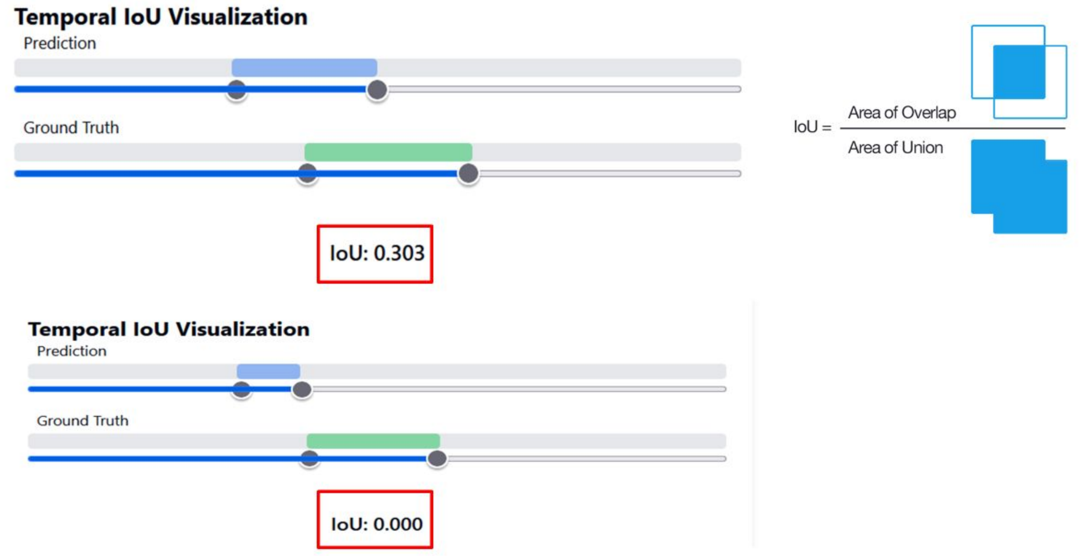
1. Existing Vid-LLMs' pretraining data do not represent Unusual Activities Sufficiently.
2. Common Metric to measure Temporal Activity Localization: Intersection over Union (IoU) fails to measure performance when the Prediction and Ground Truth spans are close but do not overlap
3. No zero-shot solution has been proposed to address the task of unusual activity localization using Large Vision Language Models (VLM) and LLMs

Vid-LLM Training Datasets

- ActivityNet
- CharadesSTA
- HowTo100M
- MSRVTT
- MSVD
- DiDeMo
- WebVid-2M

Activities

- Sports and Physical Activities
- Household and Daily Tasks
- Entertainment and Leisure
- Work and DIY Activities
- Interactions with Objects

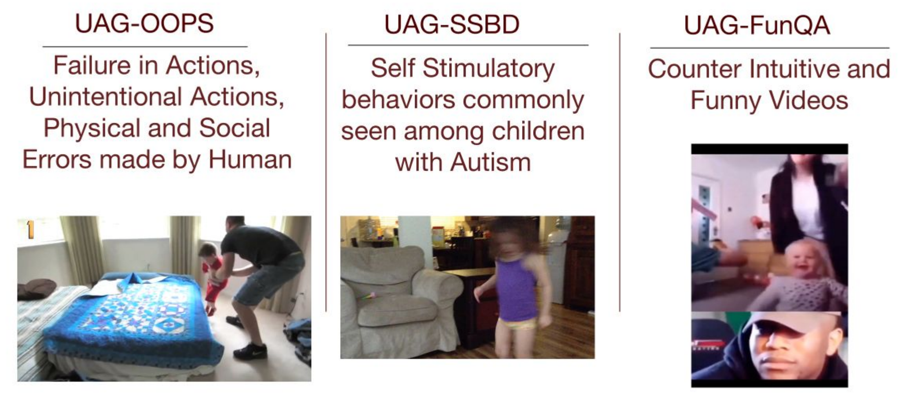


Contributions

1. We propose UAL-Bench, the first comprehensive benchmark for unusual activity localization, which includes three datasets for unusual activity localization: UAG-OOPS [2], UAG-SSBD [3], UAG-FunQA [4].

Table 1. Statistics of the proposed datasets compared to standard temporal localization dataset Charades-STA [19]. Despite being shorter in average duration, OOPS-UAG-Instruct contains more detailed descriptions than Charades-STA.

