MSCV Technical Report (Spring)

Bo Jiang*, Qi Long*, Siddharth Vohra*, Prof. Min Xu

Equal Contribution*

1 Introduction

1.1 Motivation

We aim at applying advanced computer vision techniques to tackle challenges in biomedical visual analysis.

- Automatically discover differences between two sets of medical images.
- Automatic lab-scene reasoning to identify anomalies and risks.
- Reveal protein locations inside a living cell at nanometer resolution and provide biological insights from it.

1.2 Topics

We research on three topics in total, here are some keywords, details can be found in the sections below.

- Comparative Decoding: Describing Differences in Image Sets.
- Bio-lab Video Anomaly Reasoning: video segmentation, object tracking, VLM CoT reasoning, agentic framework.
- Agentic VLMs for Biomedical Images: perception audit, quantitation via visual self-feedback & agentic spatial reasoning.

2 Comparative Decoding

2.1 Motivation

Traditional vision tasks focus on a single image at a time. A different and increasingly important question is how two *sets* of images differ from each other when described in natural language. Set-level difference captioning matters in at least three settings: comparing medical scans from healthy versus diseased patients, identifying how training data differs from deployment data, and understanding why a model performs differently across datasets. The standard approach captions every image and then asks a language model to compare the captions, which makes the result depend heavily on caption quality and effectively reduces an image-understanding problem to a language-comparison problem.

2.2 Related Work

The closest prior work is VisDiff [5], which formalizes set difference captioning: given two image sets $\mathcal{D}_A$ and $\mathcal{D}_B$, output a natural-language description that is more often true on $\mathcal{D}_A$ than $\mathcal{D}_B$. VisDiff is a two-stage proposer-ranker pipeline that uses a BLIP-2 captioner and a GPT-4 proposer to generate candidate differences, and a CLIP ranker to score how well each candidate separates the two sets. The pipeline is evaluated on VisDiffBench, which spans three difficulty levels (easy, medium, hard). Our project shares the same task definition but conditions language generation directly on the two image groups during decoding rather than on intermediate captions, removing the dependence on caption quality.

2.3 Approach

We take VisDiff [5] as the baseline and propose *comparative decoding*, which removes the intermediate captioning stage and conditions language generation directly on the two image groups. At each decoding step t, every candidate token v in the vocabulary is scored by the aggregated ratio of its likelihood under images in group A to its likelihood under images in group B:

$$y_t = \arg\max_{v \in V} \prod_{i=1,2,\ldots,n} \frac{P(y_t = v \mid I_A^i,\, y_{1:t-1})}{P(y_t = v \mid I_B^i,\, y_{1:t-1})}. \tag{1}$$

The token with the highest comparative score is selected, so visually discriminative concepts are emphasized during generation without an intermediate caption-comparison step.

2.4 Current Results

The method currently produces hypotheses that pick up the dominant visual contrast on small qualitative test sets, even when the exact ground-truth phras-

ing is not recovered. For example, given Group 1 images of oranges on a tree and Group 2 images of bird nests on a tree (ground-truth difference: tree attachment), the method generates "orange trees growing oranges" as the hypothesis. A second example contrasts pots on a stove with plates on a table (ground-truth difference: kitchen item placement) and yields "stove with pot loaded on." Quantitative evaluation against VisDiff on a larger benchmark is in progress.

3 Bio-lab Video Anomaly Reasoning

3.1 Introduction

Motivation. Building an VLM agent framework for video anomaly detection is a promising method for AI scene understanding, which is a critical building block for lab automation.

3.2 Related Work

- **Reconstruction-based**. Past work like [11, 6, 21, 20, 7] train a model on normal data to reconstruct the video. After that, the model is used to reconstruct anomalous samples to their corresponding normal counterparts and calculate the reconstruction error. The various models applied include techniques like deep neural network and diffusion based methods. It has the problem that it heavily relies on normal data for training, so it lacks generalization ability to unseen data and domains.

- **Embedding-based**. Past work like [13, 4, 10] focus on modeling the feature embeddings of normal samples and measure deviations for anomalies. This type of method typically follow the "one-class-one-model" learning paradigm, requiring plentiful normal samples for each object class to learn its distribution [9], lacking generalization and explanation ability.

