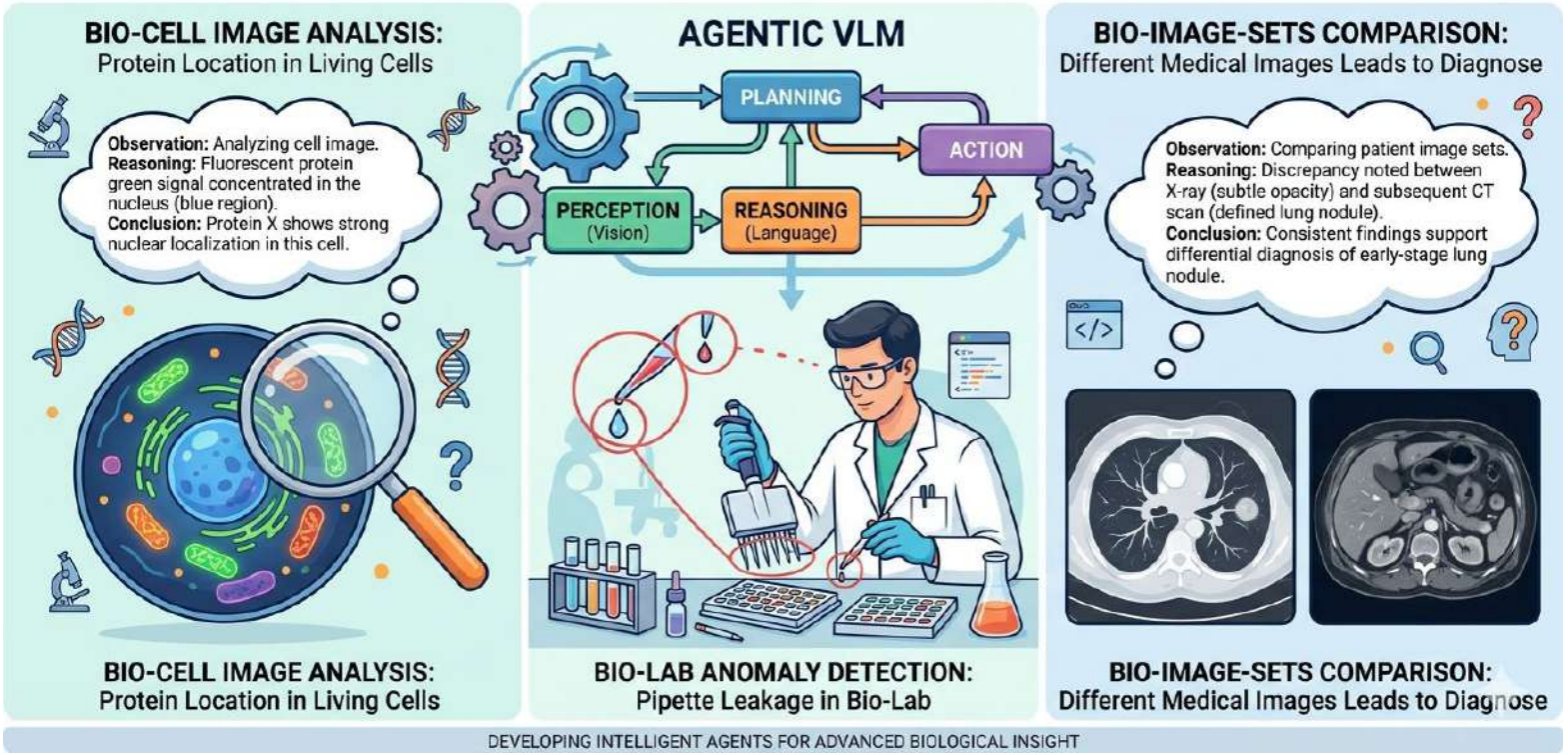
Carnegie Mellon University

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

Motivation - Biomedical Visual Analysis



Generated using AI, Source: Google Nano Banana

Problem

Bio-lab Scene Analysis

- Current Visual Anomaly Detection methods **cannot adapt to lab scenes, and rely heavily on training data**

Describing Differences in Image Sets

- Vision models excel at **single-image** analysis but struggle to detect **dataset-level** differences.

