Carnegie Mellon University

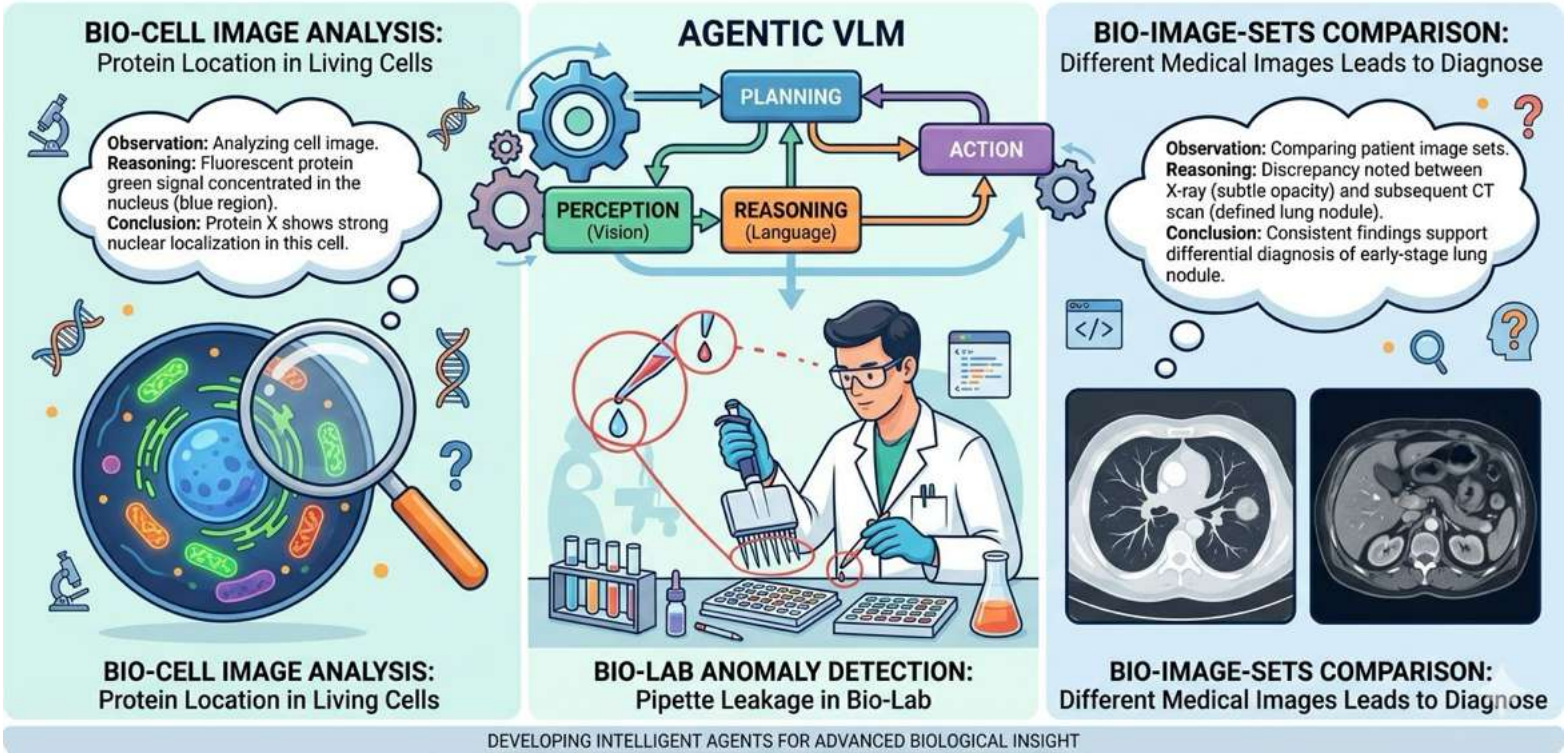
Agentic Vision-Language Models for Biomedical Anomaly Detection and Differential Biomedical Image Analysis

Bo Jiang
Qi Long
Siddharth Vohra
Prof. Min Xu

MSCV Capstone I - Spring 2026

Overview

Motivation - Biomedical Visual Analysis



Generated using AI, Source: Google Nano Banana

Comparative Decoding

Background

The Task

- Traditional vision tasks focus on **single images**
- New Taks: Describe the semantic differences between two **image sets** using natural language.