- **Multimodal**. Some recent work [12, 17, 18, 22, 9, 23, 24] proposed VLM-based methods for their generalizable and explainable merits, especially for the zero/few-shot setting. They feature sophisticated prompt design, which heavily relies on captioning videos and assigning anomaly scores by off-the-shelf VLMs.

To enhance spatio-temporal awareness, previous VAD works rely on auxiliary spatio-temporal modeling or prompting. For instance, a VLM reasoning method [17] presents a unified zero-shot framework for video anomaly analysis that chains together temporal detection, spatial localization, and textual explanation through a test-time reasoning process over foundation models, requiring no additional training or data. However, all these methods do not have good performance on video anomaly reasoning tasks, as they still largely ignore the

fact that superficial temporal caption sequences lose focus on object-centric dynamics and interactions over time, which is the core when human inspectors detect anomalous objects and events throughout industrial processes.

3.3 Method

We propose O-VAD, object centric analysis for video anomaly detection. Specifically, it contains three stages, incorporating computer vision with text generation.

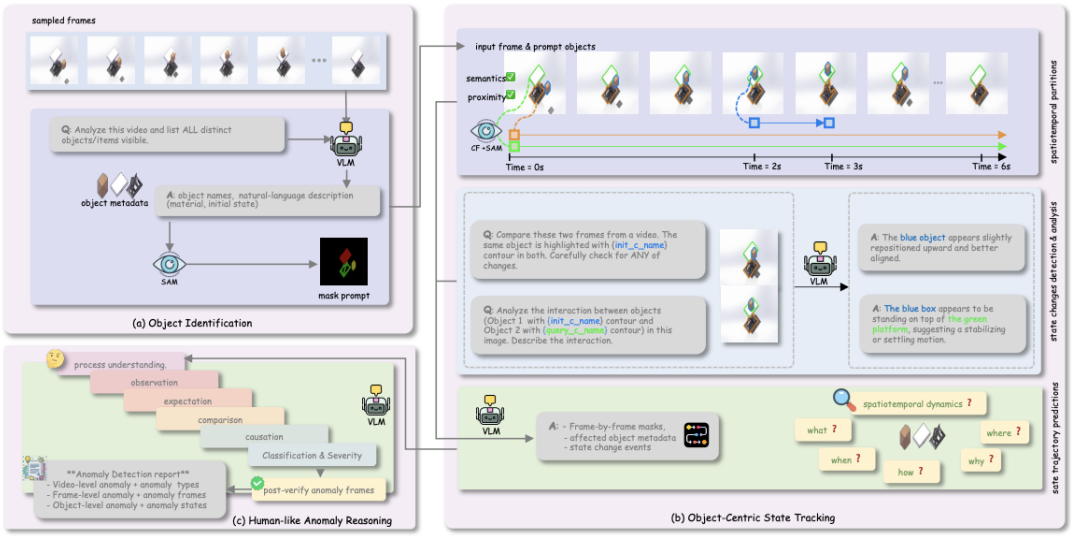

Figure 1: O-VAD: Our Bio-lab Video Anomaly Reasoning method main framework.

- **Stage 1 - Object Recognition**. Aiming at identify all objects of interest in the scene, VLM is prompted to identify objects and SAM [2] model is used for segmentation.
- **Stage 2 – Object Tracking**. TubeletGraph [19] model is used for object frame-wise state change graph generation, followed by VLM describing changes and interactions in detail.
- **Stage 3 – Anomaly Reasoning**. Given all the context gathered in Stage 1 and 2, VLM goes through CoT reasoning to output anomaly status, types, reasons and effects.

3.4 Experiments

We evaluate our method on PhysAD [15] and AutoLab [3] dataset. For video level, we compute metrics including anomaly detection accuracy, precision, re-

call, F1 score and AUROC. For type and frame level, we compute metrics including BERT score, accuracy, precision, recall, F1 score and AUROC differently for different datasets.

The results show the effectiveness of our training free generalizable method compared to other agentic reasoning methods and zero/few shot methods, and can achieve competitive results compared to traditional VAD methods, which are trained on specific domain already.

We also did some case study for detailed comparison. As shown in the case demo, our method and identify fine-grained anomalies compared to SOTA VLMs.

