MedMO: Grounding and Understanding Multimodal Large Language Model for Medical Images (Supplementary)

Ankan Deria*, Komal Kumar*, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, Imran Razzak
Mohamed bin Zayed University of Artificial Intelligence
{ankan.deria, komal.kumar}@mbzuai.ac.ae

Models: huggingface.co/collections/MBZUAI/medmo **GitHub:** github.com/genmilab/MedMO

Project Page: genmilab.github.io/MedMO-Page

1. Reward function details

1.1. Bounding Box Reward Function

For grounding tasks in the reinforcement learning stage, we employ a specialized reward function that evaluates the quality of predicted bounding boxes against ground truth annotations. This reward is computed using Hungarian matching combined with geometric metrics.

Notation and Setup. Given ground truth boxes $\mathcal{G} = \{g_j\}_{j=1}^{G}$ and predicted boxes $\mathcal{P} = \{p_i\}_{i=1}^{P}$ in XYXY format (i.e., (x_1, y_1, x_2, y_2) coordinates), we first determine the image dimensions (H, W) from the maximum extents of ground truth boxes if available, otherwise from predictions (with fallback to $(1, 1)$ if both are empty).

Pairwise Metrics. For each pair of boxes (p_i, g_j), we compute two geometric measures:

Normalized L1 Distance: The L1 distance over all four coordinates, normalized by the image perimeter:

$$L1_{ij} = \frac{|x_1^p - x_1^g| + |y_1^p - y_1^g| + |x_2^p - x_2^g| + |y_2^p - y_2^g|}{2\sqrt{H^2 + W^2}} \tag{1}$$

Generalized IoU (GIoU): We compute $\text{GIoU}_{ij} \in [-1, 1]$ following Rezatofighi et al. [34], which extends standard IoU to account for non-overlapping boxes.

Hungarian Matching. To establish optimal correspondence between predictions and ground truth, we construct a cost matrix:

$$C_{ij} = w_{\text{L1}}^m \cdot L1_{ij} + w_{\text{G}}^m \cdot (1 - \text{GIoU}_{ij}), \tag{2}$$

where $w_{\text{L1}}^m = 5.0$ and $w_{\text{G}}^m = 2.0$ are matching cost weights. We apply the Hungarian algorithm to find the minimum-cost bipartite matching, yielding $m = \min(P, G)$ matched pairs $\{(i_k, j_k)\}_{k=1}^{m}$.

Per-Match Score. For each matched pair (i_k, j_k), we compute a quality score by:

1. Mapping GIoU to $[0, 1]$: $\tilde{G}_k = \frac{\text{GIoU}_{i_k j_k} + 1}{2}$
2. Clamping L1 to $[0, 1]$: $\hat{L1}_k = \text{clip}_{[0,1]}(L1_{i_k j_k})$
3. Computing weighted blend:

$$s_k = \frac{w_{\text{L1}} \cdot (1 - \hat{L1}_k) + w_{\text{G}} \cdot \tilde{G}_k}{w_{\text{L1}} + w_{\text{G}}}, \quad s_k \in [0, 1] \tag{3}$$

where $w_{\text{L1}} = 5.0$ and $w_{\text{G}} = 2.0$ are pair score weights.

Final Reward Computation. The base reward is the coverage-normalized sum of matched pair scores:

$$\text{base} = \frac{1}{G} \sum_{k=1}^{m} s_k \tag{4}$$

We optionally apply penalties for false positives (FP) and false negatives (FN):

$$\text{penalty} = \frac{\lambda_{\text{FN}} \cdot (G - m) + \lambda_{\text{FP}} \cdot (P - m)}{\max(1, G)}, \tag{5}$$

where λ_{FN} and λ_{FP} are penalty coefficients (default: 0). The final bounding box reward is:

$$\boxed{R_{\text{bbox}} = \text{clip}_{[0,1]}\left(\text{base} - \text{penalty}\right)} \tag{6}$$

Expanding the base term:

$$\text{base} = \frac{1}{G} \sum_{k=1}^{m} \frac{w_{\text{L1}}(1 - L1_{i_k j_k}) + w_{\text{G}} \cdot \frac{\text{GIoU}_{i_k j_k} + 1}{2}}{w_{\text{L1}} + w_{\text{G}}} \tag{7}$$

Edge Cases. The reward function handles special cases as follows:

- **No ground truth boxes** ($G = 0$): $R_{\text{bbox}} = 0.5$ (neutral reward)
- **Ground truth present but no predictions** ($G > 0, P = 0$): $R_{\text{bbox}} = \text{clip}_{[0,1]}(0 - \text{penalty})$, which equals 0.0 with default penalties
- **Failed matching** (no feasible pairs): Treated as $m = 0$, where all ground truth boxes are unmatched and all predictions are false positives

This reward formulation encourages the model to produce accurate bounding box predictions through Hungarian-matched optimization of both localization (L1) and overlap quality (GIoU), while penalizing missing detections and spurious predictions.

2. Experimental Details

We conducted all experiments using the SFT_Trainer and RL (GRPO) trainer frameworks. Unless otherwise noted, we used mixed-precision training (dtype=bfloat16) on a cluster of $64\times$ AMD Instinct MI210 GPUs. Random seeds, optimizer state, and scheduler configuration were logged for full reproducibility.

2.1. Stage 1: General SFT

Parameters Details

We provide detailed experimental settings in Table 1, which we apply exclusively to training stage 1 MedMO.

Parameter	Value
Batch size	10
Gradient accumulation steps	2
Learning rate (initial)	1×10^{-5}
LR scheduler	Cosine decay
Number of epochs	1
Image resolution	768×768 pixels
dtype	bfloat16

Table 1. Training parameter details for stage 1.

