Mohamed bin Zayed University of Artificial Intelligence

CVPR JUNE 3-7, 2026 DENVER COLORADO

MedMO : Grounding and Understanding Multimodal Large Language Models for Medical Images

Ankan Deria*, Komal Kumar*, Adinath Dukre, Eran Segal, Salman Khan, Imran Razzak

Mohamed bin Zayed University of Artificial Intelligence

Problem & Solution

Problem 1: Unreliability

Existing medical MLLMs produce uncertain or hallucinated outputs due to insufficient training on expert-curated clinical data. These models lack grounding in real clinical scenarios, leading to unreliable and potentially dangerous recommendations in medical settings.

Problem 2: Poor Localization

Current models struggle with precise localization of anatomical structures and disease regions in medical images.

Problem 3: Single Modality Limitation

Most models focus on single imaging types (e.g., only radiology or pathology) rather than unified cross-modal understanding. This narrow focus prevents holistic patient assessment and limits applicability across diverse clinical scenarios.

The Impact

These limitations create a critical gap: no unified, trustworthy medical AI system exists that combines accurate understanding with spatial grounding across multiple imaging modalities and clinical domains.

Our Solution

MedMO is the first unified medical foundation model trained on 26M+ multimodal samples from 45 diverse medical datasets spanning radiology, pathology, ophthalmology, and biological systems with comprehensive multi-task supervision.

Solution: Key Innovations

Comprehensive Multimodal Foundation Model

Open-source medical foundation model trained exclusively on large-scale, domain-specific clinical data for comprehensive multimodal understanding across all major medical imaging modalities.

Unprecedented Data Scale

26M+ multimodal samples curated from 45 diverse medical datasets spanning radiology, pathology, ophthalmology, and biological systems (respiratory, cardiovascular, nervous, digestive, urinary), ensuring robust cross-modal generalization.

Progressive Four-Stage Training

Systematically building clinical knowledge from general vision-language alignment to high-resolution anatomical grounding to specialized instruction tuning and finally semantic medical report generation for enhanced alignment.

Spatial Coordination with RL

Novel reinforcement learning approach using verifiable bounding-box rewards (Hungarian Matching + Geometric Metrics) enabling precise spatial localization of anatomical structures and disease regions.

State-of-the-Art Performance

Superior results across visual question answering, clinical reasoning, report generation, and disease localization—demonstrating the effectiveness of the integrated approach.

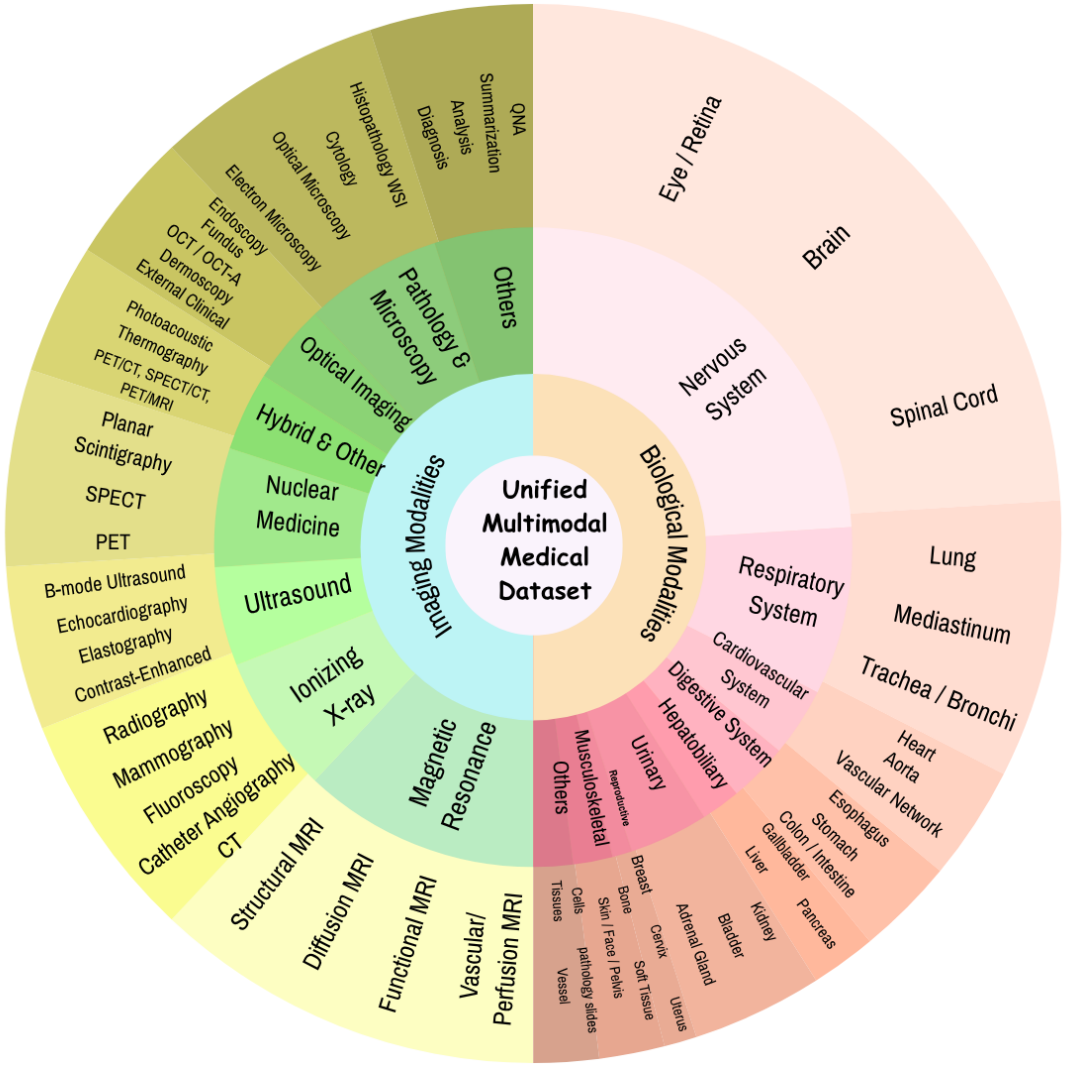
Multi-Task Supervision Across Clinical Domains

