

Scalable Training of Spatially Grounded 2D Vision–Language Models for Radiology

Yusuf Salcan^{1,4*}[0009–0005–1040–4011], Simon Ging^{1,2*}[0000–0002–5159–9792],
Robin Tibor Schirrmester³[0000–0002–5518–7445], Philipp
Arnold³[0000–0003–0686–7361], Elmar Kotter³[0000–0001–9022–6000], Behzad
Bozorgtabar^{2†}[0000–0002–5759–4896], and Thomas Brox^{1†}[0000–0002–6282–8861]

¹ Computer Vision Group, University of Freiburg, Germany
gings@cs.uni-freiburg.de

² Adaptive & Agentic AI (A3) Lab, Aarhus University, Denmark

³ Department of Radiology, Medical Center – University of Freiburg, Germany

⁴ CRIION-AI Lab, Freiburg, Germany

Abstract. We study how to train visually grounded vision–language models (VLMs) for radiology without manual spatial annotations. We introduce *RefRad2D*, a large-scale bilingual (German/English) dataset of 1.2M CT and MRI image–text pairs derived from clinical practice, with task-specific VQA and spatial grounding subsets generated automatically via LLM-based curation and automated segmentation. Trained on this data, our model *RadGrounder* jointly performs report generation, visual question answering, and spatial grounding via bounding-box detection or segmentation. On external VQA benchmarks (Slake, VQA-RAD), RadGrounder achieves competitive results with specialized medical VLMs. Adding our clinical data to the training mixture improves open-ended VQA over fine-tuning on the downstream datasets alone, showing the transferability of our dataset. Crucially, adding grounding supervision does not degrade language quality, enabling spatially verifiable outputs at no cost to VQA performance. Code and models are available at <https://github.com/lmb-freiburg/radgrounder>.

Keywords: Radiology VLM · Visual grounding · CT · MRI

1 Introduction

Vision–language models (VLMs) for radiology can generate coherent medical text, but a significant limitation remains: they cannot reliably ground their output in specific image regions. Without spatial grounding, predictions are difficult to verify against the underlying image, which limits clinical trust. Training such models for CT and MRI is further hindered by data scarcity. Approaches that consume 2D slices offer a scalable alternative to 3D volumetric methods, yet prior efforts have struggled with scale; for instance, Med-Gemini authors [16,24]

* Equal contribution.

† Equal supervision.

tailored a 2D CT slice dataset but their strict filtering yielded only 4,009 images, leaving 2D slice supervision at scale unexplored.

Recent medical VLMs achieve strong performance on radiology benchmarks, from lightweight generalists like BiomedGPT [26] and LLaVA-Med [11] to large-scale models like RadFM [22], which supports both 2D and 3D inputs. Most work on spatially grounded medical VLMs has focused on chest X-rays, including CheXagent [3] and MAIRA-2 [2]. VividMed [13] extended grounding to broader medical imaging, but relies on synthetic data from open datasets. Grounded VLMs trained on large-scale clinical CT and MRI data remain underexplored.

In this work, we introduce RadGrounder, a PaliGemma 2-based [19] multi-task VLM that performs report generation, VQA, and visual grounding on CT and MRI slices. To train it, we build RefRad2D, a bilingual (German/English) corpus of 1.2M image–text pairs derived from clinical routine via an LLM-driven curation pipeline, with spatial labels from TotalSegmentator [21,1].

Contributions.

- We present RefRad2D, a large-scale bilingual (German/English) 2D CT/MRI dataset for VLM training with 1.2M image–text pairs and automatically derived spatial grounding annotations.
- We introduce RadGrounder, a multi-task model that jointly performs VQA and visual grounding. Built on standard VLM components, it shows that 2D slice-level supervision from routine reports is a scalable strategy for training radiology VLMs. Code and pretrained models are publicly available.
- We ablate training configurations and grounding strategies, finding that (1) token-based bounding box prediction provides effective localization without requiring an auxiliary segmentation head or additional loss terms, and (2) adding spatial grounding supervision does not degrade VQA or report generation quality.
- We validate our approach on external benchmarks (VQA-RAD [10], Slake [12]), showing competitive performance with other medical VLMs while enabling verifiable, spatially grounded interpretations.