Training Dynamics

During Stage 1, optimization converges rapidly: the loss drops from ~ 11 to < 0.3 within the first ≈ 10 steps, and entropy collapses from ~ 5.3 to ~ 0.1 over the same window, indicating quickly sharpened token distributions. Mean token accuracy rises steeply from ~ 0.6 to ~ 0.95 by step ≈ 10 and then plateaus with minor oscillations thereafter. These curves reflect stable optimization under the cosine schedule, fast fit to the instruction format, and no signs of late-stage instability during the single-epoch SFT. *Unless noted, one plotted "step" corresponds to an aggregate over 100 mini-batches (logging interval = 100 batches).*

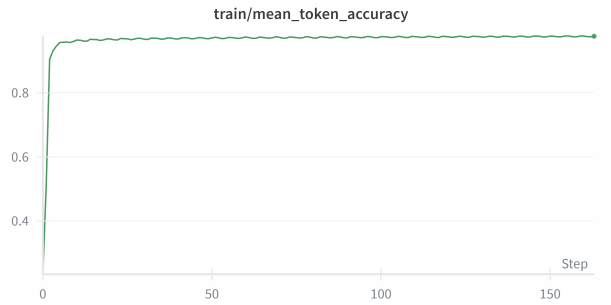


Figure 1. Stage 1 mean token accuracy vs. step (each step = 100 mini-batches). Accuracy jumps to $\sim$0.95 within $\approx$10 steps and remains stable.

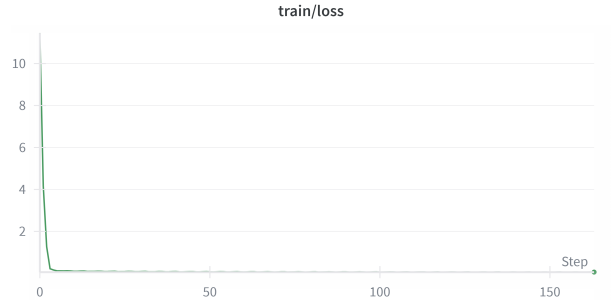


Figure 2. Stage 1 training loss vs. step (each step = 100 mini-batches). Loss declines from $\sim$11 to < 0.3 in the first $\approx$10 steps, then flattens.

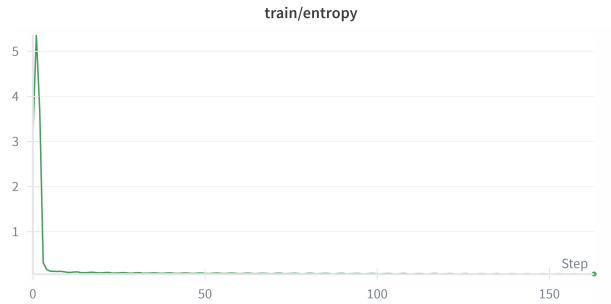


Figure 3. Stage 1 output entropy vs. step (each step = 100 mini-batches). Entropy collapses from $\sim$5.3 to $\sim$0.1 by $\approx$10 steps, indicating confident token distributions.

2.2. Stage 2: High-Resolution Image SFT

Parameters Details

We provide detailed experimental settings in Table 2, which we apply exclusively to training stage 2 MedMO.

Parameter	Value
Batch size	2
Gradient accumulation steps	8
Learning rate (initial)	8×10^{-6}
LR scheduler	Cosine decay
Number of epochs	1
Image resolution	1280×1280 pixels
dtype	bfloat16

Table 2. Training parameter details for stage 2.

Training Dynamics

During Stage 2, we fine-tuned MedMO on high-resolution (1280×1280) medical images using a combination of VQA, grounding, and report-generation datasets. Each logged step corresponds to 100 training batches. As illustrated in Figures 4–6, the model exhibits rapid convergence and stable learning behavior. Mean token accuracy (Fig. 4) increases sharply from ∼0.86 to ∼0.95 within the first few hundred steps, indicating strong adaptation to high-resolution visual–textual data. Training loss (Fig. 5) decreases quickly from ∼0.9 to ∼0.3 and then plateaus, confirming smooth optimization without overfitting. Entropy (Fig. 6) drops from ∼0.65 to ∼0.27 and remains steady, showing reduced uncertainty and confident token predictions. These results confirm that Stage 2 effectively enhances MedMO's multimodal alignment and high-resolution spatial reasoning.

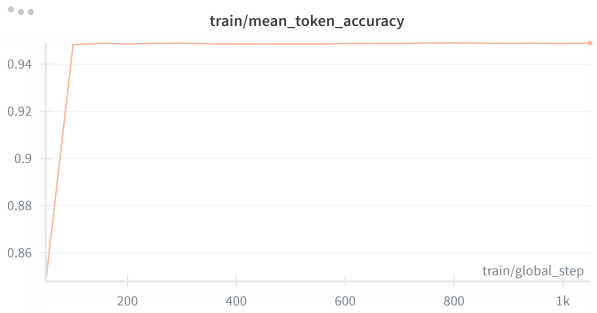


Figure 4. Stage 2 mean token accuracy vs. global step (each step = 100 mini-batches). Accuracy improves rapidly from ∼0.86 to ∼0.95, showing strong convergence and model stability.

Datasets Used

For Stage 2, we employed datasets emphasizing multimodal reasoning, high-quality medical captions, and spatial grounding. The training corpus included a diverse mix of **VQA-oriented datasets** such as *VQA-Med-2019*, *PubMed-Vision*, *NIH-VQA*, *Quilt-LLaVA-Pretrain*, *MIMIC-Ext-MIMIC-CXR-VQA*, *VQA-RAD*, *PathVQA*, *PMC-VQA*, *SLAKE*, and *CT-RATE*. We also incorporated **report-generation datasets** including *IU-Xray*, *MIMIC-CXR*, *CheXpert*, *CheXpert Plus*, *MEDPIX-ClinQA*, *ROCO*, *ROCO-V2*, and *FairVLMed* to enhance radiology-style narrative generation and image–text consistency. Finally, for **grounding and bounding-box prediction**, we used *NIH Chest X-ray*, *DeepLesion*, *GRAZPEDWRI-DX*, *SLAKE*, *Cell Microscopy (DeepCell, Bacteria, and CTC)*, and *MedSG*, which provide localized annotations for spatial reasoning and fine-grained object detection.

