LoFi: Location-Aware Fine-Grained Representation Learning for Chest X-ray

Myeongkyun Kang[1,2], Yanting Yang[1,2], and Xiaoxiao Li[1,2†]

[1] The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[2] Vector Institute, Toronto, ON M5G 0C6, Canada
myeongkyun.kang@ubc.ca, xiaoxiao.li@ece.ubc.ca

Abstract. Fine-grained representation learning is crucial for retrieval and phrase grounding in chest X-rays, where clinically relevant findings are often spatially confined. However, existing methods remain limited on these tasks, as contrastive models lack explicit region-level supervision and large vision language models often fail to generalize fine-grained representations under external validation. To address these limitations, we propose Location-aware Fine-grained representation learning (LoFi), which jointly optimizes sigmoid, captioning, and location-aware captioning losses using a lightweight large language model. The location-aware captioning loss enables region-level supervision through grounding and dense captioning objectives, thereby facilitating fine-grained representation learning. Building upon these representations, we integrate a fine-grained encoder into retrieval-based in-context learning to enhance chest X-ray grounding across diverse settings. Extensive experiments demonstrate that our method achieves superior retrieval and phrase grounding performance on MIMIC-CXR and PadChest-GR.

Keywords: Fine-Grained Representation Learning · Retrieval · Phrase Grounding · In-Context Learning · Chest X-ray

1 Introduction

Clinically relevant findings in medical imaging are often spatially confined rather than semantically distinct, necessitating fine-grained representation learning [7]. Retrieval tasks identify relevant chest X-rays among similar candidates, requiring fine-grained discrimination [28], and phrase grounding tasks localize objects based on textual descriptions, demanding fine-grained cross-modal alignment [2]. However, medical contrastive models [12, 28] and large vision language models (LVLMs) [1, 5] remain limited in their ability to capture subtle and fine-grained representations, leading to suboptimal performance on these tasks.

Contrastive chest X-ray models have been proposed to effectively leverage paired radiology reports for fine-grained, clinically relevant representation learning [9, 12]. However, these models rely on implicit attention-based local image-text alignment without explicit region-level supervision (e.g., bounding boxes);

† Corresponding author.
Code available at https://github.com/myeongkyunkang/lofi.

incorporating such supervision can alleviate this limitation. Nevertheless, compared with prior web-scale pretraining approaches [23], training autoregressive models from scratch on image-text-box triples remains challenging in the medical domain due to the scarcity of region-level annotations. Therefore, leveraging a lightweight large language model (LLM) [21] and a region-level pretrained encoder [22] is desirable to address the challenges of medical image-text-box alignment.

A further challenge is that existing models, including LVLMs, remain limited in generalizing fine-grained representations under external validation [13]. To address this challenge, retrieval-based in-context learning (ICL) has been proposed to enable adaptation to new tasks by leveraging retrieved demonstrations at inference time [25]. However, its performance largely depends on the quality of the retrieved demonstrations, underscoring the need for fine-grained representations and learning objectives that facilitate them.

We propose **Lo**cation-aware **Fi**ne-grained representation learning (LoFi), which jointly optimizes sigmoid, captioning, and location-aware captioning losses with a lightweight LLM [21] and a region-level pretrained encoder [22]. The captioning loss facilitates learning from long-form text, while the location-aware captioning loss enables region-level supervision through grounding and dense captioning objectives, thereby enhancing fine-grained representation learning. Building upon these representations, we integrate a fine-grained encoder into retrieval-based ICL to enhance chest X-ray grounding under external validation. We train our model on 208,949 radiology reports and 394,835 text-box pairs from MIMIC-CXR, and demonstrate the superiority of our fine-grained representations through retrieval tasks on MIMIC-CXR, as well as internal and external phrase grounding tasks on PadChest-GR.

In summary, the contributions are as follows: *(a)* We propose LoFi, which leverages a lightweight LLM to jointly optimize sigmoid, captioning, and location-aware captioning losses for fine-grained chest X-ray representation learning. *(b)* We integrate a fine-grained encoder into retrieval-based ICL, providing a practical solution for real-world chest X-ray grounding. *(c)* We demonstrate the effectiveness of our approach on retrieval and grounding tasks through comprehensive internal and external validation.

2 Method

Location-Aware Fine-Grained Representation Learning (LoFi) The proposed method is shown in Fig. 1(a). Given a dataset $\mathcal{D}$ of image-text-box triplets (I, T, B), where box annotations may be optional, we aim to train an image encoder E_I and a text encoder E_T, leveraging a lightweight LLM LLM. We employ 64 attention pooling modules and a projection layer to align the E_I outputs with the LLM input dimension (notation omitted for simplicity). We implement three objectives to facilitate learning from long-form text and region-level supervision: (a) a sigmoid loss, (b) a captioning loss, and (c) a location-aware captioning loss consisting of a grounding loss and a dense captioning loss.

