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

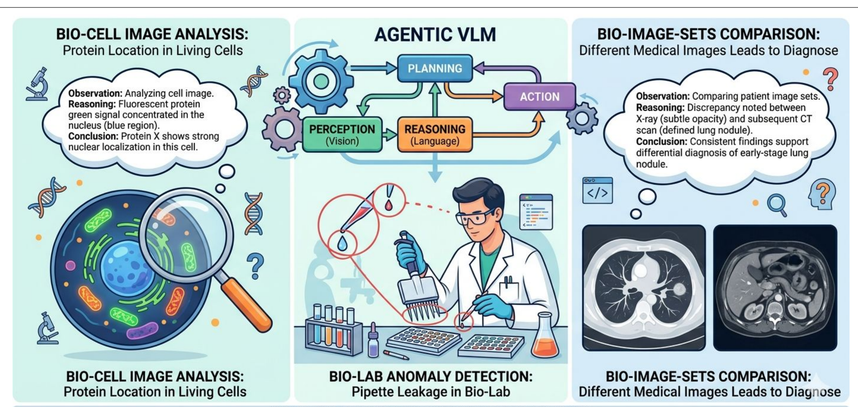
Bo Jiang*, Qi Long*, Siddharth Vohra*, Prof. Min Xu

Carnegie Mellon University, Robotics Institute

**equal contribution, names ordered alphabetically*

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



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

Comparative Decoding

Motivation: How do two sets of images differ? Understanding dataset differences is important for: analyzing datasets, diagnosing model failures, discovering hidden patterns in large visual collections.

Baseline: VisDiff: Describing Differences in Image Sets with Natural Language

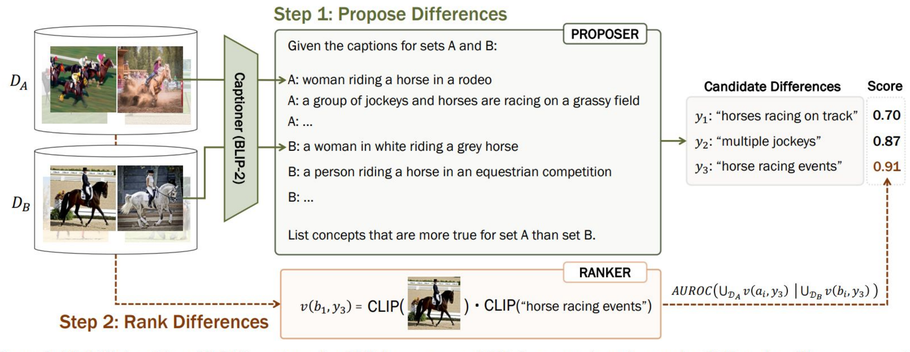


Figure 2. **VisDiff algorithm.** VisDiff consists of a *GPT-4 proposer* on *BLIP-2* generated captions and a *CLIP ranker*. The *proposer* takes randomly sampled image captions from $\mathcal{D}_A$ and $\mathcal{D}_B$ and proposes candidate differences. The *ranker* takes these proposed differences and evaluates them across all the images in $\mathcal{D}_A$ and $\mathcal{D}_B$ to assess which ones are most true.

Proposed method: We propose **comparative decoding**, where language generation is directly conditioned on two image groups during decoding. The method measures token importance by comparing how likely each word is under images from group A versus group B, allowing visually discriminative concepts to be emphasized during generation.