Visual QA • Text-Based QA • Radiology Report Generation • Disease Localization with Bounding Boxes • Anatomical Grounding • Clinical Reasoning • Diagnostic Classification • Spatial Object Detection

Dataset

- **Multi-task supervision covering diverse clinical applications:** Visual Question Answering (VQA), Text-based Medical QA, Radiology Report Generation, Disease Localization with Bounding Boxes, Anatomical Grounding, Clinical Reasoning, Diagnostic Classification, and Spatial Object Detection across radiology, pathology, ophthalmology, and emergency care scenarios.

- Comprehensive multimodal corpus spanning all major medical imaging modalities (X-ray, CT, MRI, Ultrasound, Nuclear Medicine, Optical, Pathology) and biological systems (Respiratory, Cardiovascular, Nervous, Digestive, Urinary, Musculoskeletal, etc.). 26M+ samples from 45 datasets ensure robust cross-modal generalization.



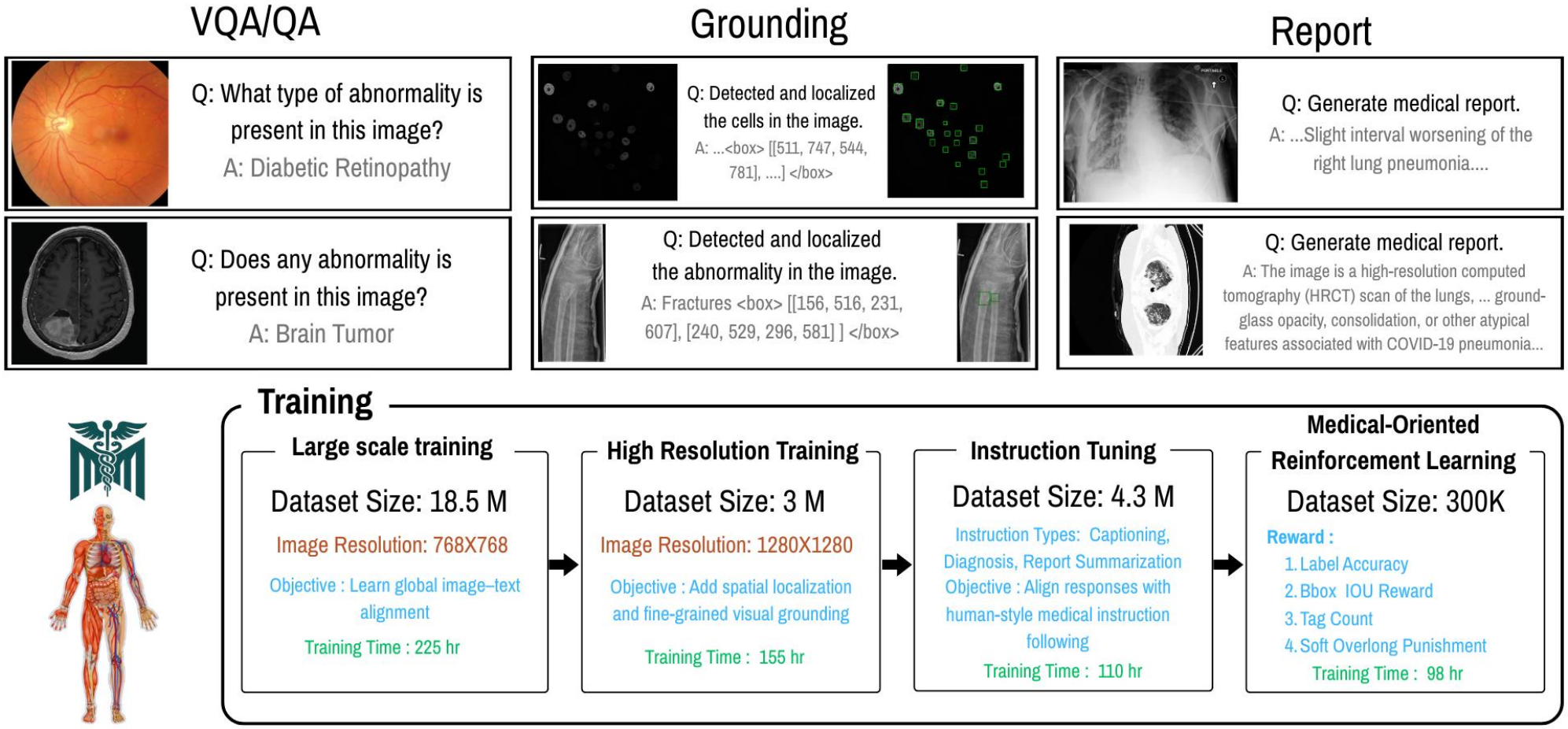
Covering Datasets & Modalities

Methodology

Base Architecture

Qwen3-VI-8B-Instruct & Qwen3-VI-4B-Instruct (Pre-trained Vision-Language Model)

Multi-Stage Progressive Training Pipeline



RL & Novel Bounding Box Reward Function

Hungarian Matching + Geometric Metrics :

- Normalized L1 Distance (resolution-invariant)