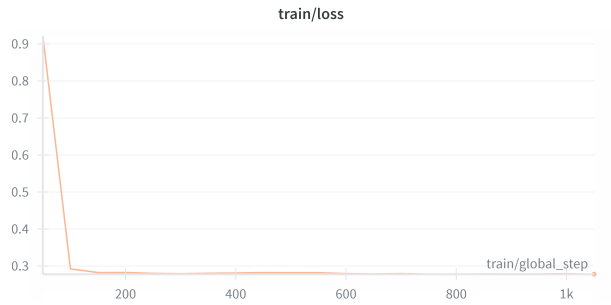


Figure 5. Stage 2 training loss vs. global step (each step = 100 mini-batches). Loss decreases from ∼0.9 to ∼0.3, confirming efficient optimization and stable convergence.

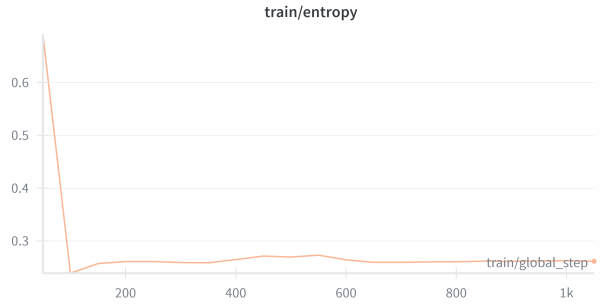


Figure 6. Stage 2 output entropy vs. global step (each step = 100 mini-batches). Entropy declines from ∼0.65 to ∼0.27, reflecting reduced uncertainty and higher confidence in predictions.

This combination allows MedMO to improve fine-grained visual grounding and detailed report synthesis under high-resolution supervision.

2.3. Stage 3: Instruction Tuning

Parameters Details

We provide detailed experimental settings in Table 3, which we apply exclusively to training stage 3 MedMO.

Training Dynamics

Stage 3 focuses on instruction tuning to enhance MedMO's clinical reasoning, comprehension, and text generation capabilities. Each step shown in the plots corresponds to 100 mini-batches. As shown in Figures 7–9, the model ex-

Parameter	Value
Batch size	14
Gradient accumulation steps	2
Learning rate (initial)	5×10^{-6}
LR scheduler	Cosine decay
Number of epochs	1
dtype	bfloat16

Table 3. Training parameter details for stage 2.

hibits smooth and stable convergence. Mean token accuracy (Fig. 7) rises steadily from ∼0.62 to ∼0.69, demonstrating improved instruction-following and cross-modal reasoning. Training loss (Fig. 8) decreases from ∼1.7 to ∼1.4 within the first few steps, while entropy (Fig. 9) declines from ∼1.55 to ∼1.38, both indicating effective optimization and improved confidence. Overall, Stage 3 consolidates multimodal understanding and instruction-following capabilities with stable convergence and balanced learning dynamics.

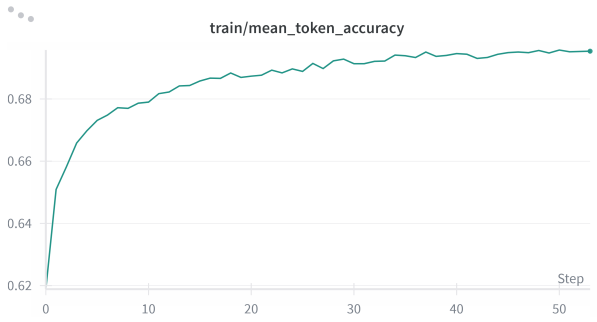


Figure 7. Stage 3 mean token accuracy vs. step (each step = 100 mini-batches). Accuracy increases gradually from ∼0.62 to ∼0.69, indicating improved instruction-following and reasoning.

Figure 8. Stage 3 training loss vs. step (each step = 100 mini-batches). Loss decreases from ∼1.7 to ∼1.4, showing smooth convergence and stable optimization.

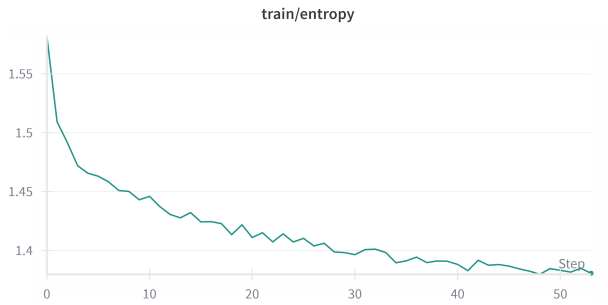


Figure 9. Stage 3 output entropy vs. step (each step = 100 mini-batches). Entropy decreases from ∼1.55 to ∼1.38, reflecting higher model confidence and stable prediction behavior.

Datasets Used

For Stage 3, we utilized datasets centered on medical instruction-following, comprehension, reasoning, and report summarization. The training corpus integrated a broad collection of **QA and understanding datasets**, including *MedQA*, *PubMedQA*, *PMC-OA*, *MedMCQA*, *PMC-InstructQA*, *MedQuAD*, *Medical-Meadow-MedQA*, *ChatDoctor-HealthCareMagic-100k*, *AlpaCare-MedInstruct-52k*, *ChatDoctor-iCliniq*, *MedReason*, *MIMIC-IV-Ext-BHC*, *Medical-R1-Distill-Data*, *medical-o1-reasoning-SFT*, *Meadow-PubMed-Causal*, *Meadow-Medical-Flashcards*, *Meadow-MediQA*, and *Meadow-Wikidoc*. These datasets collectively provide diverse factual, reasoning, and instruction-based supervision across medical, clinical, and biomedical contexts.