3D Spatial Object Recognition for Cryo-ET

- Current models classify proteins one at a time with **no spatial context**
- Raw coordinates require hours of manual analysis per tomogram with no **tools/interface** to query them

Solution

Bio-lab Anomaly Detection

- Utilize object detection and segmentation methods for video analysis
- Object tracking based anomaly detection to guide VLM reasoning

Describing Differences in Image Sets

- Use a VLM to generate captions, estimate token importance between two image groups and produce difference descriptions through comparative decoding.

3D Spatial Object Recognition for Cryo-ET

- Create a pipeline that takes a raw tomogram, does particle picking followed by APT-ViT classification, builds a full spatial map, and exposes it through four query tools

PAPER SURVEY

Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

Carnegie Mellon University

Background

Traditional Video Anomaly Detection

- reconstruction normal events from abnormal ones
- modeling normal events → hidden space comparison

Rely on Normal Data
No Rationale

VLM-based Video Anomaly Detection

- vision encoder: Vision Transformer (ViT)
- LLM backbone: causal text model

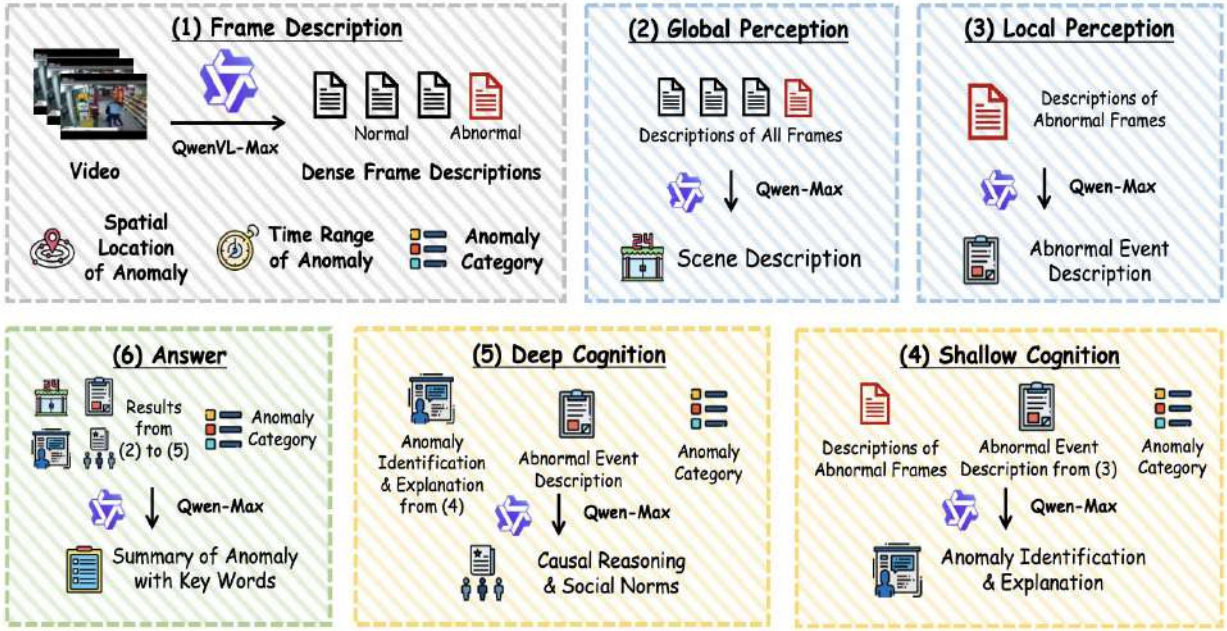
Low Accuracy

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

Video Anomaly Reasoning

Key Takeaways:

- ***VLMs** for anomaly reasoning*
- *Transform video understanding to **Agentic NLP** task*



