

Aligning Effective Tokens with Video Anomaly in Large Language Models

Supplementary Material

1. Details of Datasets

1.1. Fine-tuning Data

By utilizing the training dataset that comprises the UCF-Crimes dataset, we attain a significant data volume of 775241 instruction pairs (see Figure 1) for model fine-tuning. This comprehensive dataset enables a more rigorous assessment of the model’s efficacy in identifying and comprehending different types of anomalies.

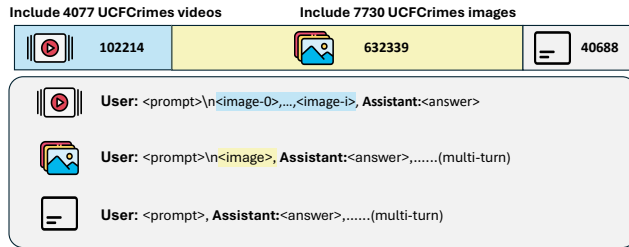


Figure 1. The distribution of the dataset. There is a huge amount of video, image and text data available for training, including 7730 images and 4077 videos from UCF-Crimes Dataset [11].

We re-construct the UCF-Crimes [11] training dataset as instruct-following format [8] for participating in model fine-tuning. Utilizing the UCF-Crimes annotation dataset UCA [16], which offers detailed annotations, we firstly segment the videos into distinct clips and associate each clip with corresponding descriptive information. Based on these descriptions, we employ ChatGPT4 [1] to generate three relevant questions, with the final question being “Are there any anomalies?”. Finally, we re-purpose the questions generated by ChatGPT4 to produce corresponding answers, which we then modify to conform to the instruct-following training data format shown in Figure 2. Our QA pairs are initially generated using GPT-4. Then each pair is carefully reviewed and validated by human annotators. These two steps help us maintain scalability and accuracy in datasets. We will continue to be committed to the exploration and utilization of efficient data [2, 12, 13], and pay more attention to details in data annotation in scenarios such as personnel security and traffic planning [4, 9, 15, 19].

1.2. Evaluation Data

We evaluate our model’s performance using two benchmarks based on UCF-Crimes and XD-Violence Question-Answer pair datasets, respectively. Initially, we extract vision information from videos and transform test videos into the `pk1` format, which facilitates the extraction visual information from the videos.

We firstly test our model on UCF-Crimes benchmarks (see Figure 3) as our in-domain evaluation, which is an existing benchmark for video anomaly understanding MLLMs. Following the structure illustrated in Figure 3, we design a new evaluation benchmark based on XD-Violence dataset. We set up Question-Answer pair file for each test video, creating four multiple-choice options that cover the anomaly type and its corresponding time interval. Among these options, only one accurately reflects the video content. We proceed to load the Question-Answer pair file, arranging the questions in a manner akin to the “User” section (Figure 4). The model then processes these questions, along with the `pk1` file containing video data, to generate the most suitable response. Upon obtaining the answer, we compute the precision by comparing the ground truth with the model’s inference results. Through our in-domain and cross-domain evaluations, different models will be more accurately tested for their reliability and robustness, thus having more opportunities to promote open-world learning [10, 18, 20, 21] and potentially promote development in more fields [5–7].

1.3. Comparison with Other Dataset

Our research prioritizes the development of comprehensive, high-quality instruction-following data. While CUVA [3] offers valuable contributions, its dataset lacks the breadth and depth of our proposed approach (Tab.1). Similarly, although HAWK [14] presents relevant work in this domain, the absence of open-source evaluation code limits direct methodological comparison. Nevertheless, our multi-dimensional comparative analysis demonstrates the superior characteristics of our dataset across several key metrics (Tab.2).

Data	Total Video Length	Questions-Answer Pairs	Num. of Clips
CUVA [3]	88.16 h	6720	2240
<i>Ours</i>	128.00 h	16308	4077

Table 1. Details of CUVA and our proposed data.

2. More Experiment Details

To assess the effectiveness of our model in spatial and temporal video anomaly understanding, we design specific prompts tailored to separately evaluate these two aspects. The results indicate that our model demonstrates strong performance in both spatial and temporal anomaly understanding, excelling in each area independently.