Why it is important?

- Medical Imaging: compare scans from healthy vs diseased patients.
- Dataset Shift Detection: identify how training data differs from real deployment data.
- Model Debugging: understand why a model performs differently across datasets.

Comparative Decoding

Limitations of Current Method:

- Strong dependence on **caption quality**; weak captions lead to weak comparisons.
- Converts an image understanding task into a **language comparison** task.

Proposed New Method:

- Compare two image groups directly during **decoding** instead of comparing captions afterward.
- At each token step, combine evidence from many images in Group A and contrast it with Group B.

Population Comparative Decoding

- At decoding step **t**, evaluate every candidate token **v** in the vocabulary.
- Compute token **likelihoods** across both groups and select tokens using an aggregated Group A vs Group B probability **ratio score**.
- Select the token with the **highest comparative score**.

Example:

- Group 1: Oranges on a tree
- Group 2: Bird nests on a tree
- Ground Truth Difference: Tree attachment (Oranges vs Bird nests)
- Generated Hypotheses: orange trees growing oranges

Bio-lab Video Anomaly Reasoning

Background

Problem Setting

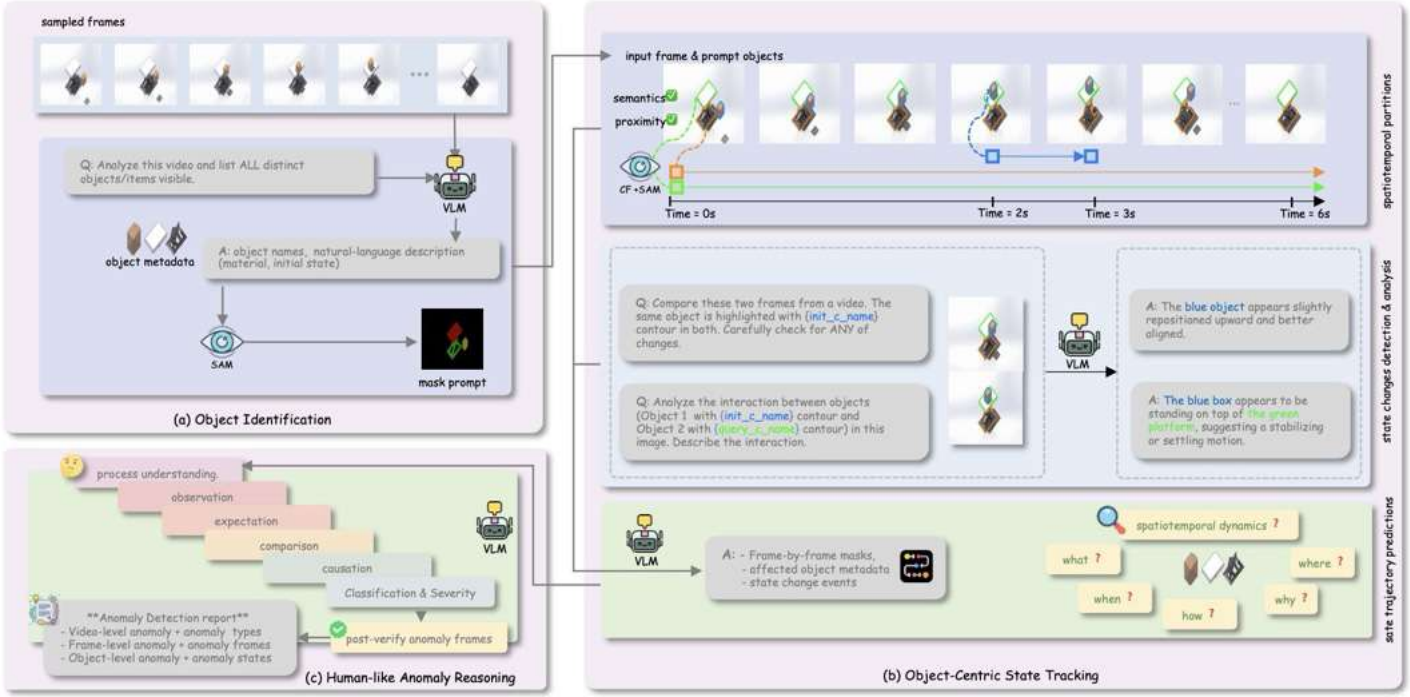
- **Video Anomaly Detection** is a critical building block for bio-lab automation