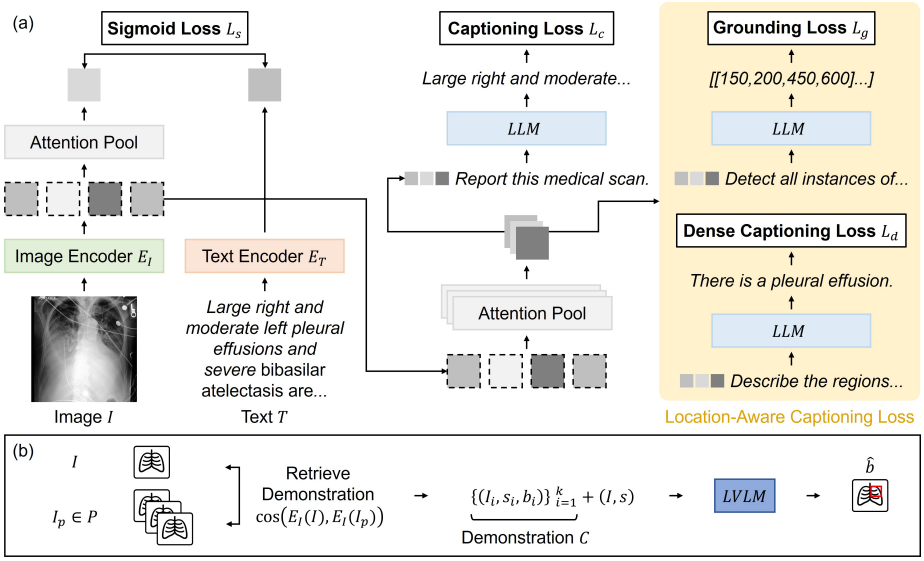


Fig. 1: Illustration of (a) location-aware fine-grained representation learning frameworks and (b) a retrieval-based in-context learning pipeline.

We adopt a sigmoid loss [22] as our contrastive objective, encouraging the model to assign higher similarity scores to matched pairs and thus learn representations that are well suited for retrieval. Formally, the sigmoid loss is defined as

$$\mathcal{L}_s = -\mathbb{E}_{(I,T)\sim\mathcal{D}} \left[\log \sigma\big(\delta(I,T)\, E_I(I)^\top E_T(T)\big)\right], \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid, and $\delta(I,T) \in \{+1,-1\}$ is a sign indicator that equals $+1$ for matched positive pairs and -1 for in-batch negative pairs. However, the sigmoid loss alone is limited in leveraging long-form text and region-level supervision of fine-grained object locations during training. To address this issue, we adapt prior works [22, 23] to our learning objective using a lightweight LLM. We denote the conditional distribution induced by the LLM as p_{LLM}, and define the autoregressive captioning loss as

$$\mathcal{L}_c = -\mathbb{E}_{(I,T)\sim\mathcal{D}} \left[\log p_{LLM}(T \mid E_I(I), \mathcal{P}_c)\right], \tag{2}$$

where $\mathcal{P}_c$ denotes a fixed prompt (e.g., *"Report this medical scan."*). Next, to incorporate location information of fine-grained objects during training, we adapt the grounding objective and the dense captioning objective to our autoregressive loss. Note that a text (e.g., a radiology report) T consists of sentences, and we denote a sentence as $s \in T$. We denote the set of bounding boxes associated with s by $b \subseteq B$, where each box is represented by the coordinates of its top-left and bottom-right corners $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$. Formally, we define the grounding loss as

$$\mathcal{L}_g = -\mathbb{E}_{(I,T,B)\sim\mathcal{D}_B,\, s\sim T} \left[\log p_{LLM}(b \mid E_I(I), \mathcal{P}_g, s)\right], \tag{3}$$

where $\mathcal{D}_B \subseteq \mathcal{D}$ denotes samples with bounding boxes, and $\mathcal{P}_g$ denotes a fixed prompt (e.g., *"Detect all instances of..."*). For the dense captioning loss, we swap the prediction target from b to s. Formally, we define the dense captioning loss as $\mathcal{L}_d = -\mathbb{E}_{(I,T,B)\sim\mathcal{D}_B,\, s\sim T}\left[\log p_{LLM}(s \mid E_I(I), \mathcal{P}_d, b)\right]$, where $\mathcal{P}_d$ denotes a fixed prompt (e.g., *"Describe the regions..."*). In summary, the total loss is defined as $\mathcal{L}_t = \mathcal{L}_s + \lambda \cdot \mathcal{L}_\tau, \tau \in \{c, g, d\}$, where τ indicates the loss type (captioning, grounding, or dense captioning), and λ is a hyperparameter for weighting the autoregressive loss (set to 5). When (I, T, B) are available, $\mathcal{L}_c$ and the location-aware captioning loss $\mathcal{L}_{g,d}$ contribute equally to the total loss.