2 Dataset

We introduce RefRad2D, a bilingual dataset with visual grounding annotations, derived from clinical logs at a university hospital. The dataset comprises 1.2 million unique image–caption pairs extracted from 945k CT and 321k MRI slices, representing a decade of clinical practice [18].

Preprocessing and Bilingual Expansion. Raw reports often contain longitudinal references (e.g., “lesion has grown”) that induce hallucinations in single-frame VLMs. We used GPT-OSS (120B) [15] to rewrite captions, removing temporal context while preserving current findings [20]. To address Anglo-centric bias [14], we translated all German reports into English using Gemma 3 (27B) [6]. To refine translation quality, we sampled outputs and judged them with a stronger model, GPT-OSS (120B), iteratively improving the translation prompt based on identified errors [4]. We train on a mix of both languages.

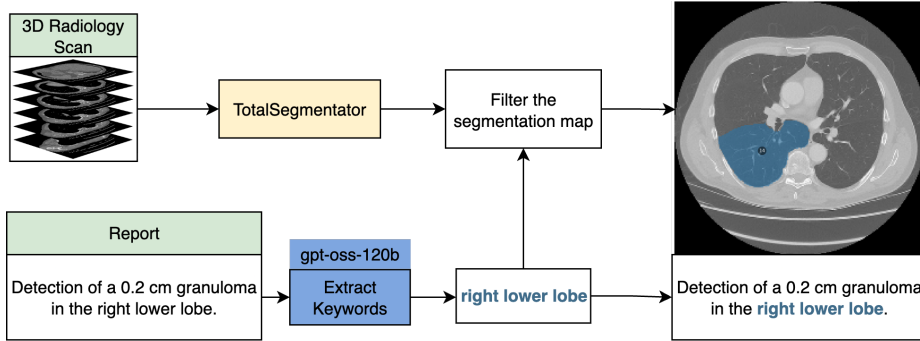


Fig. 1. Overview of the components to generate the data for visual grounding. We use TotalSegmentator [21,1] to create 3D segmentation masks. Via slicing and LLM-based keyword extraction we create dense 2D pixel-level annotations. A strict set-intersection filter ($C_T \cap C_I \neq \emptyset$) ensures high-quality matching between textual entities and visual regions.

Automated Anatomical Grounding. To generate dense, pixel-level annotations without manual labeling, we developed an automated pipeline using 181,362 CT and 36,026 MRI 3D volumes. This produced the RefRad2D-Grounded subset: 236,157 grounded slice–text pairs (217,692 CT, 18,465 MRI).

Anatomical Segmentation. We processed full 3D volumes using TotalSegmentator [21,1] to generate masks for 117 CT and 50 MRI classes. These were harmonized into a unified schema of $C = 121$ classes by merging shared anatomies. Volumes were sliced to match the VLM’s 2D input format.

Keyword-Mask Matching. To link visual regions to text (Fig. 1), we used GPT-OSS (120B) to extract anatomical mentions from each caption and map them to the same $C = 121$ class schema used by TotalSegmentator. Let $C_T \subseteq \{1, \dots, C\}$ denote the classes mentioned in caption T , C_I the classes with masks in slice I , and M_I^T the segmentation mask for this class in this slice. We construct a valid training triplet (I, T, M_I^T) only if $C_T \cap C_I \neq \emptyset$. For detection, we convert the masks into bounding boxes by taking the box that covers the mask.

Synthetic VQA Generation. For VQA training, we generated the RefRad2D-VQA dataset (~ 9.6 M pairs). Using Gemma 3 (27B), we derived 5 QA pairs per image directly from clinical findings (Open, Yes/No, Multiple Choice) and 3 QA pairs from the slice metadata [11].

3 Method

We propose RadGrounder, a multi-task Vision–Language Model (VLM) that generates radiology reports and localizes findings through bounding-box detection or segmentation.