Motivation

- Building an **VLM agent framework** for video anomaly detection is a promising method for AI scene understanding
- Compared to traditional methods, we reduce reliance on normal data and offer anomaly reasoning along with detection
- Compared to VLM based methods, we achieve deeper scene understanding

Method

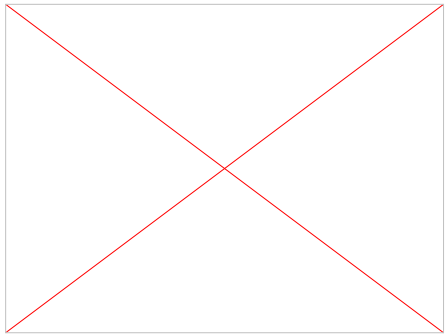
- Stage 1 – Object Recognition
- Stage 2 – Object Tracking
- Stage 3 – Anomaly Reasoning

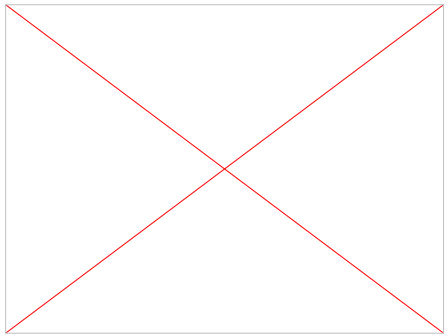


Results

Method	Phys-AD (video / type) Level	Acc	P	R	F1	AUC	BERT	LiquidAD (video / frame) 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

Demo





❌: ***Closed plastic bottles*** *release liquid during manipulation, indicating* ***rupture/puncture or seal failure.***

🤖: *The* ***4th pipette*** *tip/channel repeatedly shows* ***reduced/absent liquid column and inconsistent dispensing compared to the other tips****. Evidence includes…. This indicates a channel-specific failure in aspiration/retention/dispensing*

Agentic VLMs for Biomedical Image Analysis

Carnegie Mellon University

Motivation

Problem Setting:

- Frontier VLMs make **unreliable predictions** on scientific imagery
- Standard RAG retrieves on the **question**, ignoring per-image evidence

Goals:

- Frontier VLMs make **unreliable predictions** on scientific imagery
- Standard RAG retrieves on the **question**, ignoring per-image evidence

Part 1: Do VLMs See, or Read the Colormap?