2.1. Spatial Understanding

There are some interaction cases with our model as shown in Figure 5. We incorporate the localization prompts “top, bot-

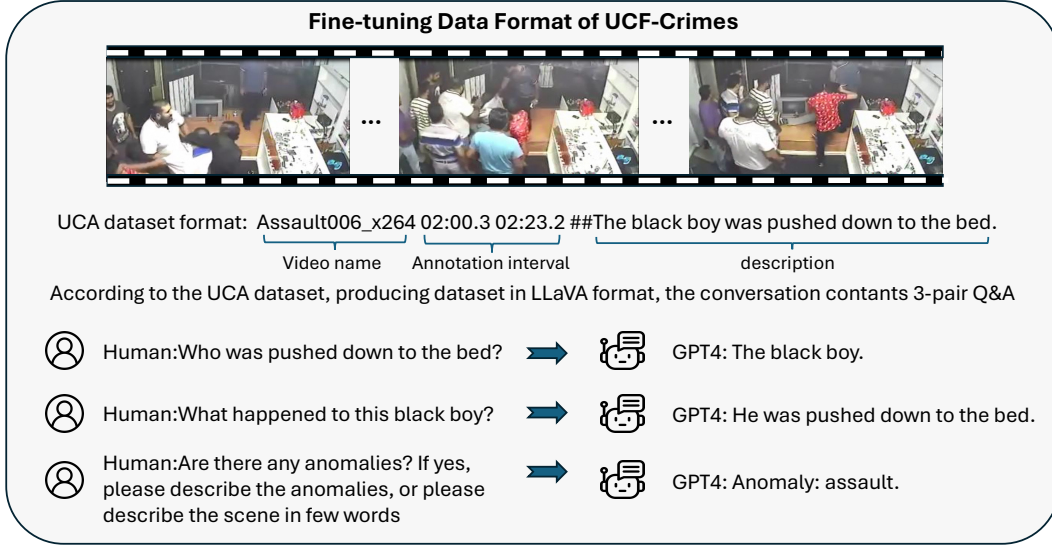


Figure 2. The fine-tuning process for the UCF-Crimes [11] dataset involves generating 3 Question-Answer pairs based on the UCA dataset [16] and converting them to the LLaVA format. Videos in the UCF-Crimes dataset and their corresponding cut-frame images are then added to the complete fine-tuning dataset.

Feature	Hawk [14]	HolmesVAU [17]	Ours
Structure	Lack conversations	Lack conversations	Multiple QA pairs
Clarity of Anomaly Judgment	Does not directly address anomaly judgment	Clear anomaly descriptions but overly focused, lacks contrast with normal scenes	Clear (e.g., "Normal" or specific anomaly descriptions)
Description	Overly general, lacks specific scene details	Excessive and repetitive details	Moderate, specific, and concise
Diversity	Focuses on the importance of anomaly handling, lacks diversity	Focuses on anomaly events, lacks descriptions of normal scenes	Covers various question types, rich in content

Table 2. A Comparative analysis of Structural elements, Judgment clarity, Descriptive quality, and Content diversity between other methods and our proposed data.

tom, centre, right, left" into the questions and inquire about both the anomaly and its localization. The responses reveal that our model exhibits exceptional precision in pinpointing the specific localization where anomalies occur, demonstrating its remarkable accuracy in anomaly understanding. For example, the anomalies in video `Explosion_016_x264` show an explosion. Our model can correctly describe the localization of smoke and give a reasonable description.

2.2. Temporal Understanding

Illustrated in Figure 6 are several interaction cases of temporal anomaly understanding. Using video `Shooting021_x264` as an example, the video contains just one type of anomaly that is clearly visible in the scene. Consequently, our model effectively delivers the accurate anomaly description along with the corresponding time slot. For instances where anomalies are less obvious (like `Shoplifting016_x264` in Figure 6) or when multiple

anomalies exist within a video (like `Arrest030_x264` in Figure 6), we incorporate the anomaly type into the prompt. This approach assists the model in accurately determining the time intervals for each anomaly.


2.3. Details of Temporal Classifier


We use video-level captions (with temporal information) from UCF-Crime training set to label whether each frame contains anomalies. These labels are used to train a three-layer MLP binary classifier (lr=1e-3, bs=32, SGD). The classifier is trained independently and then frozen after being integrated into the framework.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

UCF-Crimes Dataset Evaluation

 **Video:** Explosion005_x264






 **QA json format:**
"video_id": "Explosion008_x264",
"choices": {**"A":** "Normal", **"B":** "Vandalism 00:10 00:24 00:35 00:49", **"C":** "Vandalism 00:20 00:29",
"D": "Explosion 00:33 00:41"}
User: Known common crime types are: 'Shooting','Arson','Arrest','Fighting','RoadAccidents','Stealing','Burglary','Explosion','Robbery','Vandalism','Shoplifting','Abuse','Assault'. Each option below is the crime type and the start and end time of each criminal behavior video clip (if the video has no crime type and is normal, there is no corresponding start and end time of the video clip). A.Normal B.Vandalism 00:10 00:24 00:35 00:49 C.Vandalism 00:20 00:29 D.Explosion 00:33 00:41 \n Among the four options A, B, C, D above, which is the best answer?
Answer: The best answer is D. Explosion.