In addition, we incorporated **summarization and clinical reporting datasets** such as *Medical-Meadow-Cord19*, and *mimic-ext-bhc*. These datasets focus on long-form radiology and biomedical report synthesis, improving contextual understanding, summarization, and domain-specific narrative generation.

Together, this combined corpus strengthens MedMO's instruction-tuned reasoning, factual grounding, and text–image comprehension, enabling robust performance across diverse medical instruction and report-generation scenarios.

2.4. Stage 4: Reinforcement Learning (Spatial Grounding)

Parameters Details

- Reward functions: Label accuracy, bounding-box IoU (Δ), tag count, and soft-overlong-punishment.
- Image resolution: dynamic (no fixed resize or bounding-box rescaling).
- Epsilon (policy perturbation) = 0.15.
- Epsilon_high (upper bound) = 0.25.
- Number of training epochs = 2.

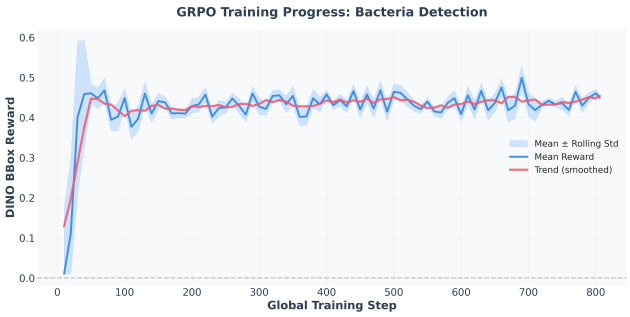


Figure 10. **DAPO training progress for bounding-box detection.** Mean bounding-box reward (blue) with ± rolling standard deviation (shaded) and smoothed trend (red). The consistent upward trajectory indicates effective policy optimization and stable improvement in spatial localization accuracy.

- Number of batch size = 2.
- Gradient accumulation steps = 4.
- Number of generations per prompt = 8.
- Maximum prompt length = 2048 tokens.
- Maximum completion length = 1024 tokens.

Implementation & Reproducibility Notes

- Optimizer: AdamW with default betas (0.9, 0.999) and weight decay = 0.1.
- Warm-up steps = 10% of total training steps per stage.
- Seed: All runs initialized with a fixed seed (e.g., 42) per stage; randomness only arises from data shuffling and augmentations.

Training Dynamics

During Stage 4, MedMO was trained with reinforcement learning using the DAPO [50] algorithm to refine its spatial grounding and bounding-box localization capabilities. Each global step aggregates multiple rollouts sampled per instruction prompt. As shown in Figure 10, the bounding-box reward rises sharply from nearly zero to ∼0.45 within the first 100 steps, indicating rapid adaptation of the policy to spatial localization signals. Beyond this point, the mean reward curve (blue) stabilizes around 0.42–0.45 with moderate oscillations, while the smoothed trend (red) shows a consistent upward trajectory, reflecting incremental performance gains and robust reward optimization. The steady variance band (rolling standard deviation) demonstrates that exploration remains controlled throughout training, preventing reward collapse or policy drift. Overall, the DAPO stage successfully enhances the model's spatial precision and stability in bounding-box generation tasks such as bacteria and lesion detection.

Datasets Used

For Stage 4, we utilized datasets providing explicit spatial supervision and precise bounding-box annotations for medical object detection and grounding tasks. These include *NIH Chest X-ray*, *DeepLesion*, *Bacteria Segmentation*, *CTC (Cell Tracking Challenge)*, *SLAKE*, *GRAZPEDWRI-DX*, and *MedSG*, which collectively cover anatomical structures, lesions, and microscopic cellular regions. The DAPO objective leverages bounding-box IoU and label-accuracy rewards derived from these datasets to iteratively refine spatial alignment and improve localization precision. This stage significantly enhances MedMO's visual grounding ability, leading to robust disease localization and fine-grained spatial reasoning across diverse medical modalities.

3. Dataset Collection

We curated a unified multimodal corpus comprising **45 datasets** spanning radiology, pathology, ophthalmology, dermatology, and surgical imaging, totaling more than **26M samples**. At the core lies the **MedTrinity** dataset [45], which contributes **18.5M** publicly available instruction-following pairs. This large-scale collection integrates both image–text and text-only medical data, enabling tasks such as captioning, visual question answering (VQA), clinical reasoning, and visual grounding.

The model was trained through four progressive stages. In **Stage 1**, we used the MedTrinity dataset to establish foundational multimodal understanding across diverse imaging modalities. **Stage 2** incorporated additional VQA, grounding, and captioning datasets, and trained the model with high-resolution medical images to enhance visual reasoning and fine-grained spatial grounding.

Stage 3 focused on medical text-only instruction data to strengthen clinical knowledge and language understanding. Finally, **Stage 4** employed reinforcement learning with bounding-box supervision to further refine localization and grounding capabilities.

The datasets encompass a broad spectrum of imaging modalities (X-ray, CT, MRI, ultrasound, optical, and nuclear imaging) and biological systems (chest, brain, heart, liver, kidney, eye, colon, and tissue), ensuring comprehensive anatomical and modality coverage. For grounding supervision, we incorporated datasets containing bounding-box annotations, including *NIH Chest X-ray*, *DeepLesion*, *Bacteria*, *Wrist X-ray (boneanomaly, fracture etc.)*, *CT*, and *Cell Microscopy (DeepCell)*. This diverse corpus collectively supports robust multimodal alignment, spatial reasoning, and medical instruction tuning.

Table 4 summarizes the datasets used in MedMO's training pipeline, grouped according to their primary role in each stage.

Note. Several other publicly available datasets such as

TCGA [40], *VALSET* [39], *MAMA-MIA* [13], *LLD-MMRI* [26], *CPD* [42], *CISC* [12], *CT-RATE* [14], *KIPA22* [48], and *PTCGA* [20] are already included in MedTrinity and were not trained on separately.