Dataset	# of Videos	Avg Duration (seconds)	Avg Description length (words)
UAG-OOPS	1,589	8.34	92
UAG-SSBD	75	90	7
UAG-FunQA	172	7.26	5
OOPS-UAG-Instruct	3,778	9.83	93.52
Charades-STA [19]	3,720	30.59	33



2. We introduce a new metric, Temporal Distance, $TD \leq P$ to address the limitations in existing metric, providing more reliable evaluation in certain scenarios

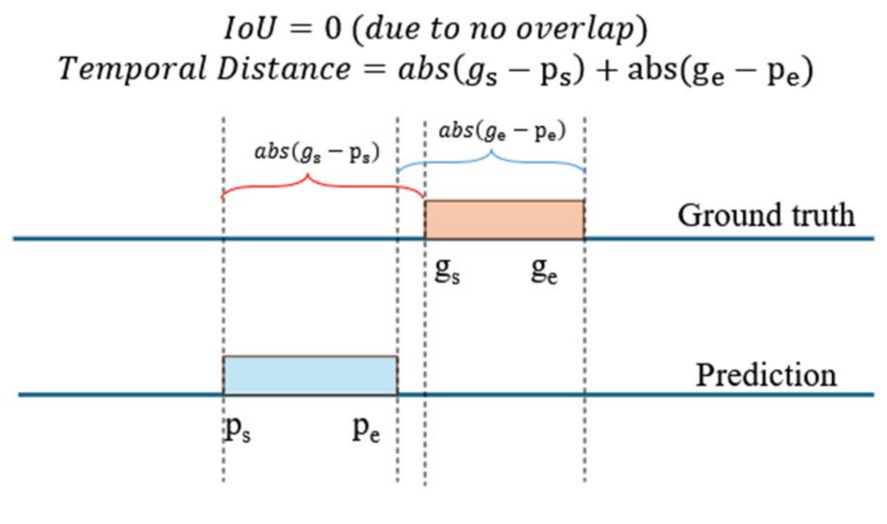
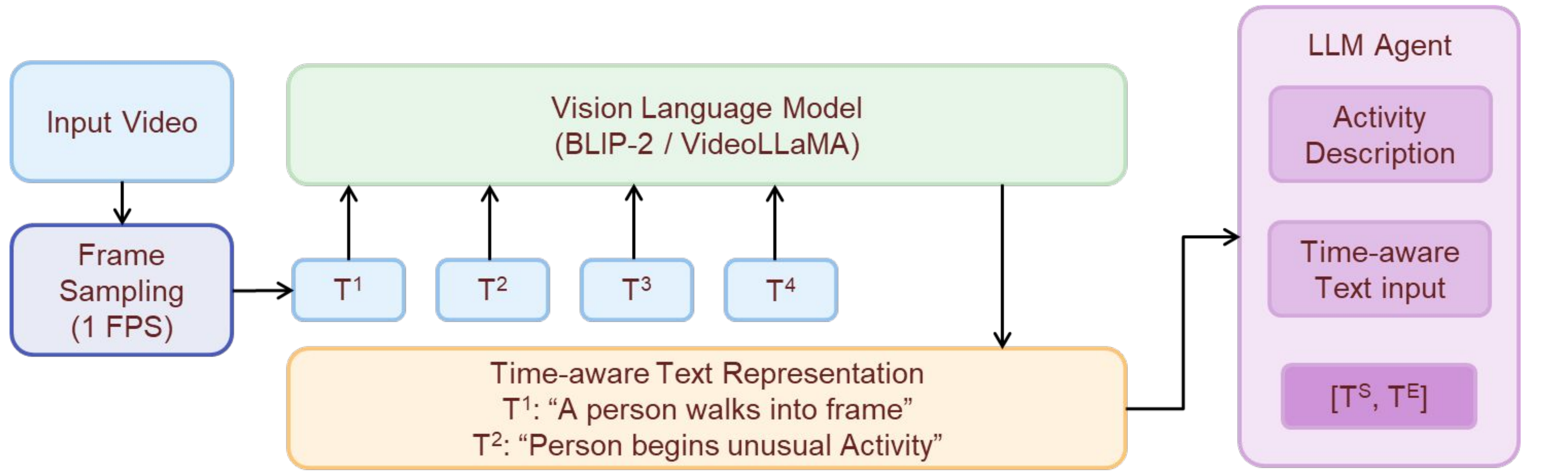


Figure 2. An illustration of our proposed Temporal Distance.

3. To address the challenge of no zero-shot solution available for unusual activity localization, we introduce a novel integration of Language and Vision Models (VLM-LLM)



Results

Our VLM-LLM approach outperforms existing Vid-LLM in short-span unusual events (UAG-FunQA)

Table 2. Overall performance comparison of the Video-LLM, VLM-LLM and Fine-tuned VLM approaches on three unusual activity localization benchmarks: UAG-OOPS, UAG-SSBD and UAG-FunQA. For the $R@1, IoU \geq m$ and $R@1, TD \leq p$ metrics, higher scores indicate better performance, while for the mTD metric, the lower scores are better.