Video → VLM → frame **Caption**

frame **Caption** → LLM → scene and abnormal description

abnormal frame **Caption** → LLM → anomaly **Reasoning** & categ

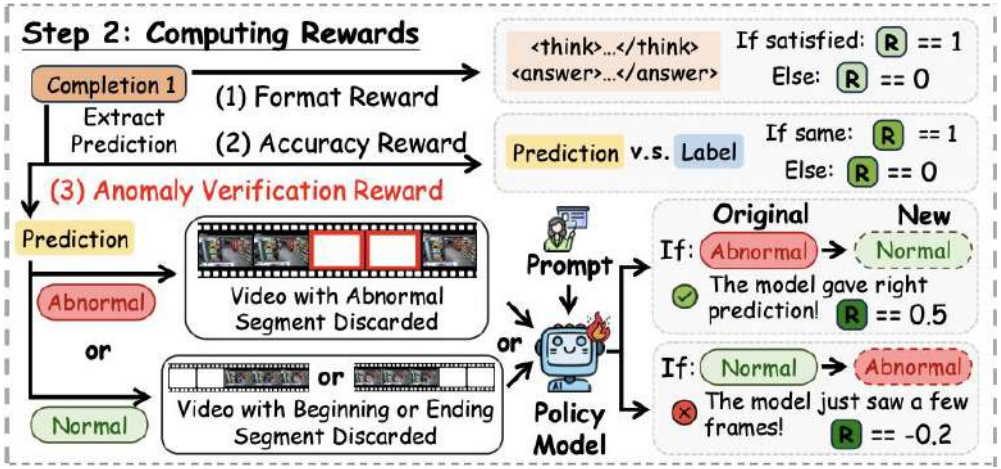
all texts → LLM → final **Answer**

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

VLM Training

Key Takeaways:

- ***SFT & RL** training*



Reasoning & Answer → SFT (1,755 samples)

Reasoning & Answer → parse format + label match + video trim comparison for **reward** → RL (6,448 samples)

$$\mathcal{L}_{SFT}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\log P_\theta(y_t^{(i)} \mid y_{<t}^{(i)}, x^{(i)})$$

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{q,O\}}\left[\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}A_i,\ \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right)\right], \qquad A_i = \frac{r_i - \text{mean}(R)}{\text{std}(R)}$$

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

Experiments

Method	Params.	Anomaly Reasoning			Anomaly Detection				
		BLEU-2	METEOR	ROUGE-2	Acc	F1	mIoU	R@0.3	R@0.5
Open-Source video MLLMs									
InternVideo2.5 [65]	8B	0.110	0.264	0.109	0.715	0.730	0.417	0.458	0.424
InternVL3 [92]	8B	0.124	0.286	0.116	0.779	0.756	0.550	0.613	0.540
VideoChat-Flash [27]	7B	0.012	0.084	0.047	0.683	0.487	0.536	0.538	0.358
VideoLLaMA3 [82]	7B	0.066	0.200	0.092	0.665	0.624	0.425	0.451	0.419
LLaVA-NeXT-Video [89]	7B	0.094	0.238	0.104	0.651	0.423	0.576	0.601	0.585
Qwen2.5-VL [57]	7B	0.113	0.264	0.116	0.761	0.730	0.567	0.610	0.563
Open-Source video reasoning MLLMs									
Open-R1-Video [63]	7B	0.060	0.179	0.084	0.793	0.790	0.559	0.642	0.540
Video-R1 [14]	7B	0.135	0.317	0.132	0.624	0.694	0.334	0.392	0.328
VideoChat-R1 [28]	7B	0.128	0.287	0.123	0.793	0.790	0.559	0.642	0.540
MLLM-based VAD methods									
Holmes-VAD [84]	7B	0.003	0.074	0.027	0.565	0.120	-	-	-
Holmes-VAU [85]	2B	0.077	0.182	0.075	0.490	0.371	-	-	-
HAWK [50]	7B	0.042	0.156	0.042	0.513	0.648	-	-	-
Proprietary MLLMs									
Claude3.5-Haiku [2]	-	0.097	0.253	0.098	0.580	0.354	0.518	0.543	0.524
GPT-4o [40]	-	0.154	0.341	0.133	0.711	0.760	0.472	0.565	0.476
Gemini2.5-Flash [51]	-	0.133	0.308	0.120	0.624	0.707	0.370	0.437	0.358
Proprietary reasoning MLLMs									
Gemini2.5-pro [52]	-	0.145	0.356	0.137	0.829	0.836	0.636	0.722	0.638
QVQ-Max [56]	-	0.142	0.318	0.121	0.702	0.747	0.430	0.503	0.412
o4-mini [42]	-	0.106	0.263	0.109	**0.884**	**0.875**	0.644	0.736	0.631
Vad-R1 (Ours)	7B	**0.233**	**0.406**	**0.194**	0.875	0.862	**0.713**	**0.770**	**0.706**