4. Quantitative Results

4.0.1. SOTA Comparison of MedMO for QA

Table 5 summarizes MedMO's performance across medical VQA and Text QA benchmarks for all four variants: MedMO-4B, MedMO-4B-Next, MedMO-8B, and MedMO-8B-Next.

VQA Benchmarks. MedMO-8B-Next achieves the highest VQA average of **72.7%**, outperforming all open-source competitors including Fleming-VL-8B (66.1%) and Lingshu-7B (55.1%) by **+6.6%** and **+17.6%**, respectively. It sets new state-of-the-art scores on MMMU-Med (**69.3%**), VQA-RAD (**86.4/68.0**), SLAKE (**83.0/81.6**), and OMVQA (**93.3%**). **MedMO-4B-Next** also surpasses Fleming-VL-8B with a VQA average of **68.5%**, achieving competitive scores on PMC-VQA (**75.7%**) and OMVQA (**90.6%**) despite its smaller scale. The base variants MedMO-4B (45.4%) and MedMO-8B (63.2%) show consistent improvement with scale, with MedMO-8B notably achieving the second-best PathVQA score (56.3%).

Text QA Benchmarks. MedMO-8B-Next achieves a Text QA average of **60.1%**, outperforming Fleming-VL-8B (45.7%) by **+14.4%**. It leads on MMLU-Med (**80.2%**), MedQA (**83.8%**), and MedXpertQA (20.9%), demonstrating strong clinical reasoning and knowledge integration. **MedMO-8B** achieves the highest QA average among all models including *Next* variants at **61.3%**, leading on MedMCQA (**65.0%**), MedQA (**84.3%**), and Medbullets (**66.5/60.2**), suggesting its base instruction tuning yields strong reasoning without RL fine-tuning overhead. **MedMO-4B-Next** achieves a QA average of **55.0%**, surpassing Fleming-VL-8B (45.7%) by **+9.3%** and even matching or exceeding Lingshu-7B (53.1%) on several benchmarks including PubMedQA (**78.2%**). Overall, all MedMO variants consistently outperform same-scale open-source models, with larger and *Next* variants delivering substantial improvements across both VQA and QA tasks.

4.1. SOTA Comparison of MedMO for Report Generation

Table 6 evaluates medical report generation across four datasets using semantic (ROUGE-L, CIDEr) and model-based (RaTE, Semb) metrics.

MIMIC-CXR. MedMO-8B-Next achieves the highest CIDEr of **143.4** and strong RaTE (**57.7%**) and Semb (**51.5%**), outperforming Fleming-VL-8B (132.5, 56.7%, 33.6%) on all metrics except ROUGE-L, where Fleming leads (35.7% vs. 32.6%). **MedMO-8B** achieves the second-best CIDEr (140.0) with the highest Semb among all models (50.0%), confirming that MedMO generates reports with superior semantic fidelity and clinical coherence. **MedMO-4B-Next** (CIDEr: 96.7, Semb: 34.3%) and **MedMO-4B** (CIDEr: 92.6, Semb: 31.6%) also outperform most open-source baselines despite their smaller scale.

CheXpert Plus. MedMO-8B-Next achieves the highest CIDEr (**88.3**) and RaTE (**48.1%**) and Semb (**43.8%**), surpassing Fleming-VL-8B (82.2, 47.1%, 40.1%) across all model-based metrics. **MedMO-8B** achieves the second-best CIDEr (87.5) and Semb (42.2%). While MedGemma-4B-IT leads on ROUGE-L (27.1% vs. 25.7%), MedMO's superior CIDEr and Semb scores indicate better semantic coherence and clinical accuracy over lexical overlap.

IU-Xray. Fleming-VL-8B leads on IU-Xray with CIDEr 198.6, RaTE 66.0%, and Semb 51.3%. **MedMO-8B-Next** achieves competitive performance (CIDEr: 171.9, RaTE: 56.0%, Semb: 43.1%), and **MedMO-8B** ranks second on ROUGE-L (37.0%) and CIDEr (169.7%). **MedMO-4B-Next** shows a strong improvement over the base 4B variant, achieving CIDEr 147.8 and Semb 49.4%, while Lingshu-7B leads on ROUGE-L (41.2%) among open-source models.

Med-Trinity. On Med-Trinity, which spans diverse modalities including CT, MRI, ultrasound, and pathology, **MedMO-8B-Next** achieves the highest ROUGE-L (**38.5%**) and CIDEr (**272.1**), while **MedMO-8B** leads on RaTE (**53.0%**) and Semb (39.2%). Both variants dramatically outperform all baselines — MedMO-8B-Next's CIDEr of 272.1 is over **3**× higher than the next best open-source model, Qwen2.5VL-7B (81.5), underscoring MedMO's exceptional capability in multi-modal medical report generation. **MedMO-4B-Next** also delivers strong performance (CIDEr: 183.8), surpassing all non-MedMO baselines.

5. Qualitative Results

To complement the quantitative analyses presented in the main text, Figures 11–14 provide qualitative insights into our method's performance across diverse medical imaging scenarios. These visualizations illustrate representative predictions, highlighting both successful cases and challenging examples under varied clinical conditions.

6. Overall Training Summary

Across the four stages, MedMO progressively improves from general multimodal alignment (Stage 1) to high-resolution spatial reasoning and grounding (Stage 2), instruction-tuned language understanding (Stage 3), and reinforcement-driven grounding refinement (Stage 4). Together, these stages establish a robust, domain-aware foundation model for diverse medical imaging tasks.

Table 4. Overview of datasets used in **MedMO** training. Datasets are grouped by category, each contributing to distinct training objectives such as image captioning, multimodal and text-based instruction tuning, and spatial grounding.