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

Example 2:

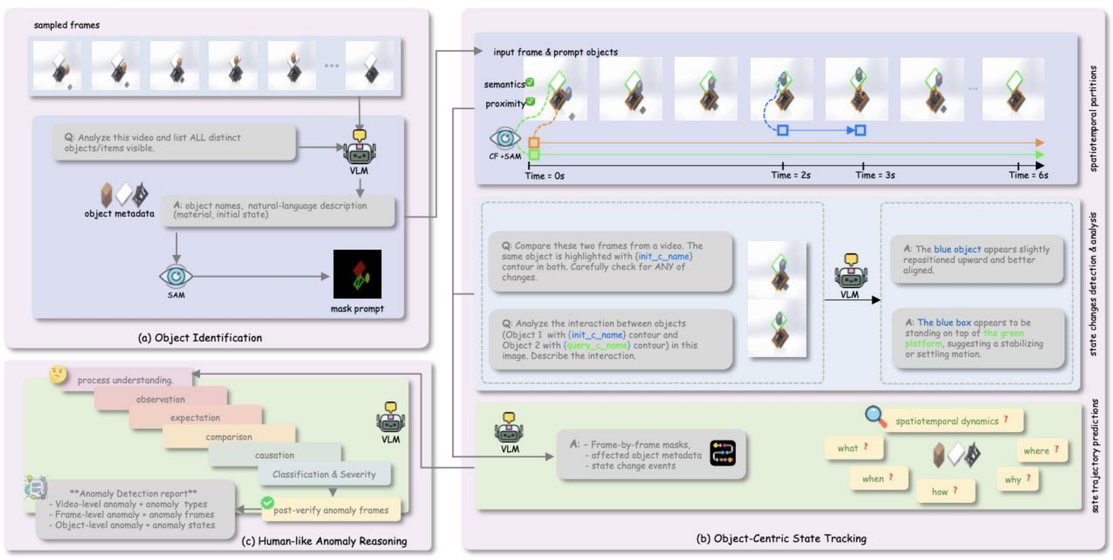
- Group 1: Oranges on a tree
- Group 2: Bird nests on a tree
- Ground Truth Difference: Tree attachment (Oranges vs Bird nests)
- Generated Hypotheses: orange trees growing oranges

Example 1:

- Group 1: Pots on a stove
- Group 2: Plates on a table
- Ground Truth Difference: Kitchen item placement (Stove vs Table)
- Generated Hypotheses: stove with pot loaded on

Bio-lab Video Anomaly Reasoning

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.

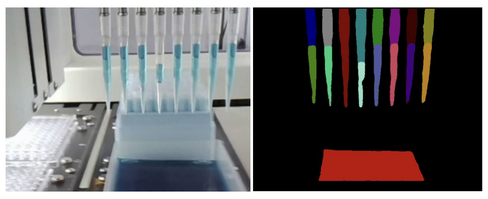


Framework. We propose O-VAD, object centric analysis for video anomaly detection. Specifically, it contains three stages, incorporating computer vision with text generation.

- **Stage 1 – Object Recognition**. Aiming at identify all objects of interest in the scene, VLM is prompted to identify objects and SAM [b1] model is used for segmentation.
- **Stage 2 – Object Tracking**. TubeletGraph [b2] model is used for object frame-wise state 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.

Experiment. We evaluate our method on AutoLab [b3] dataset, on one 4-way A40 GPU.

Method	Level	LiquidAD (video / frame) Acc	P	R	F1	AUC
MNAD.p† [23]	video	0.600	0.676	0.806	0.735	0.468
S3R† [30]	video	0.633	0.750	0.400	0.522	0.516
Qwen3-VL-32B [1]	video	0.150	0.886	0.075	0.139	0.489
	frame	0.876	0.000	0.000	0.000	0.499
URF-ZS-HVAA [19]	video	0.356	1.000	0.033	0.065	0.423
	frame	0.701	0.444	0.033	0.062	0.476
O-VAD (Ours)	video	0.868	0.910	0.948	0.929	0.692
	frame	0.458	0.431	0.614	0.507	0.512