		Phys-AD (video / type)						LiquidAD (video / frame)					
Method	**Level**	Acc	P	R	F1	AUC	BERT	**Level**	Acc	P	R	F1	AUC
Traditional VAD (with training)													
MNAD.p† [23]	video	0.583	0.659	0.790	0.694	0.481	-	video	0.600	0.676	0.806	0.735	0.468
S3R† [30]	video	0.533	0.370	0.505	0.413	0.612	-	video	0.633	0.750	0.400	0.522	0.516
Open-Source VLMs (training-free)													
Qwen3-VL-32B [1]	video	0.454	0.717	0.368	0.486	0.513	-	video	0.150	0.886	0.075	0.139	0.489
	type	-	-	-	-	-	0.798	frame	0.876	0.000	0.000	0.000	0.499
Agentic Reasoning (training-free)													
URF-ZS-HVAA [19]	video	0.336	0.689	0.051	0.069	0.499	-	video	0.356	1.000	0.033	0.065	0.423
	type	-	-	-	-	-	-	frame	0.701	0.444	0.033	0.062	0.476
O-VAD (Ours)	video	0.592	0.724	0.625	0.621	0.584	-	video	0.868	0.910	0.948	0.929	0.692
	type	-	-	-	-	-	0.803	frame	0.458	0.431	0.614	0.507	0.512

Figure 2: Experiment of O-VAD on PhysAD and AutoLab datasets.

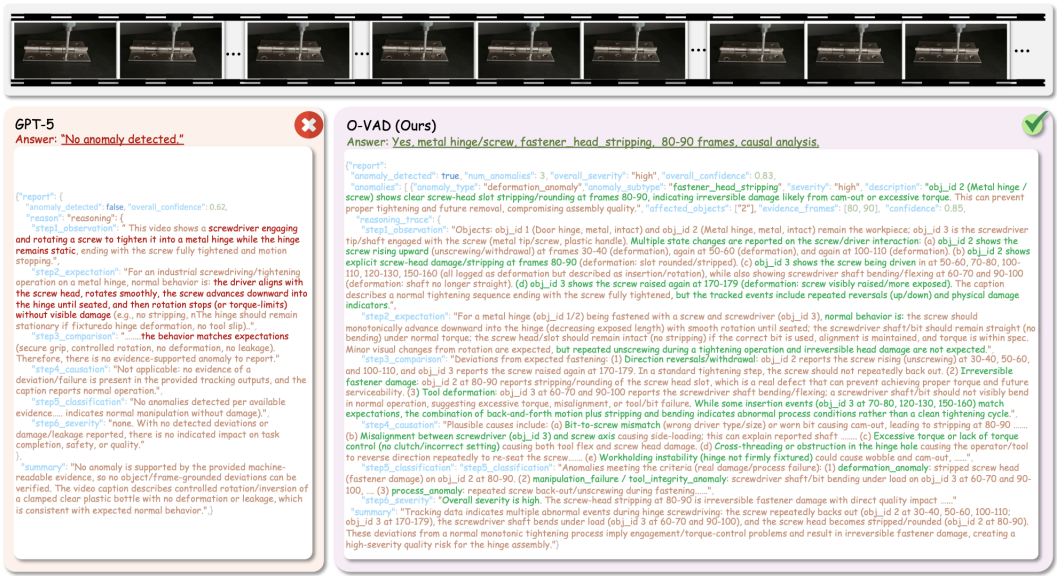

Figure 3: Case Study of O-VAD.

4 Agentic VLMs for Biomedical Image Analysis

4.1 Introduction

Before vision-language models can serve as autonomous readers of biomedical imagery, three capabilities must hold simultaneously: faithful perception of the underlying scalar field, reliable quantitative tool use, and structured spatial reasoning over volumetric data. The original proposal targeted each capability with a separate sub-project. Over the course of the semester these have consolidated into two ongoing projects: a perception-audit project (*Mutual Colormap Dependence*, MCD) and a tool-grounded retrieval project (*Computation-Conditioned Retrieval*, CCR). The standalone quantitation arm (BBBC039 nuclei counting via a six-phase visual self-feedback agent) is now retained inside the CCR project as a negative datapoint that fails the computability boundary condition, and is no longer tracked as an independent contribution.