Architecture. Our architecture builds upon the PaliGemma 2 (3B) framework [19], comprising a SigLIP-So400m [25] vision encoder and a Gemma-2B

language decoder [5]. The model processes an image $I \in \mathbb{R}^{H \times W \times 3}$ into visual tokens Z_v , which are concatenated with text tokens Z_t and are processed autoregressively. We investigate two grounding strategies on top of this foundation: token-based bounding-box detection, which the model can perform natively via text generation, and an auxiliary segmentation head.

Bounding-Box Detection. For spatial grounding, we treat detection as a text-generation task. We extend the vocabulary with coordinate tokens discretized into 512 bins and class-identifying tokens. The model generates a structured sequence:

$$\langle \text{p bbox} \rangle [\text{LOC}] \text{id}=\langle \text{segID} \rangle \text{KEYWORD} \langle \text{/p} \rangle \quad (1)$$

where [LOC] denotes the bounding box coordinates $\langle y_{min}, x_{min}, y_{max}, x_{max} \rangle$ and id maps to our unified schema of $C = 121$ anatomical classes. To resolve ambiguity in slices containing multiple instances of the same organ (e.g., multiple lymph nodes), we employ a class-wise merging strategy, predicting a single union bounding box for all instances of a class.

Segmentation Head. We also explore pixel-level grounding via a lightweight mask decoder following VividMed [13] and SAM [9]. When the model generates a $\langle \text{/seg} \rangle$ token, its hidden state is projected into a 256-dim prompt that drives a Two-Way Transformer [9] over the image embeddings to produce a binary mask. Unlike detection, this approach requires an auxiliary decoder and an additional segmentation loss.

Training Objective. The model is trained end-to-end with auto-regressive cross-entropy $\mathcal{L}_{txt}$ for next-token prediction. Detection grounding is trained entirely through $\mathcal{L}_{txt}$, as bounding-box coordinates are generated as tokens. For the segmentation variant, we add an auxiliary loss:

$$\mathcal{L} = \mathcal{L}_{txt} + \lambda_{seg}(\mathcal{L}_{focal} + \mathcal{L}_{dice}) \quad (2)$$

where $\mathcal{L}_{focal}$ and $\mathcal{L}_{dice}$ handle the class imbalance inherent in medical segmentation. We employ a dynamic sampling strategy where batches alternate between report generation, VQA, and grounded tasks.

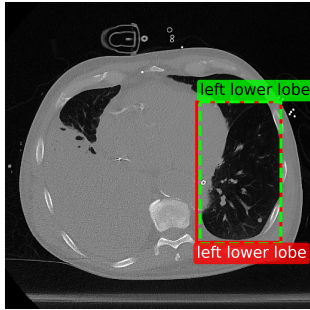
LLMScore. Standard n-gram metrics (e.g., CIDEr) struggle with semantic equivalence and the complexities of medical text. We developed LLMScore by adopting the “LLM-as-a-judge” paradigm [27,7]. We use Gemma 3 (27B) as the evaluator to compare the generated report against the ground truth. The judge assesses clinical factuality and semantic correctness on a 5-point scale normalized to $[0, 1]$ and outputs a brief textual justification for its decision. To validate this metric, three radiologists independently scored 200 QA pairs from VQA-RAD. Inter-annotator agreement was high (ordinal Krippendorff’s $\alpha = 0.958$) [8], and LLMScore achieved a Pearson correlation of $r = 0.977$ with the mean human scores, indicating strong alignment with expert judgment.

Spatial-Semantic Evaluation (G-IoU). Standard Intersection over Union (IoU) assumes a fixed vocabulary and fails when a model generates a keyword that is semantically correct but lexically distinct from the ground truth. We propose Grounding-IoU (G-IoU) to jointly measure spatial and semantic fidelity.