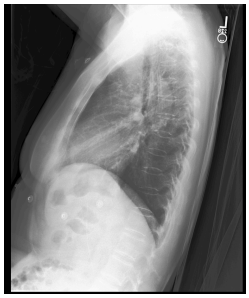
Category	Datasets	Purpose / Usage
Medical Caption Data	*MedTrinity* [45], *IU-Xray* [10], *MIMIC-CXR* [19], *CheXpert* [16], *CheXpert Plus* [7], *MEDPIX-ClinQA* [38], *ROCO* [32], *ROCO-V2* [35], *FairVLMed* [27]	Used for large-scale image–text alignment, caption-based supervision, and radiology-style report modeling across diverse imaging modalities.
Medical Multimodal Instruction Data	*VQA-Med-2019* [3], *PubMed-Vision* [9], *NIH-VQA* [36], *Quilt-LLaVA-Pretrain* [37], *MIMIC-Ext-MIMIC-CXR-VQA* [4], *VQA-RAD* [21], *PathVQA* [15], *PMC-VQA* [53], *SLAKE* [25], *CT-RATE* [6]	Facilitates multimodal instruction tuning for VQA, diagnosis, reasoning, and clinical summarization, improving image–text comprehension and task-driven responses.
Medical Text Instruction Data	*MedQA* [47], *PubMedQA* [18], *PMC-OA* [24], *MedMCQA* [31], *PMC-InstructQA* [53], *MedQuAD* [5], *Medical-Meadow-MedQA* [17], *ChatDoctor-HealthCareMagic-100k* [22], *AlpaCare-MedInstruct-52k* [52], *ChatDoctor-iCliniq* [23], *MedReason* [44], *MIMIC-IV-Ext-BHC* [1], *Medical-R1-Distill-Data* [8], *medical-o1-reasoning-SFT* [8], *Meadow-PubMed-Causal* [49], *Meadow-Medical-Flashcards* [28], *Meadow-MediQA* [33], *Meadow-Wikidoc* [29], *Medical-Meadow-Cord19* [43], *mimic-ext-bhc* [2]	Provides text-only instruction and QA supervision to enhance factual reasoning, language understanding, and medical knowledge grounding across clinical and biomedical contexts.
Medical Bounding Box Data	*NIH Chest X-ray* [11], *DeepLesion* [46], *GRAZPEDWRI-DX* [30], *SLAKE* [25], *Cell Microscopy (DeepCell, Bacteria, CTC)* [41], *MedSG* [51]	Provides explicit spatial grounding and disease-localization supervision with bounding-box annotations, enabling fine-grained object detection and improved spatial reasoning in medical imagery.

Table 5. Performance comparison across medical **VQA** and **Text QA** benchmarks. **Bold** and underline indicate the best and second-best results, respectively. OMIVQA and MedXQA refer to the OmniMedVQA and MedXpertQA benchmarks.

	VQA Benchmarks								Text QA Benchmarks							
Models	**MMMU-Med**	**VQA-RAD (closed/all)**	**SLAKE (closed/all)**	**PathVQA (all)**	**PMC-VQA**	**OMVQA**	**MedXQA**	**Avg.**	**MMLU-Med**	**PubMedQA**	**MedMCQA**	**MedQA**	**Medbullets (op4/op5)**	**MedXQA**	**SGPQA**	**Avg.**
Fleming-VL-8B	63.3	78.4/56.4	86.9/80.0	56.5	64.3	88.2	21.6	66.1	71.8	74.0	51.8	53.7	40.5/37.3	12.1	24.9	45.7
Qwen3VL-8B	61.4	54.1/31.2	34.3/15.0	14.6	52.3	77.2	24.8	40.5	79.0	70.4	60.0	66.1	56.1/47.7	15.1	34.7	53.6
MedMO-4B	54.6	50.9/35.0	41.0/30.0	42.4	50.6	79.7	24.8	45.4	75.7	78.0	58.0	78.5	57.5/47.7	16.4	29.4	55.1
MedMO-4B-Next	58.7	79.7/59.6	78.0/74.0	**73.3**	**75.7**	90.6	27.0	68.5	74.8	**78.2**	58.1	78.3	57.4/47.6	16.5	29.5	55.0
MedMO-8B	64.6	72.3/64.7	70.6/70.0	56.3	59.4	84.8	26.2	63.2	**81.0**	77.6	**65.0**	**84.3**	**66.5/60.2**	19.9	**36.0**	**61.3**
MedMO-8B-Next	**69.3**	**86.4/68.0**	**83.0/81.6**	56.3	74.1	**93.3**	**42.9**	**72.7**	80.2	75.6	62.0	83.8	65.2/57.8	**20.9**	35.5	60.1

Table 6. Comparison of medical report generation performance on MIMIC-CXR, CheXpert Plus, IU-Xray, and Med-Trinity using semantic (ROUGE-L, CIDEr) and model-based (RaTE, Semb) metrics. Models highlighted in green denote our proposed MedMO, which achieves the best overall performance across all datasets.

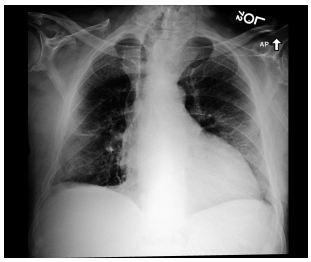
	MIMIC-CXR				CheXpert Plus				IU-Xray				Med-Trinity			
Models	**ROUGE-L**	**CIDEr**	**RaTE**	**Semb**	**ROUGE-L**	**CIDEr**	**RaTE**	**Semb**	**ROUGE-L**	**CIDEr**	**RaTE**	**Semb**	**ROUGE-L**	**CIDEr**	**RaTE**	**Semb**
Fleming-VL-8B	**35.7**	132.5	56.7	33.6	26.1	82.2	47.1	40.1	**44.9**	**198.6**	**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-4B-Next	28.3	96.7	52.0	34.3	23.5	74.5	42.6	38.7	38.0	147.8	62.0	49.4	26.3	183.8	49.5	38.6
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
MedMO-8B-Next	32.6	**143.4**	**57.7**	**51.5**	25.7	**88.3**	**48.1**	**43.8**	31.8	171.9	56.0	43.1	**38.5**	**272.1**	**53.8**	**40.7**



Ground Truth : The heart is normal in size. The mediastinal and hilar contours are unremarkable. There is no pleural effusion or pneumothorax. The lungs appear clear. No evidence of acute process.