4.2 Related Work

Our work on biomedical VLM reliability builds on three lines of prior research. On 3D medical image analysis, M3D [1] introduces M3D-LaMed, a 3D ViT coupled with an LLM and trained on 120K scan-report pairs from M3D-Cap, demonstrating that volumetric encoders can drive report generation, VQA, and language-guided segmentation on chest CT; its single-turn VQA setup, however, leaves multi-step agentic spatial reasoning unexplored, which motivates the iterative tool-grounded loop in our CCR-Agent. On color robustness in VLMs, ColorBench [16] catalogues substantial color-robustness gaps across 32 VLMs under natural-image color transformations. Our MCD project asks a narrower invariance question grounded in scientific imagery: rather than reporting accuracy gaps under arbitrary recoloring, we measure the conditional mutual information between a finite orbit of monotonic, intensity-preserving colormaps and the model's answer for a fixed image-question pair. On retrieval-augmented generation, the original RAG framework [14] retrieves passages keyed on the user question; CCR instead conditions retrieval on a deterministically computed fact derived from per-instance structured data, distinguishing it from hypothesis-keyed retrieval such as HyDE [8], where the key is generated from parametric memory rather than grounded in the input.

4.3 Project 1: Mutual Colormap Dependence (MCD)

Scientific images encode a scalar measurement field through a colormap, but a VLM used as a reader of that field should not change its answer when only the colormap changes. We define MCD as the conditional mutual information $I(c; a \mid I, q)$ in nats between a finite orbit of monotonic, intensity-preserving lookup tables and the model's answer for a fixed image-question pair (I, q). MCD is non-negative, equals zero if and only if the model is colormap-invariant, and is upper-bounded by $\log K$ for an orbit of size K. Under black-box

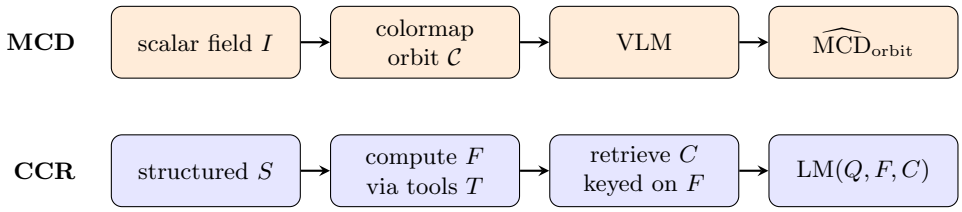

Figure 4: Two complementary audits for agentic biomedical VLMs. *Top*: MCD measures whether the answer distribution carries information about a scientifically irrelevant display variable across a finite orbit of intensity-preserving colormaps. *Bottom*: CCR replaces question-keyed retrieval with retrieval keyed on a deterministically computed fact F obtained by running tools on per-instance structured data S.

API constraints we report a finite-orbit audit score together with stochasticity, deterministic-decoding, and same-render diagnostics that calibrate decoding noise. The current audit covers GPT-5.4, Gemini 3.1 Pro, Claude Sonnet 4.6, and Qwen3-VL-235B on SLAKE medical VQA, a ChestX-ray14 robustness proxy, and a BBBC039 exact-count emission stress test. The project has been submitted to EMNLP 2026.

4.4 Project 2: Computation-Conditioned Retrieval (CCR)

Standard retrieval-augmented generation retrieves passages keyed on the question Q. When the answer depends on per-instance structured data S (3D coordinate sets, tabular records, scientific graphs), question-keyed retrieval can return passages irrelevant to the actual answer. CCR is the three-step pattern $F = \mathrm{compute}(S, Q, T)$, $C = \mathrm{retrieve}(F, K)$, $A = \mathrm{LM}(Q, F, C)$, in which the retrieval key is a deterministically computed fact rather than the question. The central contribution under development is a falsifiable predictor with three boundary conditions for when CCR is the active mechanism: decoupling, computability, and retrieval-required. The current evaluation suite covers four datapoints: cryo-electron tomography QA (CryoBioQA), HybridQA, OTT-QA, and a synthetic decoupled benchmark (TableBench-Decoupled-30) constructed in this project. The project is targeting EMNLP 2026.