Prompt: <image> Clinical information: Status post sleeve gastrectomy, bypass, with increasing abdominal pain. Question: Internal hernia? Caption this image:
Ground Truth: Nonspecific increased mesenteric lymph nodes, up to 10mm.
Prediction: Increased mesenteric lymph nodes up to 1.2 cm.
Metrics:
CIDEr: 1.0877
LLM Score: 0.7500
LLM Reason: The candidate and reference both identify increased mesenteric lymph nodes. The sizes are very close (1.2cm vs 10mm = 1cm), indicating a mostly right answer.

Fig. 2. Qualitative example of RadGrounder’s report generation on an abdominal CT slice (soft tissue window). The model successfully identifies increased mesenteric lymph nodes and accurately estimates their size. Despite a low traditional CIDEr score (1.08), the LLM-as-a-judge correctly assigns a high factual score (0.75) by recognizing the semantic equivalence of the measurements (1.2 cm vs. 10 mm).



Prompt: <image> Clinical information: Severe pneumonia in a patient with COPD Gold 4. Question: Extent of the pneumonia? Caption this image and detect keywords:
Ground Truth: Approximately 6 mm measuring round lesion in the **left lower lobe**.
Prediction: Nodular densities in the left lung, exemplified in the **left lower lobe** with 5 mm, most likely postinflammatory.
Metrics:
CIDEr: 0.1373
LLM Score: 0.7500
LLM Reason: The candidate and reference both identify a lesion in the left lower lobe and provide similar size estimations (5mm vs 6mm). The candidate adds 'nodular densities' and 'postinflammatory' which aren't explicitly in the reference but are plausible given the clinical context.
G-IoU: 0.9497
mIoU: 0.9497

Fig. 3. Detection grounding performance on a chest CT (lung window). RadGrounder accurately localizes a nodular lesion in the left lower lobe, generating a bounding box that strongly aligns spatially and semantically with the ground truth (G-IoU: 0.9497).

We first match predicted and ground-truth entities using the cosine similarity of their text embeddings. The G-IoU is then computed as the sum of their spatial IoUs weighted by this semantic similarity. The score is normalized by the maximum number of predicted or ground-truth entities, penalizing both spatial hallucinations (false positives) and diagnostic omissions (false negatives). We then average the per-sample G-IoU over all samples in the respective test set.

Table 1. Quantitative results on the internal RefRad2D test set for Report Captioning, VQA, and spatial grounding tasks. The table compares our RadGrounder detection model (row 4) against PaliGemma 2 [19] (PaliG. 2), Gemma 3 [6], and MedGemma [17] (MedG.) (top), and ablates various training configurations (bottom). SL.: SigLIP vision encoder. We highlight **values within the confidence interval of the highest score** separately for the baseline comparison and the ablations.

Training data	Report				VQA					Grounded
Report VQA Slake VQA-RAD Detect. Segment	CIDEr	LLM			F1 Open	LLM Open	Acc. Closed			G-IoU
PaliG. 2 (3B) [19]	0.6 0.1	2.6 0.5			12.7 1.9	22.8 2.3	51.5 2.7			—
Gemma 3 (4B) [6]	0.1 0.0	21.7 1.3			5.5 0.8	31.8 2.7	13.0 1.6			—
MedG. (4B) [17]	0.4 0.1	21.8 1.3			9.0 1.2	40.6 2.9	16.6 2.2			—
✓ ✓ ✓ ✓ ✓	37.5 3.3	30.3 1.4			48.9 2.8	60.1 2.6	94.6 1.4			43.6 1.6
+ Unfreeze finet. SL.	34.6 3.3	28.0 1.4			47.8 2.8	60.1 2.5	95.1 1.5			43.4 1.7
+ Unfreeze orig. SL.	36.9 3.2	29.9 1.4			49.2 2.8	60.8 2.6	94.0 1.5			45.2 1.7
✓ ✓ ✓ ✓ ✓	36.9 3.2	30.1 1.4			50.4 2.8	64.1 2.7	95.6 1.3			36.9 1.6
✓ ✓	39.0 3.4	30.7 1.5			49.2 2.8	60.3 2.6	94.5 1.5			—
	0.1 0.0	1.5 0.5			15.7 2.2	21.8 2.5	32.4 2.2			—
✓ ✓ ✓ ✓	37.3 3.3	30.6 1.4			50.1 2.8	61.2 2.6	94.4 1.5			—
✓ ✓ ✓ ✓	39.2 3.3	30.6 1.5			49.1 2.8	60.2 2.6	94.5 1.5			43.6 1.6