Key Takeaways:

- *Achieve **SOTA** performance*
- *Can **NOT** solve bio-lab domain*

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
s)						
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

Evaluation on our dataset gives same-as-random-guessing performance

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

Conclusion

VLM agentic reasoning framework works well

LLM training methods works well

Lack bio-medical lab scene-specific video pre-analysis

Huang et al.,Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought, NIPS 2025

VisDiff: Describing Differences in Image Sets with Natural Language

Dunlap, Lisa & Zhang, Yuhui & Wang, Xiaohan & Zhong, Ruiqi & Darrell, Trevor & Steinhardt, Jacob & Gonzalez, Joseph & Yeung-Levy, Serena. (2024). Describing Differences in Image Sets with Natural Language. 24199-24208. 10.1109/CVPR52733.2024.02284.

Carnegie Mellon University

VisDiff: Describing Differences in Image Sets with Natural Language

Goal: Given two datasets **A** and **B**, generate a description y that is **more true** for **A** than **B**.

Example:

Challenges: We want a description that can **most effectively differentiate** between the two sets. **“birthday party”** is a valid difference, but **“people posing for a picture”** better separates the sets.

Dunlap, Lisa & Zhang, Yuhui & Wang, Xiaohan & Zhong, Ruiqi & Darrell, Trevor & Steinhardt, Jacob & Gonzalez, Joseph & Yeung-Levy, Serena. (2024). Describing Differences in Image Sets with Natural Language. 24199-24208. 10.1109/CVPR52733.2024.02284.

VisDiff Framework

Method Overview:

Two-stage pipeline: **Proposer → Ranker**

- Step 1: Generate candidate descriptions
- Step 2: Evaluate which description best separates the datasets

Proposer:

- Image-based: Vision model compares images directly
- Feature-based: Embedding difference
- Caption-based: Generate captions then compare text

Ranker:

- image-based (VQA): Use a VQA model to check if the description matches the image.
- caption-based: Generate image captions and test if the description appears in the caption.
- feature-based (CLIP): Use CLIP similarity between image and text embeddings.

Dunlap, Lisa & Zhang, Yuhui & Wang, Xiaohan & Zhong, Ruiqi & Darrell, Trevor & Steinhardt, Jacob & Gonzalez, Joseph & Yeung-Levy, Serena. (2024). Describing Differences in Image Sets with Natural Language. 24199-24208. 10.1109/CVPR52733.2024.02284.