- Generalized IoU (GIoU) ∈ [-1, 1]
- Coverage-normalized scoring

$$L1_{ij} = \frac{|x_1^p - x_1^g| + |y_1^p - y_1^g| + |x_2^p - x_2^g| + |y_2^p - y_2^g|}{2\sqrt{H^2 + W^2}}$$

Cost Matrix: $$C_{ij} = w_{\text{L1}}^m \, L1_{ij} + w_{\text{G}}^m \, (1 - \text{GIoU}_{ij}) w_{\text{L1}}^m = 5, \; w_{\text{G}}^m = 2$$

Per-Match Quality Score: $$s_{ij} = \frac{w_{\text{L1}} \, (1 - \text{clip}_{[0,1]}(L1_{ij})) + w_{\text{G}} \left(\frac{\text{GIoU}_{ij} + 1}{2} \right)}{w_{\text{L1}} + w_{\text{G}}}$$

The reward is a coverage normalized sum with optional FP/FN penalties (Pen):

$$\text{B} = \frac{1}{G} \sum_{(i,j) \in M} s_{ij}, \; \text{Pen} = \frac{\lambda_{\text{FN}}(G - |M|) + \lambda_{\text{FP}}(P - |M|)}{\max(1, G)}$$

Final Reward: $$R_{\text{bbox}} = \text{clip}_{[0,1]} \left(\text{B} - \text{Pen} \right)$$

We adapted clip higher strategy for DAPO

Results

Visual Question Answering (VQA)

- 60.8% average accuracy — +21.3% over baseline, within 0.6% of SOTA Fleming-VL

Text-Based Medical QA

- 61.3% accuracy — +7.7% vs baseline, +15.6% vs Fleming-VL

Medical Report Generation

- CIDEr 140.0, Semb 50.0 -best semantic coherence and clinical accuracy.
- Produces grounded radiology reports with proper medical terminology.