The Question

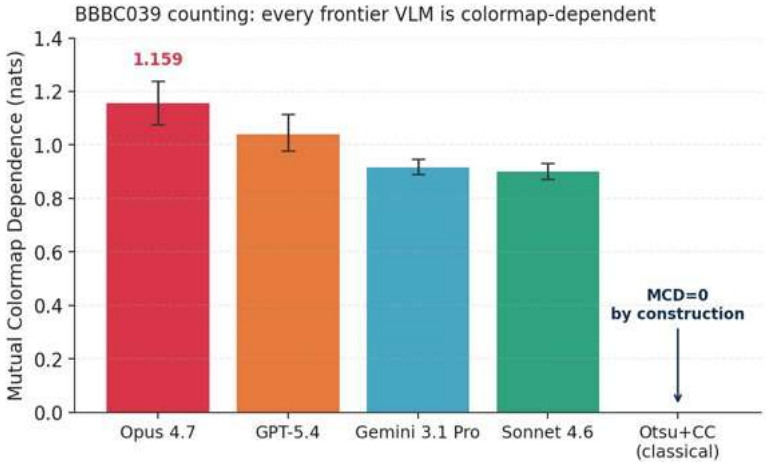
If a VLM truly perceives the scalar field, it should give the same answer regardless of colormap.

Our Measure

- **MCD**: mutual information (nats) between colormap and answer
- MCD = 0 iff colormap-invariant

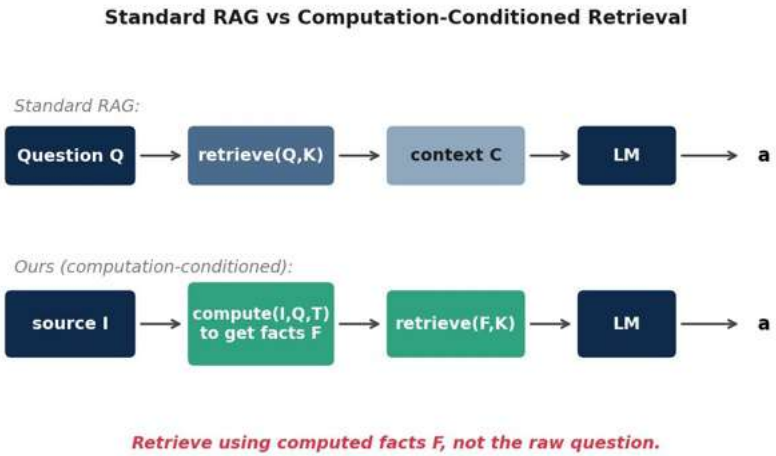
The Claim

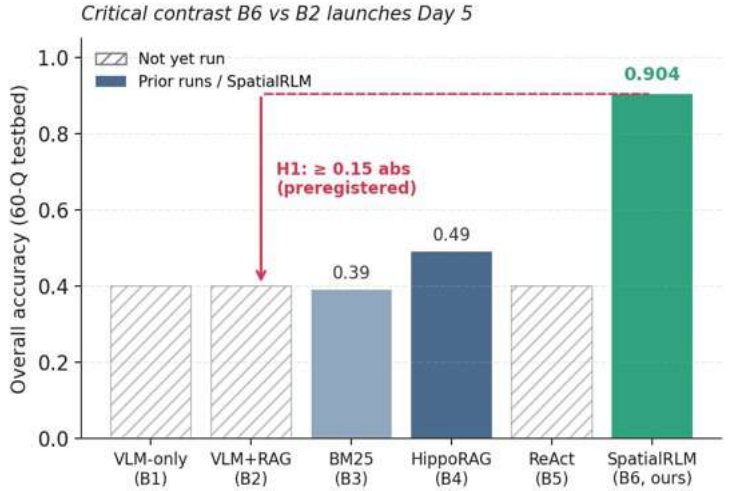
Every frontier VLM is colormap-dependent. Classical CV achieves MCD=0 by construction at similar error.



Fix in progress: LoRA adapter on Qwen3-VL trained on random monotonic splines; target 30% MCD reduction with 2pp max clean-accuracy loss.

Part 2: Computation-Conditioned Retrieval





- **SpatialRLM** (cryo-ET): agent runs spatial tools, retrieves UniProt+PubMed using computed facts F, answers biological questions
- **Preregistered H1**: computation-conditioned beats question-conditioned by 0.15+ absolute accuracy, both seeing the image
- **VELM** (microscopy): tool-only pipeline nearly matches full agent; language-model loops are not the bottleneck

THANK YOU