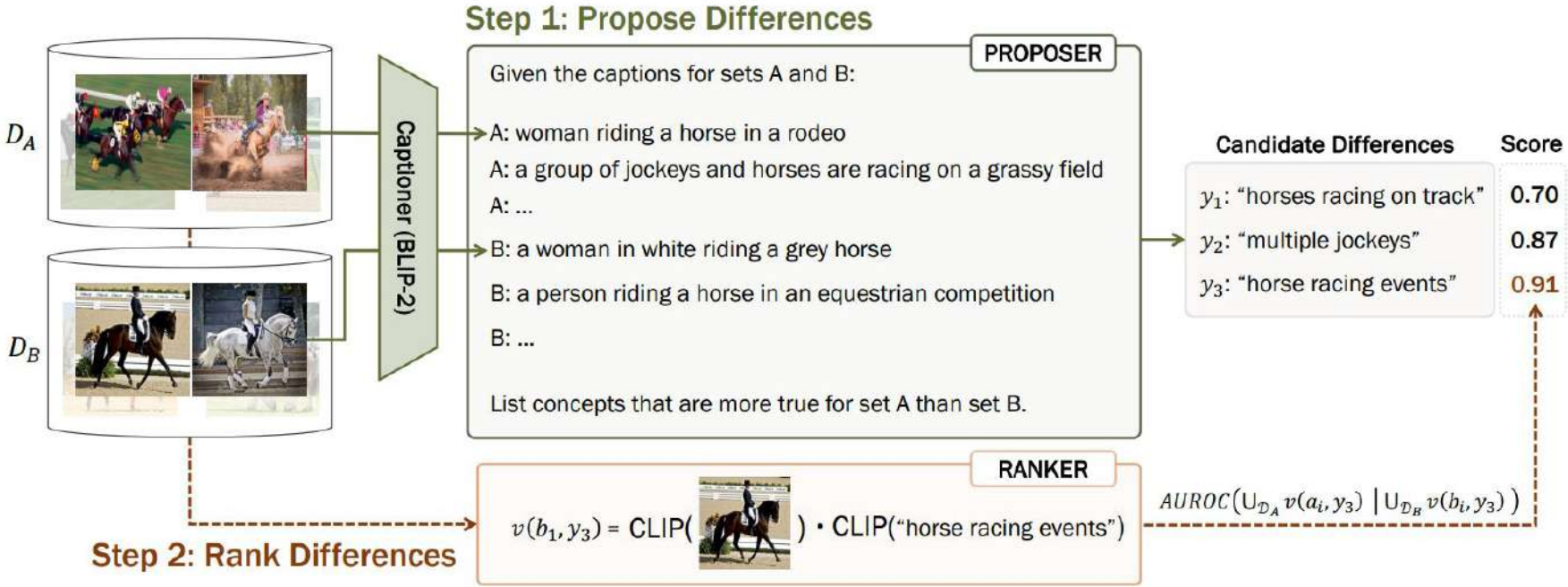


Figure 2. **VisDiff algorithm.** VisDiff consists of a *GPT-4 proposer* on *BLIP-2* generated captions and a *CLIP ranker*. The *proposer* takes randomly sampled image captions from $\mathcal{D}_A$ and $\mathcal{D}_B$ and proposes candidate differences. The *ranker* takes these proposed differences and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

Dunlap, Lisa & Zhang, Yuhui & Wang, Xiaohan & Zhong, Ruiqi & Darrell, Trevor & Steinhardt, Jacob & Gonzalez, Joseph & Yeung-Levy, Serena. (2024). Describing Differences in Image Sets with Natural Language. 24199-24208. 10.1109/CVPR52733.2024.02284.

Dataset and Results

VisDiffBench: a benchmark designed to evaluate algorithms that describe differences between image sets.

The dataset includes three difficulty levels:

- **Easy** – obvious differences: **dogs vs cats**
- **Medium** – fine-grained differences: **SUVs vs sedans**
- **Hard** – subtle semantic differences: **yoga vs meditation**

Proposer	Ranker	ImageNet-R/*		PIS-Easy		PIS-Medium		PIS-Hard	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Feature (BLIP-2)	Feature (CLIP)	0.68	0.85	0.48	0.69	0.13	0.33	0.12	0.23
Image (LLaVA-1.5)	Feature (CLIP)	0.27	0.39	0.71	0.81	0.39	0.49	0.28	0.43
Caption (BLIP-2 + GPT-4)	Caption (Vicuna-1.5)	0.42	0.70	0.60	0.92	0.49	0.77	0.31	0.61
Caption (BLIP-2 + GPT-4)	Image (LLaVA-1.5)	0.78	0.88	0.78	**0.99**	0.58	0.80	0.38	0.62
Image (GPT-4V)	Feature (CLIP)	**0.86**	0.92	**0.95**	**1.00**	**0.75**	**0.87**	0.57	0.74
Caption (BLIP-2 + GPT-4)	Feature (CLIP)	0.78	**0.96**	0.88	**0.99**	**0.75**	**0.86**	**0.61**	**0.80**

Conclusion

Key takeaway:

The paper introduces the new task of **Set Difference Captioning** and shows that combining **VLMs** and **LLMs** can automatically discover meaningful semantic differences between large image datasets.