Figure 3. The evaluation format of the UCF-Crimes dataset involves extracting video information and generating a Question-Answer pairs file. Subsequently, the model is inferred with questions to obtain corresponding answers.

XD-Violence Dataset Evaluation

 **Video:** v=BXR3d22BhHs_#00-07-00_00-08-00_label_B4-0-0




 **QA json format:**
"video_id": "v=BXR3d22BhHs_#00-07-00_00-08-00_label_B4-0-0",
"choices": {**"A":** "Riot 00:04 01:05", **"B":** "Abuse 00:13 00:23 00:40 00:59", **"C":** "Riot 00:00 00:59",
"D": "Riot 00:09 00:22"}
User: Known common crime types are: 'Shooting','Abuse','Explosion','Fighting','CarAccidents','Riot'. Each option below is the crime type and the start and end time of each criminal behavior video clip. A.Riot 00:04 01:05 B.Abuse 00:13 00:23 00:40 00:59 C.Riot 00:00 00:59 D.Riot 00:09 00:22 \n Among the four options A, B, C, D above, which is the best answer?
Answer: The best answer among the four options A, B, C, and D is D. Riot.

Figure 4. The evaluation process for the XD-Violence dataset involves extracting video data and creating a Question-Answer pairs file. The model is then inferred using questions, which allows for the generation of corresponding answers.

- [2] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv preprint arXiv:2406.19568*, 2024. 1
- [3] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. Uncovering what, why and how: A comprehensive benchmark for causation understanding of video anomaly, 2024. 1
- [4] Xinyang Jiang, Fufu Yu, Yifei Gong, Shizhen Zhao, Xiaowei Guo, Feiyue Huang, Wei-Shi Zheng, and Xing Sun. Devil's in the detail: Graph-based key-point alignment and embedding for person re-id. *arXiv preprint arXiv:2009.05250*, 2020. 1
- [5] Yifan Liang and Hongbin Li. Los signal identification for passive multi-target localization in multipath environments. *IEEE Signal Processing Letters*, 30:1597–1601, 2023. 1
- [6] Yifan Liang and Hongbin Li. Source localization with spatially distributed active and passive sensors. In *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2024.
- [7] Yifan Liang, Cengcang Zeng, and Hongbin Li. Los path

Interaction about Spatial Understanding

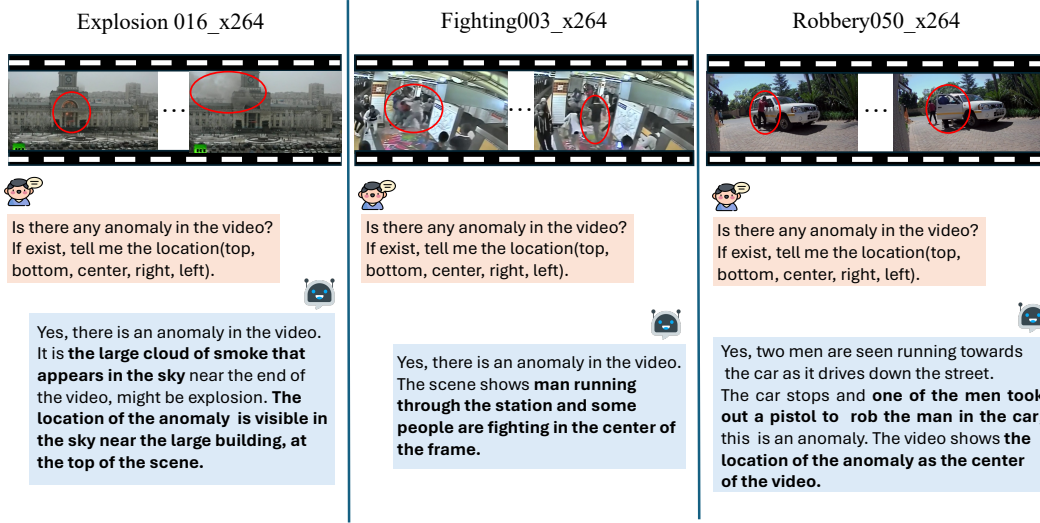


Figure 5. Accurate instances of spatial anomaly detection, illustrating our model’s proficiency in identifying the precise location of anomaly occurrences and emphasizing its effectiveness in detecting spatial anomalies.