	Methods	$R@1, IoU \geq m$ m=0.3	m=0.5	m=0.7	$mIoU(0-1)$	$R@1, TD \leq p(sec)$ p=0	p=1	p=3	p=5	$mTD(sec)$
UAG-OOPS	VideoChat2 [34]	16.49	5.98	1.95	0.12	0.00	1.01	8.37	23.10	9.31
	Video-ChatGPT [39]	25.49	10.70	3.15	0.18	0.00	1.32	8.37	25.61	11.50
	VideoLLaMA [62]	**40.72**	**20.77**	**6.23**	**0.27**	**0.06**	**2.01**	14.85	33.79	11.22
	BLIP2-LlaMA3 [32,41](ours)	19.07	7.17	2.45	0.15	0.00	1.38	**27.00**	53.74	5.85
	VideoLlaMA-LlaMA3 [41,62] (ours)	19.38	7.93	2.08	0.15	0.00	1.89	26.12	**55.13**	**5.72**
	VideoLlaMA 7B (Fine-tuned)	2.96	0.50	0.19	0.04	0.00	1.26	10.07	22.84	14.09
	Random	12.21	4.47	1.45	0.10	0.00	0.31	2.77	5.29	24.10
UAG-SSBD	VideoChat2	2.88	0.96	0.00	0.02	0.00	0.00	1.92	2.88	139.63
	Video-ChatGPT	4.81	2.88	0.00	0.03	0.00	0.00	0.96	2.88	93.99
	VideoLlaMA	**15.38**	**8.65**	1.92	**0.11**	0.00	0.00	**3.85**	**6.73**	96.55
	BLIP2-LlaMA3 (ours)	1.92	1.92	1.92	0.03	0.00	0.00	0.96	1.92	**68.05**
	VideoLlaMA-LlaMA3 (ours)	2.88	0.96	0.00	0.03	0.00	0.00	0.96	4.81	70.38
	VideoLlaMA 7B (Fine-tuned)	0.00	0.00	0.00	0.01	0.00	0.00	0.00	1.92	105.27
	Random	10.58	5.77	**3.85**	0.10	0.00	0.00	1.92	1.92	87.73
UAG-FunQA	VideoChat2	12.79	4.65	3.49	0.08	0.00	2.33	23.84	44.77	7.48
	Video-ChatGPT	1.16	0.58	0.00	0.01	0.00	0.00	22.67	44.77	53.42
	VideoLlaMA	2.91	0.58	0.00	0.02	0.00	0.00	4.07	8.72	31.64
	BLIP2-LlaMA3 (ours)	**18.60**	**9.30**	**5.23**	**0.12**	0.00	**9.30**	39.53	60.47	5.43
	VideoLlaMA-LlaMA3 (ours)	12.21	4.65	2.33	0.09	0.00	5.23	**44.19**	**65.70**	**4.93**
	VideoLlaMA 7B (Fine-tuned)	6.40	2.33	0.0	0.01	0.00	5.81	29.65	47.67	8.19
	Random	5.81	1.74	0.58	0.05	0.00	0.00	1.74	4.65	27.27

Nine observations to guide future research.

1. VLM-LLM excels in localizing short-span unusual activities, outperforming existing Vid-LLMs in short video datasets.
2. VLM-LLM provides highly accurate and coherent explanations, enhancing the interpretability of model predictions.
3. VLM-LLM outperforms most vid-LLMs in standard temporal activity localization benchmarks like Charades-STA.
4. Our benchmark datasets present challenges comparable to the Charades-STA dataset for unusual activity localization.
5. The $IoU \geq m$ metric becomes unreliable for evaluating short-span videos, requiring specialized metrics.
6. There are trade-offs between model complexity and performance, especially in terms of inference time for VLM-LLM. Yet it yield 2X accuracy boost compared to Vid-LLMs.
7. Long-duration diagnosis videos, like those in UAG-SSBD, require tailored models for accurate interpretation.
8. Instruction-tuning suffers due to the lack of time-awareness in the video encoder, impacting performance.
9. Explicit content in annotations can trigger model refusals, requiring careful wording during annotation.

References

[1] J. Gao, et al (2017) "Tall: Temporal activity localization via language query." In: ICCV
[2] Dave Epstein et. al. (2020) "Oops! predicting unintentional action in video" In: CVPR
[3] Shyam Rajagopalan et. al. (2013) "Self-stimulatory behaviours in the wild for autism diagnosis." In: Proceedings of the IEEE International Conference on Computer Vision Workshops
[4] Binzhu Xie et al.(2025) "Funqa: Towards surprising video comprehension".In: European Conference on Computer Vision