Model	NIH	DeepLession	Bacteria	MedSG (multi_view)	MedSG (object_tracking)	MedSG (referring)	Avg.
InternVL3-8B	10.1	0.00	0.7	6.3	13.0	3.3	5.6
Fleming-VL-8B	0.00	0.00	8.3	42.0	36.7	16.6	17.2
Lingshu-7B	5.3	0.7	0.00	28.3	38.7	10.4	13.9
Qwen3VL-8B	**16.4**	0.00	9.16	8.4	17.8	31.4	13.8
MedSG-Bench	–	–	–	55.0	62.1	60.4	-
MedMO-8B	8.83	**38.5**	**54.6**	**75.8**	**77.2**	**70.1**	**54.2**

Models	VQA Benchmarks								Text QA Benchmarks							
	MMMU-Med	VQA-RAD (all)	SLAKE (all)	PathVQA (all)	PMC-VQA	OMVQA	MedXQA	Avg.	MMLU-Med	PubMedQA	MedMCQA	MedQA	Medbullets (op4/op5)	MedXQA	SGPQA	Avg.
Closed-source Models																
GPT-4.1	75.2	65.0	72.2	55.5	55.2	75.5	45.2	63.4	89.6	75.6	77.7	89.1	77.0	30.9	49.9	70.0
Claude Sonnet 4	74.6	67.6	70.6	54.2	54.4	65.5	43.3	61.5	91.3	78.6	79.3	92.1	80.2	33.6	56.3	73.1
Gemini-2.5-Flash	76.9	68.5	75.8	55.4	55.4	71.0	52.8	65.1	84.2	73.8	73.6	91.2	77.6	35.6	53.3	69.9
Open-source Models																
BiomedGPT	24.9	16.6	13.6	11.3	27.6	27.9	–	–	–	–	–	–	–	–	–	–
Med-R1-2B	34.8	39.0	54.5	15.3	47.4	–	21.1	–	51.5	66.2	39.1	39.9	33.6	11.2	17.9	37.0
MedVLM-R1-2B	35.2	48.6	56.0	32.5	47.6	77.7	20.4	45.4	51.8	66.4	39.7	42.3	33.8	11.8	19.1	37.8
MedGemma-4B-IT	43.7	72.5	76.4	48.8	49.9	69.8	22.3	54.8	66.7	72.2	52.2	56.2	45.6	12.8	21.6	46.8
LLaVA-Med-7B	29.3	53.7	48.0	38.8	30.5	44.3	20.3	37.8	50.6	26.4	39.4	42.0	34.4	9.9	16.1	31.3
HuatuoGPT-V-7B	47.3	67.0	67.8	48.0	53.3	74.2	21.6	54.2	69.3	72.8	51.2	52.9	40.9	10.1	21.9	45.6
BioMediX2-8B	39.8	49.2	57.7	37.0	43.5	63.3	21.8	44.6	68.6	75.2	52.9	58.9	45.9	13.4	25.2	48.6
Qwen2.5VL-7B	50.6	64.5	67.2	44.1	51.9	63.6	22.3	52.0	73.4	76.4	52.6	57.3	42.1	12.8	26.3	48.7
InternVL2.5-8B	53.5	59.4	69.0	42.1	51.3	81.3	21.7	54.0	74.2	76.4	52.4	53.7	42.4	11.6	26.1	48.1
InternVL3-8B	59.2	52.9	62.4	39.0	53.8	79.1	22.4	52.7	77.5	75.4	57.7	62.1	50.2/42.8	13.1	31.2	51.2
Lingshu-7B	54.0	43.0	33.2	41.9	54.2	82.9	**26.9**	48.0	69.6	75.8	56.3	63.5	**62.0/53.8**	16.4	27.5	53.1
Fleming-VL-8B	63.3	56.4	**80.0**	**56.5**	**64.3**	**88.2**	21.6	**61.4**	71.8	74.0	51.8	53.7	40.5/37.3	12.1	24.9	45.7
Qwen3VL-8B	61.4	31.2	15.0	14.6	52.3	77.2	24.8	39.5	79.0	70.4	60.0	66.1	56.1/47.7	15.1	34.7	53.6
MedMO-4B	54.6	35.0	30.0	42.4	50.6	79.7	24.8	45.3	75.7	**78.0**	58.0	78.5	57.5/47.7	16.4	29.4	55.1
MedMO-8B	**64.6**	**64.7**	70.0	56.3	59.4	84.8	26.2	60.8	**81.0**	77.6	**65.0**	**84.3**	**66.5/60.2**	**19.9**	**36.0**	**61.3**