🤖: *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*

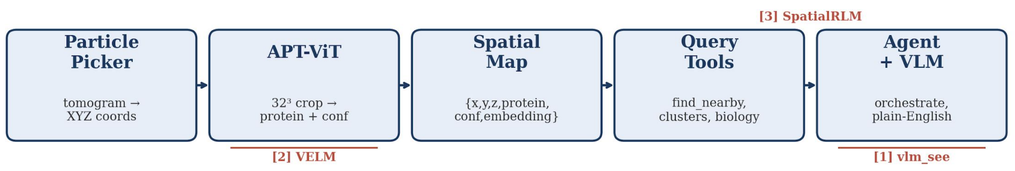
[b1] Nicolas Carion et al., 2025, SAM 3: Segment Anything with Concepts
[b2] Yihong Sun et al., 2025, Tracking and Understanding Object Transformations (NeurIPS)
[b3] Ali Dabouei et al., 2025, Deep video anomaly detection in automated laboratory setting

Agentic VLMs for Biomedical Image Analysis

Motivation. 3 capabilities must hold before VLMs can interpret biomedical images at cell level: faithful perception, reliable quantitation, & spatial reasoning. Each workstream improves one

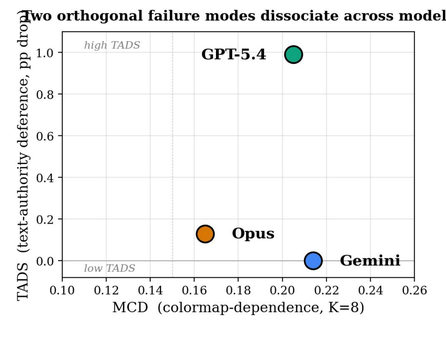
Framework. Integrating case study: end-to-end cryo-ET cell interpretation pipeline.

- **Particle Picker**: raw tomogram → XYZ coordinates
- **APT-ViT**: 32×32×32 crop → protein class + confidence
- **Spatial Map**: batch-classify all proteins → {x, y, z, protein, confidence, embedding}
- **Query Tools**: find_nearby, analyze_distribution, lookup_biology, analyze_clusters
- **Agent + VLM**: orchestrates tools, inspects uncertain cases, answers in plain English



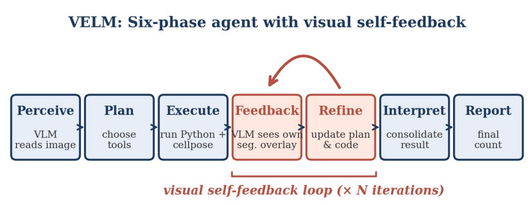
[1] VLM_SEE: Perception Audit

Motivation. Do VLMs see scientific imagery or shortcut via colormap priors and metadata?
Method. Causal 2×2 probe (pixels × metadata); two metrics, MCD (color-dependence) and TADS (text-authority deference).
Finding. Failure modes dissociate across model families (see figure).



[2] VELM: Quantitation via Visual Self-Feedback

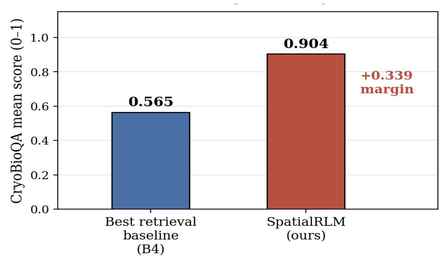
Motivation. VLMs describe microscopy fluently but cannot count what they describe.
Method. Six-phase agent with a Feedback→Refine loop on its own segmentation overlay.
Finding. On BBBC039 nuclei counting, MAE ratio 0.589 (kill-gate ≤ 0.70 passed).



Result on BBBC039 nuclei counting
MAE ratio vs. single VLM call = 0.589
(kill-gate ≤ 0.70 PASSED)

[3] SpatialRLM: Agentic Spatial Reasoning

Motivation. 3D spatial reasoning over tomograms needs on-demand computation + KB retrieval; RAG and long-context degrade as particle counts grow.
Method. Agent with persistent Python REPL + 7 spatial tools; new 60-question CryoBioQA benchmark (5 question types).
Finding. 0.904 vs. 0.565 best baseline (+0.339 margin, single seed, Claude Opus 4.7).



Caveat: B4 narrowly beats SpatialRLM on T5 comparative Qs (0.911 vs 0.867).