4 Experiments

Implementation Details. All models were trained on a single NVIDIA H100 GPU for 6 epochs with a batch size of 24 and 12 gradient accumulation steps, using the Adafactor optimizer with a learning rate of 5e-5. We apply random scale cropping, contrast adjustment, and intensity shifts as data augmentation. Training with the frozen SigLIP encoder on the full dataset mixture took approximately 2.5 days.

Datasets and Protocols. The RefRad2D training split contains 760,409 CT and 256,197 MRI image-report pairs. The RefRad2D-VQA subset provides 8 QA pairs per image, and the RefRad2D-Grounded training subset contains 174,655

Table 2. Performance on Slake [12] and VQA-RAD [10]. SL.: SigLIP vision encoder, PaliG.: PaliGemma, MedG.: MedGemma. We used the train/test split from Xu et al. [23] due to overlap between train and test images in the original VQA-RAD splits [24]. *Reported results are not available for that split. We highlight **values within the confidence interval of the highest score** separately for the baseline comparison and the ablations.

Training data		Slake				VQA-RAD			
Report	VQA Slake VQA-RAD Detect. Segment	F1	Recall Open	LLM Open	Acc. Closed	F1	LLM Open	Acc. Closed	
PaliG. 2 (3B) [19]		24.5	24.3	34.2 3.5	58.1	37.9	26.0 4.1	55.6	
Gemma 3 (4B) [6]		40.2	33.3	22.3 3.2	53.0	33.6	11.1 3.5	33.6	
MedG. (4B) [17]		72.3	63.3	54.9 3.8	87.6	49.9	20.8 4.2	69.1	
BiomedGPT-B [26]		85.2	—	—	89.9	*	—	*	
LLaVA-Med [11]		—	87.1	—	86.8	*	—	*	
RadFM (14B) [22]		84.4	—	—	—	*	—	—	
Med-Gemini [24]		75.8	—	—	84.8	*	—	78.8	
✓ ✓ ✓ ✓ ✓		87.7 2.4	88.1 2.4	88.8 2.3	90.3 2.7	50.7 3.5	44.3 4.8	64.7 4.6	
Unfreeze finet. SL.		88.4 2.4	88.9 2.4	89.9 2.3	90.8 2.6	48.7 3.4	44.3 4.8	62.7 4.6	
Unfreeze orig. SL.		87.6 2.3	88.2 2.3	88.6 2.2	91.7 2.7	50.2 3.4	44.9 4.8	64.5 4.6	
✓ ✓ ✓ ✓ ✓		86.4 2.5	86.9 2.5	87.3 2.5	89.6 2.8	50.4 3.5	46.6 5.2	65.9 4.4	
✓ ✓		19.2 2.9	20.3 3.0	31.9 3.4	63.0 4.5	39.6 3.4	33.0 4.5	52.9 4.7	
✓ ✓		86.9 2.5	87.6 2.5	87.8 2.4	88.4 3.2	47.9 3.4	28.6 4.2	68.1 4.4	
✓ ✓ ✓ ✓		87.2 2.5	87.6 2.4	88.8 2.3	90.8 2.7	49.2 3.4	42.8 4.7	63.2 4.5	
✓ ✓		18.3 2.8	20.2 3.0	31.4 3.4	58.8 4.7	40.2 3.4	35.9 4.5	52.9 4.9	

CT and 14,721 MRI spatially grounded triplets. All internal evaluation sets were standardized to 2,000 test samples. For external benchmarking, we use Slake [12] (9,849 training / 2,070 testing QA pairs) and VQA-RAD [10] (3,064 training / 451 testing QA pairs). For open-ended evaluation we use LLMscore (Sec. 3). We report 95% bootstrap confidence intervals ($B=10,000$) over test set samples.