Interaction about Temporal Understanding

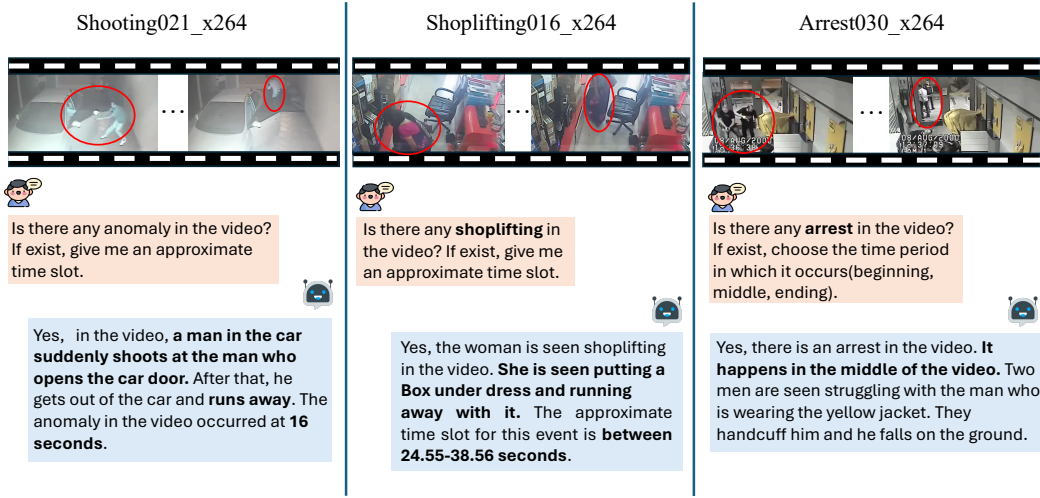


Figure 6. Successful instances of temporal anomaly detection, showcasing our model’s strong ability to identify the approximate time of anomaly occurrence and highlighting its effectiveness in detecting temporal anomalies.

identification in multipath environments with an unknown number of passive targets. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 1740–1744. IEEE, 2024. 1

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[9] Zhao Shizhen, Gao Changxin, Zhang Jun, Cheng Hao, Han Chuchu, Jiang Xinyang, Guo Xiaowei, Zheng Wei-Shi, Sang Nong, and Sun Xing. Do not disturb me: Person re-identification under the interference of other pedestrians. In

European Conference on Computer Vision (ECCV), 2020. 1

[10] Zhao Shizhen, Gao Changxin, Shao Yuanjie, Li Lerenhan, Yu Changqian, Ji Zhong, and Sang Nong. Gtnet: Generative transfer network for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1

[11] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

- [12] Haoru Tan, Sitong Wu, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [1](#)
- [13] Haoru Tan, Sitong Wu, Wei Huang, Shizhen Zhao, and Xiaojuan Qi. Data pruning by information maximization. In *International Conference on Learning Representations (ICLR)*, 2025. [1](#)
- [14] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies, 2024. [1](#), [2](#)
- [15] Yimo Yan, Yang Deng, Songyi Cui, Yong-Hong Kuo, Andy HF Chow, and Chengshuo Ying. A policy gradient approach to solving dynamic assignment problem for on-site service delivery. *Transportation Research Part E: Logistics and Transportation Review*, 178:103260, 2023. [1](#)
- [16] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset, baselines, and challenges, 2023. [1](#), [2](#)
- [17] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity, 2024. [2](#)
- [18] Shizhen Zhao and Xiaojuan Qi. Prototypical votenet for few-shot 3d point cloud object detection. In *Advances in Neural Information Processing Systems*, 2022. [1](#)
- [19] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [1](#)
- [20] Shizhen Zhao, Jiahui Liu, Xin Wen, Haoru Tan, and Xiaojuan Qi. Equipping vision foundation model with mixture of experts for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. [1](#)
- [21] Shizhen Zhao, Xin Wen, Jiahui Liu, Chuofan Ma, Chunfeng Yuan, and Xiaojuan Qi. Learning from neighbors: Category extrapolation for long-tail learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. [1](#)