Retrieval-Based In-Context Learning for Grounding The retrieval-based in-context learning (ICL) is illustrated in Fig. 1(b). Given an image query I and a sentence (or phrase) query s, the instruction-tuned model $LVLM$ predicts a set of bounding boxes $\hat{b}$ on I corresponding to s, i.e., $\hat{b} = LVLM(I, s)$. We use ICL to adapt $LVLM$ to new tasks without fine-tuning by leveraging a small set of demonstrations at inference time. The $LVLM$ predicts $\hat{b}$ conditioned on the set of demonstrations C, i.e., $\hat{b} = LVLM(I, s \mid C)$, where $C = \{(I_i, s_i, b_i)\}_{i=1}^{k}$ denotes a set of k image-sentence-box triples. Following retrieval-based approaches [25], we construct C by selecting the top-k most relevant triplets from the candidate pool. Given a candidate pool P and an image encoder E_I, we retrieve relevant demonstrations by computing $\cos\left(E_I(I), E_I(I_p)\right)$ for all $I_p \in P$, where $\cos(\cdot, \cdot)$ denotes cosine similarity. Since ICL performance relies on the quality of retrieved demonstrations, the fine-grained representations captured by E_I play a crucial role in effective grounding, underscoring the value of fine-grained representation learning in our framework.

Discussion In phrase grounding, inter-observer variability leads to inconsistent bounding box annotations, posing challenges for model training and evaluation [2]. Retrieval-based ICL is particularly effective in this setting, as it mitigates annotation inconsistency at inference time without additional fine-tuning. In contrast, when training is feasible, models can be fine-tuned with $\mathcal{L}_g$, thereby further enhancing performance on downstream tasks. Across both settings, strong fine-grained representations are essential for achieving high performance, which aligns with our training objective.

Implementation Details We adopted a pretrained SigLIP2-400M [22] with an input size of 512 as E_I and E_T, and used Gemma-3-270M [21] as the LLM. Following prior work [12, 24], we randomly sampled 4-8 sentences from radiology reports for contrastive learning due to the 64 token limit of the text encoder. We normalized bounding box coordinates to $[0, 1000]$ and converted them into strings to compute autoregressive $\mathcal{L}_g$ and $\mathcal{L}_d$. We applied LoRA [8] with rank 16 to the query, key, value, and output layers of E_I and E_T, and with rank 4 to the query and value layers of the LLM. We trained the model for 10 epochs with a batch size of 16 using cosine annealing with an initial learning rate of 3e-4.

Table 1: Retrieval performance on the MIMIC-CXR dataset. R@K denotes Recall at K, where K is the cutoff rank.

Model	Image-to-Text					Text-to-Image				
	R@1	R@5	R@10	R@20	R@40	R@1	R@5	R@10	R@20	R@40
SigLIP2 [22]	0.08	0.95	1.56	2.59	4.32	0.27	0.98	1.60	2.95	4.69
BiomedCLIP [27]	0.82	2.10	3.78	7.07	12.66	0.58	1.77	3.17	6.00	11.02
BMC-CLIP [15]	0.12	0.53	1.15	2.55	4.44	0.05	0.80	1.65	2.63	4.78
MedSigLIP [19]	4.61	14.27	21.92	32.69	44.82	2.38	10.33	17.47	26.43	38.48
CARZero [12]	11.64	29.52	40.17	51.19	63.98	9.67	26.02	36.24	48.40	62.37
RadIR [28]	7.36	22.62	32.48	44.70	57.44	7.11	22.70	32.61	45.44	59.91
Ours	13.90	35.57	47.29	60.20	72.53	11.91	31.30	42.36	54.46	66.71

For retrieval-based ICL, we used the instruction-tuned MedGemma-1.5-4B [19] with an input size of 896, motivated by our empirical observation that general-purpose LVLMs (e.g., Qwen [10]) have limited instruction-following capabilities for chest X-rays. Other LVLMs [14, 13, 4, 1, 5] had limited context lengths (e.g., 4096 tokens) and therefore could not support ICL.