Vision Encoder Strategy (Tab. 1, 2). Before multi-task training, we adapt the SigLIP encoder to the medical domain via contrastive pretraining on ReRad2D. We then ablate whether to keep this adapted encoder frozen or unfrozen during the subsequent LLM training, and compare against unfreezing the original (non-adapted) SigLIP weights. All three strategies perform comparably within confidence intervals on both internal and external benchmarks. Since freezing the encoder reduces memory footprint and accelerates training, we adopt it as our standard configuration.

Dataset Mixture and Grounding (Tab. 1, 2). Baseline models score near zero on report captioning CIDEr, though LLMscore is less affected (e.g., MedGemma: 0.4 vs. 21.8). CIDEr penalizes any n-gram mismatch with the ref-

erence, while LLMScore better captures semantic correctness (Fig. 2). The low baselines are expected: these models were not trained on our data, and the task requires generating hospital-specific text from a single slice without clinical context. Token-based detection (Fig. 3) achieved higher G-IoU (43.6 vs. 36.9) than the segmentation head while requiring no additional loss terms or decoder parameters. Neither grounding strategy degrades VQA or report generation quality compared to the no-grounding baseline. Mixture ablations show clear domain effects: training exclusively on internal data achieved the highest report generation score (CIDEr: 39.2) but failed to generalize to external VQA (Slake F1: 18.3). Conversely, adding our internal corpus to external fine-tuning improved open-ended VQA on most metrics (e.g., Slake F1: 87.7 vs. 86.9; VQA-RAD F1: 50.7 vs. 47.9). The exception was VQA-RAD closed accuracy, where the external-only model scored higher (68.1% vs. 64.7%).

Grounding Error Analysis. The 88 anatomical classes with test support are heavily long-tailed: the five most frequent cover half of all instances. Class frequency correlates only mildly with per-class G-IoU (Pearson $r=0.34$). The model misses 18.7% of ground-truth findings (false negatives) and produces 13.1% spurious predictions (false positives). Among matched boxes, 54.7% are fully correct. Another 28.6% localize correctly but name the wrong structure, 4.6% name the right structure but mislocalize, and 12.1% get both wrong. Spatial and textual errors correlate only weakly ($r=0.29$).

Comparison to State-of-the-Art (Tab. 2). Using our standard configuration (frozen encoder, full multi-task mixture), RadGrounder achieves competitive performance on external benchmarks. On Slake, our model reaches F1 87.7 and Closed Accuracy 90.3, comparable to BiomedGPT-B [26] (85.2 / 89.9) and LLaVA-Med [11] (86.8 Acc), while surpassing Med-Gemini [16,24] (75.8 / 84.8) and MedGemma [17] (72.3 / 87.6). On VQA-RAD, our model achieves the highest Open F1 (50.7) among compared methods.

5 Conclusion

We presented RadGrounder, a visually grounded VLM for radiology trained without manual spatial annotations, and RefRad2D, a bilingual dataset of 1.2M CT and MRI image-text pairs with automatically derived spatial labels. On external VQA benchmarks, RadGrounder achieves competitive performance with specialized medical VLMs, and adding our clinical data improves over fine-tuning on downstream data alone. We explored two grounding strategies—token-based detection and auxiliary segmentation—finding that detection provides effective localization without additional model components or loss terms, and that neither strategy degrades VQA quality. RadGrounder thus enables spatially verifiable predictions at no cost to language performance.

Limitations. Raw images and reports cannot be shared for privacy reasons. The VQA and grounding annotations are generated automatically and require expert validation before clinical use. Our training data comes from a single hospital, and multi-center validation is needed to confirm generalization. The spa-

tial grounding targets anatomical structures (via TotalSegmentator) rather than pathological findings, limiting clinical utility for lesion-level localization. Finally, comparing detection and segmentation grounding is inherently difficult, as the two approaches solve different geometric tasks (box overlap vs. pixel-level mask agreement), making direct comparison via a single metric nontrivial.