MedMO : The lungs are clear. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities. No acute cardiopulmonary process.

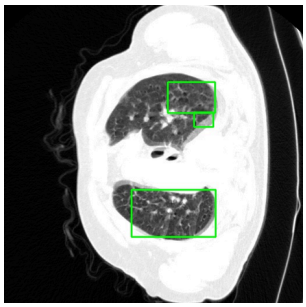
Figure 11. **Qualitative result example #1.** We show model predictions compared against ground truth annotations. The input medical image is displayed on the left, with corresponding text outputs on the right. Correct predictions are highlighted (highlighted in green) to demonstrate alignment with clinical ground truth, while differences indicate areas for potential improvement.



Ground Truth : Frontal and lateral chest radiographs again demonstrate a vascular stent and surgical clips. Moderate to severe cardiomegaly is unchanged. There is no definite focal consolidation. Pleural and parenchymal scarring have been more fully evaluated by CTA of the chest of . A small right pleural effusion is seen. There is no appreciable pneumothorax. The visualized upper abdomen is unremarkable. Rightward deviation of the trachea is consistent with left lobe thyroid enlargement. No definite focal consolidation. Small right pleural effusion.

MedMO : There is a stent in the right brachiocephalic vein. There is a small right pleural effusion. There is no focal consolidation or pneumothorax. There is mild cardiomegaly. There is no pulmonary edema. There is a small hiatal hernia. There is a right upper lobe nodule, better seen on the prior CT. Small right pleural effusion. No focal consolidation.

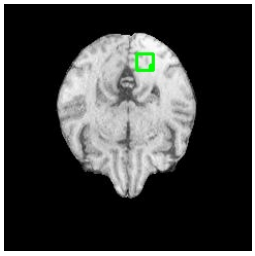
Figure 12. **Qualitative result example #2.** We show model predictions compared against ground truth annotations. The input medical image is displayed on the left, with corresponding text outputs on the right. Correct predictions are highlighted (highlighted in green) to demonstrate alignment with clinical ground truth, while differences indicate areas for potential improvement.



Ground Truth: The image is a high-resolution CT scan of the thoracic region, specifically showing the lungs and the heart. The heart is centrally located, with the lungs on either side. The region of interest, located in the lower-middle right-center of the image, shows an area of the lung that is abnormal in appearance, characterized by ground-glass opacity, consolidation, or other atypical features indicative of a COVID-19 infection. This region is adjacent to normal lung tissue, which may suggest that the disease has not spread extensively and could be at an early or localized stage. The proximity to normal lung tissue could imply a potential for further spread or influence on adjacent areas, although the image alone does not confirm the directionality of this relationship.

MedMO: The image is a high-resolution computed tomography (HRCT) scan of the chest, showing the lungs, heart, and part of the spine. The lungs are the primary organs visible, with the heart centrally located between them and the spine visible posteriorly. The region of interest, located in the lower-middle right-center of the image, shows an area of abnormality within the lung tissue. This area exhibits a texture and density that differ from the surrounding lung parenchyma, suggesting the presence of a pathological process, likely related to COVID-19 as indicated by the caption. The abnormal region is adjacent to normal lung tissue, which may be either unaffected by the disease or could potentially be at risk of being affected due to proximity. The presence of the disease in this region could be indicative of an early stage of infection, which is consistent with the knowledge that COVID-19 can present with ground-glass opacities and consolidation in the lungs, often in the peripheral distribution of the disease.

Figure 13. **Qualitative result example #3.** We show model predictions compared against ground truth annotations. The input medical image is displayed on the left, with corresponding text outputs on the right. Correct predictions are highlighted (highlighted in green) to demonstrate alignment with clinical ground truth, while differences indicate areas for potential improvement.



Ground Truth: The image is a transverse slice of a magnetic resonance imaging (MRI) scan of the brain, showing various brain structures such as the cerebral cortex, ventricles, and other intracranial structures. A region of interest is located at the center of the brain, which is the cerebral hemispheres, and it appears to be a transverse slice through the axial plane, likely at the level of the cerebral cortex, possibly involving the cerebral cortex and adjacent brain tissue. The region of interest, located centrally in the image, there is an area that appears to be abnormal, with an area ratio of 0.4% of the image, indicating a potential abnormality. This region is characterized by a difference in tissue character, which may suggest a variation in tissue density or composition, such as altered signal intensity, possibly due to the presence of a brain tumor. The abnormality is situated in the cerebral cortex, which is the central part of the brain, which could be affecting or be affected by the disease process, potentially disrupting the normal brain tissue surrounding it. The abnormality's relationship to other brain structures could be indicative of a pathological process that may be influencing or being influenced by adjacent brain tissue, such as causing a mass effect or being influenced by surrounding brain structures, potentially altering their function or being affected by them.

MedMO: The image is a magnetic resonance imaging (MRI) scan of the brain, showing various brain structures such as the cerebral cortex, ventricles, and possibly the basal ganglia or thalamus, without any medical devices visible. A region of interest is located at center part of the image horizontally and upper-middle part of image vertically, occupying approximately 0.4% of the total area. The region of interest, located in the cerebral cortex, exhibits an abnormality that differs in appearance from the surrounding brain tissue, potentially indicating a brain tumor. This abnormality could be affecting or be affected by adjacent brain structures due to its proximity, which may suggest an interaction or influence on nearby tissue, although the exact relationship depends on the nature of the tumor and its growth pattern. The MRI scan, with its high soft-tissue contrast, is instrumental in identifying such abnormalities.

Figure 14. **Qualitative result example #4.** We show model predictions compared against ground truth annotations. The input medical image is displayed on the left, with corresponding text outputs on the right. Correct predictions are highlighted (highlighted in green) to demonstrate alignment with clinical ground truth, while differences indicate areas for potential improvement.