4.5 Current Results

- **MCD audit signal.** Every audited frontier VLM currently shows a positive finite-orbit score on all three scientific panels. On the two medical display-invariance panels, scores so far range from 0.13 to 0.47 nats. Cross-colormap disagreement, recast as a non-conformity score, currently beats random abstention and one-call self-reported confidence on SLAKE and ChestX-ray14 at the six-call orbit budget for every audited model.

- **CCR cryo-ET case study.** On the 42-question CryoBioQA core set, the CCR-Agent reaches 85.7% versus 19.0% for VLM+RAG, a +66.7-point gap. Three text-only retrievers (BM25, HippoRAG, LightRAG) all currently score 0/42, ruling out alternative-design retrieval as the source of the lift. The full-loop variants with and without retrieval are statistically indistinguishable, which we currently read as isolating tool-grounded computation rather than retrieval keying as the active mechanism on this domain.

- **CCR positive demonstration.** On the synthetic TBD-30 benchmark constructed to satisfy all three boundary conditions, Full CCR currently reaches $27/30 = 0.900$ and matches the Oracle-F upper bound across three frontier backbones, while every retrieval-disabling variant collapses at or below 0.033.

- **Boundary-conditions truth table so far.** HybridQA and OTT-QA violate decoupling and currently show null effects; Case A microscopy violates computability and currently shows inverted effects (tools hurt); cryo-ET violates the retrieval-required condition and currently shows the loop active with retrieval acting as a no-op; TBD-30 satisfies all three conditions and currently shows F-keyed retrieval as load-bearing.

References

[1] Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. M3D: Advancing 3D medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026.

[3] Ali Dabouei, Jishnu Parayil Shibu, Vibhu Dalal, Chengzhi Cao, Andy MacWilliams, Joshua Kangas, and Min Xu. Deep video anomaly detection in automated laboratory setting. *Expert Systems with Applications*, 271:126581, 2025.

[4] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer, 2021.

[5] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[6] Lei Fan, Junjie Huang, Donglin Di, Anyang Su, Maurice Pagnucco, and Yang Song. Revitalizing reconstruction models for multi-class anomaly detection via class-aware contrastive learning. *arXiv preprint arXiv:2412.04769*, 2024.

[7] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023.

[8] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Findings of the Association for Computational Linguistics (ACL)*, 2023.

[9] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large

vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

[10] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20405–20415, 2025.

[11] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024.

[12] Chao Huang, Benfeng Wang, Jie Wen, Chengliang Liu, Wei Wang, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *arXiv preprint arXiv:2505.19877*, 2025.

[13] Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, and Byung Jun Kang. Reconpatch: Contrastive patch representation learning for industrial anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2052–2061, 2024.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Wenqiao Li, Yao Gu, Xintao Chen, Xiaohao Xu, Ming Hu, Xiaonan Huang, and Yingna Wu. Towards Visual Discrimination and Reasoning of Real-World Physical Dynamics: Physics-Grounded Anomaly Detection . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30409–30419, Los Alamitos, CA, USA, June 2025. IEEE Computer Society.

[16] Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. ColorBench: Can VLMs see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[17] Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *arXiv preprint arXiv:2511.00962*, 2025.

[18] Shiwei Lin, Chenxu Wang, Xiaozhen Ding, Yi Wang, Boyuan Du, Lei Song, Chenggang Wang, and Huaping Liu. A vlm-based method for visual anomaly detection in robotic scientific laboratories. In *2025 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 34–39. IEEE, 2025.

[19] Yihong Sun, Xinyu Yang, Jennifer J Sun, and Bharath Hariharan. Tracking and understanding object transformations. *Advances in Neural Information Processing Systems*, 2025.

[20] H Zhang, Z Wang, Z Wu, and YG Jiang. Diffusionad: norm-guided one-step denoising diffusion for anomaly detection (2023).

[21] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16699–16708, 2024.

[22] Shifang Zhao, Yiheng Lin, Lu Han, Yao Zhao, and Yunchao Wei. Omniad: Detect and understand industrial anomaly via multimodal reasoning. *arXiv preprint arXiv:2505.22039*, 2025.

[23] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025.

[24] Shu Zou, Xinyu Tian, Lukas Wesemann, Fabian Waschkowski, Zhaoyuan Yang, and Jing Zhang. Unlocking vision-language models for video anomaly detection via fine-grained prompting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4223–4233, 2026.