Future Work. Natural extensions include multi-center training to improve generalization and pathology-level grounding using lesion annotations, which would enable clinically actionable spatial predictions beyond anatomical localization. Since we operate on independent 2D slices, enforcing inter-slice spatial coherence for volumetric consistency is another natural direction.

Acknowledgments. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 417962828, 539134284, and 499552394 – SFB 1597 – Small Data, as well as through EFRE (FEIH_2698644) and the state of Baden-Württemberg.



Baden-Württemberg



Co-funded by
the European Union

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akinci D’Antonoli, T., Berger, L.K., Indrakanti, A.K., Vishwanathan, N., Weiss, J., et al.: TotalSegmentator MRI: Robust sequence-independent segmentation of multiple anatomic structures in MRI. *Radiology* **314**(2) (Feb 2025). <https://doi.org/10.1148/radiol.241613>
2. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., et al.: MAIRA-2: Grounded radiology report generation (2024), arXiv:2406.04449
3. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., et al.: CheX-agent: Towards a foundation model for chest X-ray interpretation. In: AAAI 2024 Spring Symposium Series on Clinical Foundation Models (2024)
4. Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., et al.: AlpacaFarm: A simulation framework for methods that learn from human feedback. In: *Advances in Neural Information Processing Systems*. vol. 36 (2023)
5. Gemma Team: Gemma: Open models based on Gemini research and technology (2024), arXiv:2403.08295
6. Gemma Team: Gemma 3 technical report (2025), arXiv:2503.19786
7. Ging, S., Bravo, M.A., Brox, T.: Open-ended VQA benchmarking of vision-language models by exploiting classification datasets and their semantic hierarchy. In: *The Twelfth International Conference on Learning Representations* (2024)
8. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* **1**(1), 77–89 (2007). <https://doi.org/10.1080/19312450709336664>
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., et al.: Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3992–4003. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00371>

10. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data* **5**(1), 180251 (Nov 2018). <https://doi.org/10.1038/sdata.2018.251>
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., et al.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*. vol. 36 (2023)
12. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. pp. 1650–1654. IEEE (2021). <https://doi.org/10.1109/ISBI48211.2021.9434010>
13. Luo, L., Tang, B., Chen, X., Han, R., Chen, T.: VividMed: Vision language model with versatile visual grounding for medicine. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 1800–1821. Association for Computational Linguistics, Albuquerque, New Mexico (2025). <https://doi.org/10.18653/v1/2025.naacl-long.89>
14. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics* **9**(1), 12 (Mar 2018). <https://doi.org/10.1186/s13326-018-0179-8>
15. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), arXiv:2508.10925
16. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., et al.: Capabilities of Gemini models in medicine (2024), arXiv:2404.18416
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., et al.: MedGemma technical report (2025), arXiv:2507.05201
18. Smith-Bindman, R., Kwan, M.L., Marlow, E.C., Theis, M.K., Bolch, W., et al.: Trends in use of medical imaging in US health care systems and in ontario, canada, 2000–2016. *JAMA* **322**(9), 843–856 (09 2019). <https://doi.org/10.1001/jama.2019.11456>
19. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., Zhai, X.: PaliGemma 2: A family of versatile VLMs for transfer (2024), arXiv:2412.03555
20. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature Medicine* **29**(8), 1930–1940 (Aug 2023). <https://doi.org/10.1038/s41591-023-02448-8>
21. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence* **5**(5), e230024 (2023). <https://doi.org/10.1148/ryai.230024>
22. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications* **16**(1), 7866 (Aug 2025). <https://doi.org/10.1038/s41467-025-62385-7>
23. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., et al.: ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders (2023), arXiv:2308.01317
24. Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., et al.: Advancing multimodal medical capabilities of Gemini (2024), arXiv:2405.03162

25. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11941–11952. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.01100>
26. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., et al.: A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* **30**(11), 3129–3141 (Nov 2024). <https://doi.org/10.1038/s41591-024-03185-2>
27. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., et al.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track. vol. 36 (2023)