Models	MIMIC-CXR				CheXpert Plus				IU-Xray				Med-Trinity			
	ROUGE-L	CIDEr	RaTE	Semb	ROUGE-L	CIDEr	RaTE	Semb	ROUGE-L	CIDEr	RaTE	Semb	ROUGE-L	CIDEr	RaTE	Semb
Closed-source Models																
GPT-4.1	9.0	82.8	51.3	23.9	24.5	78.8	45.5	23.2	30.2	124.6	51.3	47.5	–	–	–	–
Claude Sonnet 4	20.0	56.6	45.6	19.7	22.0	59.5	43.5	18.9	25.4	88.3	55.4	41.0	–	–	–	–
Gemini-2.5-Flash	25.4	80.7	50.3	29.7	23.6	72.2	44.3	27.4	33.5	129.3	55.6	50.9	–	–	–	–
Open-source Models																
Med-R1-2B	19.3	35.4	40.6	14.8	18.6	37.1	38.5	17.8	16.1	38.3	41.4	12.5	–	–	–	–
MedVLM-R1-2B	20.3	40.1	41.6	14.2	20.9	43.5	38.9	15.5	22.7	61.1	46.1	22.7	–	–	–	–
MedGemma-4B-IT	25.6	81.0	52.4	29.2	**27.1**	79.0	47.2	29.3	30.8	103.6	57.0	46.8	–	–	–	–
LLaVA-Med-7B	15.0	43.4	12.8	18.3	18.4	45.5	38.8	23.5	18.8	68.2	40.9	16.0	–	–	–	–
HuatuoGPT-V-7B	23.4	69.5	48.9	20.0	21.3	64.7	44.2	19.3	29.6	104.3	52.9	40.7	–	–	–	–
BioMediX2-8B	20.0	52.8	44.4	17.7	18.1	47.9	40.8	21.6	19.6	58.8	40.1	11.6	–	–	–	–
Qwen2.5VL-7B	24.1	63.7	47.0	18.4	22.2	62.0	41.0	17.2	26.5	78.1	48.4	36.3	23.5	81.5	44.9	38.3
InternVL2.5-8B	23.2	61.8	47.0	21.0	20.6	58.5	43.1	19.7	24.8	75.4	51.1	36.7	13.5	47.1	42.5	12.8
InternVL3-8B	22.9	66.2	48.2	21.5	20.9	65.4	44.3	25.2	22.9	76.2	51.2	31.3	12.9	46.6	42.2	3.7
Lingshu-7B	30.8	109.4	52.1	30.0	26.5	79.0	45.4	26.8	41.2	180.7	57.6	48.4	16.0	74.5	44.4	24.0
Fleming-VL-8B	**35.7**	132.5	56.7	33.6	26.1	82.2	47.1	40.1	**44.9**	**198.6**	**66.0**	**51.3**	13.1	35.8	41.9	18.1
Qwen3VL-8B	25.1	77.9	50.3	33.4	21.9	67.4	44.4	37.9	25.0	91.44	52.5	42.9	20.2	69.9	45.9	33.6
MedMO-4B	26.0	92.6	49.8	31.6	15.1	62.3	36.6	34.2	26.6	94.0	42.1	41.3	22.5	152.6	47.8	34.3
MedMO-8B	31.7	**140.0**	**57.1**	**50.0**	23.6	**87.5**	**47.3**	**42.2**	31.1	169.7	45.3	41.3	**37.0**	**270.4**	**53.0**	**39.2**

Conclusion & Impact

Key Achievements & Impact

- ✓ Open-source unified foundation model combining multimodal understanding with verifiable spatial grounding across 45 datasets
- ✓ State-of-the-art performance across VQA (62.6%), Text QA (61.5%), Report Generation (CIDEr: 140.0), and Grounding (54.2% IoU)
- ✓ 3× superior disease localization compared to prior medical MLLMs through novel geometric reward learning
- ✓ Scalable, open-source release (MedMO-8B/4B) establishing reliable foundation for clinical AI development and deployment

Check out MedMO-Next: Leading medical foundation model with strong bounding-box grounding