References

[1] Asad Aali, Dave Van Veen, YI Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, et al. Mimic-iv-ext-bhc: labeled clinical notes dataset for hospital course summarization. *PhysioNet*, 1(0):10–13026, 2024. 7

[2] Asad Aali, Dave Van Veen, Yamin Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash Tehrani, Jangwon Kim, and Akshay Chaudhari. Mimic-iv-ext-bhc: Labeled clinical notes dataset for hospital course summarization (version 1.2.0), 2025. 7

[3] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. *CLEF (working notes)*, 2(6):1–11, 2019. 7

[4] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. 7

[5] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23, 2019. 7

[6] Harald Brodoefel, Christof Burgstahler, Ilias Tsiflikas, Anja Reimann, Stephen Schroeder, Claus D Claussen, Martin Heuschmid, and Andreas F Kopp. Dual-source ct: effect of heart rate, heart rate variability, and calcification on image quality and diagnostic accuracy. *Radiology*, 247(2):346–355, 2008. 7

[7] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024. 7

[8] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. 7

[9] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria, 2025. Association for Computational Linguistics. 7

[10] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. 7

[11] Ross W Filice, Anouk Stein, Carol C Wu, Veronica A Arteaga, Stephen Borstelmann, Ramya Gaddikeri, Maya Galperin-Aizenberg, Ritu R Gill, Myrna C Godoy, Stephen B Hobbs, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset. *Journal of digital imaging*, 33(2):490–496, 2020. 7

[12] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *ArXiv*, abs/2003.10778, 2020. 6

[13] Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with expert segmentations. *arXiv e-prints*, pages arXiv–2406, 2024. 6

[14] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024. 6

[15] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *ArXiv*, abs/2003.10286, 2020. 7

[16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 7

[17] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. 7

[18] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *ArXiv*, abs/1909.06146, 2019. 7

[19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 7

[20] Masakata Kawai, Noriaki Ota, and Shinsuke Yamaoka. Large-scale pretraining on pathological images for fine-tuning of small pathological benchmarks. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 257–267. Springer, 2023. 6

[21] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 7

[22] Lavita AI. Chatdoctor-healthcaremagic-100k. https://huggingface.co/datasets/lavita/ChatDoctor-HealthCareMagic-100k, 2023. Accessed: 2025-11-17. 7

[23] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 7

[24] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023. 7

[25] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021. 7

[26] Meng Lou, Hanning Ying, Xiaoqing Liu, Hong-Yu Zhou, Yuqin Zhang, and Yizhou Yu. Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. *Neural Networks*, page 107228, 2025. 6

[27] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024. 7

[28] MedAlpaca. medical_meadow_medical_flashcards. https://huggingface.co/datasets/medalpaca/medical_meadow_medical_flashcards, 2023. Accessed: 2025-11-17. 7

[29] MedAlpaca. medical_meadow_wikidoc. https://huggingface.co/datasets/medalpaca/medical_meadow_wikidoc, 2023. Accessed: 2025-11-17. 7

[30] Eszter Nagy, Michael Janisch, Franko Hržić, Erich Sorantin, and Sebastian Tschauner. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. *Scientific data*, 9(1):222, 2022. 7

[31] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 248–260. PMLR, 2022. 7

[32] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189, Cham, 2018. Springer International Publishing. 7

[33] Raymond D Ratliff. Meadows in the sierra nevada of california: state of knowledge. 1985. 7

[34] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 658–666, 2019. 1

[35] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1), 2024. 7

[36] Mourad Sarrouti. Nlm at vqa-med 2020: Visual question answering and generation in the medical domain. In *CLEF (Working Notes)*, 2020. 7

[37] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. 7

[38] I Siragusa, S Contino, ML Ciura, R Alicata, and R MedPix Pirrone. 2.0: A comprehensive multimodal biomedical data set for advanced ai applications. arxiv 2024. *arXiv preprint arXiv:2407.02994*. 7

[39] Yuri Tolkach, Lisa Marie Wolgast, Alexander Damanakis, Alexey Pryalukhin, Simon Schallenberg, Wolfgang Hulla, Marie-Lisa Eich, Wolfgang Schroeder, Anirban Mukhopadhyay, Moritz Fuchs, et al. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. *The Lancet Digital Health*, 5(5): e265–e275, 2023. 6

[40] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 6

[41] Simon van Vliet, Annina R Winkler, Stefanie Spriewald, Bärbel Stecher, Martin Ackermann, et al. Spatially correlated gene expression in bacterial groups: the role of lineage history, spatial gradients, and cell-cell interactions. *Cell systems*, 6(4):496–507, 2018. 7

[42] Patrick Wagner, Maximilian Springenberg, Marius Kröger, Rose KC Moritz, Johannes Schleusener, Martina C Meinke, and Jackie Ma. Semantic modeling of cell damage prediction: a machine learning approach at human-level performance in dermatology. *Scientific Reports*, 13(1):8336, 2023. 6

[43] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, 2020. Association for Computational Linguistics. 7

[44] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *ArXiv*, abs/2504.00993, 2025. 7

[45] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations*, 2025. 5, 7

[46] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501–036501, 2018. 7

[47] Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*, 2024. 7

[48] YongchengYAO. Kipa22. https://huggingface.co/datasets/YongchengYAO/KiPA22, 2025. Accessed: 2025-11-17. 6

[49] Bei Yu, Yingya Li, and Jun Wang. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China, 2019. Association for Computational Linguistics. 7

[50] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025. 5

[51] Jingkun Yue, Siqi Zhang, Zinan Jia, Huihuan Xu, Zongbo Han, Xiaohong Liu, and Guangyu Wang. Medsg-bench: A benchmark for medical image sequences grounding. *arXiv preprint arXiv:2505.11852*, 2025. 7

[52] Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application, 2023. 7

[53] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 7