Limitations:

- **Dependence on LLM**
- **Difficulty with Subtle Differences**

Next steps:

Instead of prompting an LLM, we analyze which **caption tokens** are most distinctive between two image sets.

M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

Carnegie Mellon University

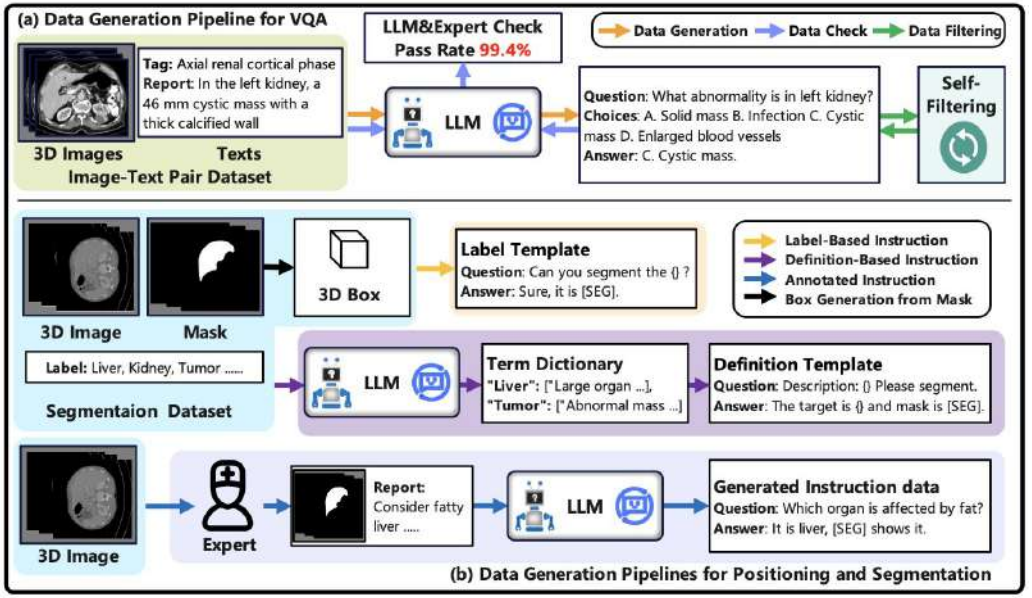
Background & Introduction

- **CT scans are inherently 3D:** hundreds of slices forming a full body volume
- **Prior AI analyzed 2D slices only:** slice-by-slice, one at a time - inefficient
- **Depth and spatial context completely lost:** no understanding of how regions relate
- **Tumor location relative to organs:** critical for surgery, invisible to 2D models
- **RadFM supported 3D but only text generation + poor performance + 13B parameters**
- **Costly slice-by-slice workarounds or models failed on 3D entirely**
- **No large-scale 3D training dataset existed publicly**
- **No benchmark to evaluate 3D medical AI fairly**

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

M3D-Data

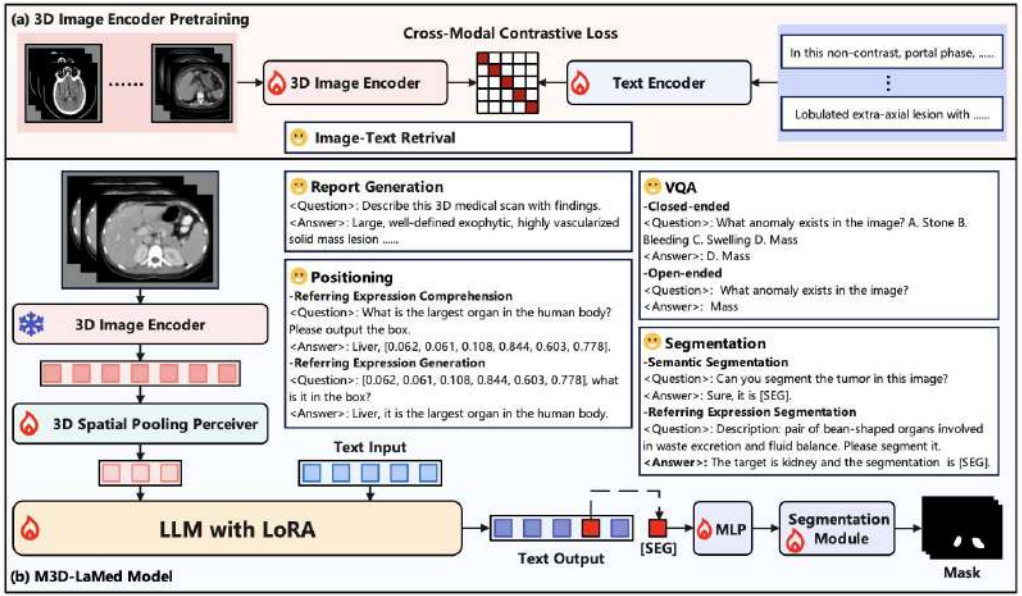
- **M3D-Cap:** 120K scans + reports pairs (from Radiopaedia)
- **M3D-VQA:** 509K QA pairs MCQ generated by Qwen-72B
- **M3D-Seg:** 5,772 segmented 3D CTs Organ/tumor masks from 25 public datasets
- **M3D-RefSeg:** 210 expert-annotated scans



Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

M3D-LaMed

- Pre-train a 3D ViT (MONAI): pair with BERT text encoder, feed 120K scan-report pairs, contrastive loss, encoder learns to understand 3D medical images
- Train the perceiver: freeze encoder and LLaMA, task is report generation, perceiver learns to bridge both
- Fine-tune everything: encoder fully, perceiver fully, LLaMA (LoRA)
- When LLaMA outputs [SEG] token, SegVol takes that vector and draws the precise 3D mask



Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

M3D-Bench

- First-ever comprehensive 3D medical AI benchmark
- 8 tasks across 5 categories: Image-text retrieval, report generation, VQA, positioning, and segmentation.
- Traditional metrics (BLEU, ROUGE, BERT-Score) measure word and phrase overlap.
- LLM-based scoring to evaluate meaning and content overlap on a 0-100 scale.
- Different measurement for each output type: Accuracy for multiple choice, text similarity for generated text, IOU for bounding boxes, Dice score for segmentation masks.
- Test set: 2,000 scan-report pairs from M3D-Cap for retrieval and report generation; 2,000 scans and 13,791 QA pairs for VQA
- Segmentation test set: 20% holdout from AbdomenCT-1K within M3D-Seg, plus ACT-1K for out-of-distribution testing
- Fully open benchmark: Any future 3D medical AI system can be evaluated against the same standard, enabling meaningful progress tracking across the field.

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

Results & Limitations

Results:

- VQA: 75.78% vs RadFM's 19.79%
- Retrieval R@1: 64% vs PMC-CLIP's 9% (100 samples)
- Report generation LLM score: 8.49 vs RadFM's 4.32
- First ever 3D language-guided segmentation

Limitations:

- Single-turn VQA only: no multi-step reasoning, no agentic behavior
- Chest CT only: organ scale, clean signal, pre-segmented volume
- Dataset collection bias: training data sourced from specific institutions
- Language bias: training reports predominantly in English

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

Key takeaways

- 3D ViT → LLM as a pattern for getting language out of volumetric biological data works
- Spatial structure in 3D must be preserved during compression, informs how we store and cluster embeddings in the spatial map
- Vision and language can be bridged with a small number of tokens: 256 visual tokens were enough
- Multi-task training generalizes better than single-task, informs our decision to build 4 query tools rather than one monolithic function

Bai, F., Du, Y., Huang, T., Meng, M. Q.-H., & Zhao, B. (2024, March 31). *M3D: Advancing 3D medical image analysis with Multi-Modal Large Language models*. arXiv.org. https://arxiv.org/abs/2404.00578

THANK YOU