3 Experiments and Results

Datasets **MIMIC-CXR** is a dataset consisting of 377,110 chest X-rays and 227,835 radiology reports [11]. We utilized the findings and impression sections annotated by an LLM in [26] and excluded non-frontal and low-quality chest X-rays for our experiments using [6]. For grounding annotations, a total of 394,835 A^{++}-rated text-box pairs from 153,602 studies in MIMIC-Ext [17] were used. As the dataset provides machine-generated silver-standard ground-truth annotations, we applied a weighted box fusion strategy [20] to merge duplicate boxes. Notably, unlike typical phrase grounding datasets for abnormal findings (e.g., PadChest-GR [2]), MIMIC-Ext consists of anatomical-level bounding box annotations, providing comparatively larger bounding boxes. The dataset was used to train our model on the official training split, and the model's fine-grained representations were evaluated on 2,432 patient studies from the official test split via retrieval tasks. **PadChest-GR** is a dataset consisting of 4,555 chest X-rays paired with grounding reports containing text-box annotations [2]. We exclusively used annotations with non-empty bounding boxes in our phrase grounding experiments. We used 1,238 text-box pairs from 604 chest X-rays in the official test split for evaluation and 4,341 text-box pairs from 2,096 chest X-rays for training (or as the candidate pool).

3.1 Retrieval Experiments

We evaluated the fine-grained representations of our model on the MIMIC-CXR dataset using image-to-text and text-to-image retrieval tasks. We split each radiology report into five-sentence sub-reports and evaluated report-level retrieval

Table 2: Phrase grounding performance on the PadChest-GR dataset for external validation using retrieval-based ICL with MedGemma and k demonstrations.

Model	k=20					k=40				
	Ro/L	Ro/S	F@0.5	P@0.5	R@0.5	Ro/L	Ro/S	F@0.5	P@0.5	R@0.5
SigLIP2 [22]	52.50	42.14	17.75	18.61	16.96	56.87	44.70	21.47	22.53	20.51
BiomedCLIP [27]	56.19	42.39	18.43	19.57	17.42	60.14	45.40	22.06	22.94	21.24
BMC-CLIP [15]	53.24	42.20	17.15	18.16	16.24	56.71	44.66	20.92	21.95	19.99
MedSigLIP [19]	58.79	44.69	21.77	22.62	20.97	61.63	46.89	24.51	25.12	23.93
RadIR [28]	56.03	42.13	17.50	18.54	16.57	56.34	41.72	17.04	17.77	16.37
Ours	59.10	44.49	22.48	23.40	21.63	63.55	48.00	25.34	25.98	24.72

performance using Recall at K (R@K), where K is the cutoff rank [24]. We compared our method against several contrastive-based representation learning models, including SigLIP2 [22], BiomedCLIP [27], and BMC-CLIP [15], as well as models trained on MIMIC-CXR, such as MedSigLIP [19], CARZero [12], and RadIR [28]. We loaded the publicly released checkpoints of the comparison methods and evaluated them on the test set.

Table 1 shows the retrieval performance on the MIMIC-CXR dataset. General medical contrastive models achieve lower retrieval performance than models tailored to chest X-ray data, highlighting the advantage of modality-specific approaches. MedSigLIP exhibits lower accuracy than CARZero and RadIR, which can be attributed to its training on medical data alongside a large proportion of non-medical images. Our method outperforms state-of-the-art contrastive models trained on MIMIC-CXR in both retrieval directions, demonstrating its effectiveness in learning fine-grained representations.

3.2 Phrase Grounding Experiments

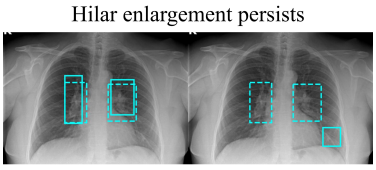
We conducted phrase grounding experiments on the PadChest-GR dataset to evaluate fine-grained representations in external validation using retrieval-based ICL and downstream performance in internal validation using pretrained parameters. In external validation, PadChest-GR was used exclusively for evaluation, whereas in internal validation, it was used for both model training and evaluation. We reported the localization (Ro/L) and shape-matching (Ro/S) scores of robust detection outcome [16], as well as precision (P@0.5), recall (R@0.5), and F1 score (F@0.5) at an intersection over union threshold of 0.5 (rather than average precision, which relies on confidence scores) [10]. All models were evaluated under a single-class setting to facilitate comparison with each other.

External Validation We evaluated the phrase grounding performance of contrastive models [22, 27, 15, 19, 28] using retrieval-based ICL with MedGemma [19]. We assessed performance across different settings by varying the number of demonstrations (k=20 and k=40). CARZero [12] was excluded as it only supports cross-modal similarity. Additionally, we compared our method (*w/* ICL)

Table 3: Phrase grounding performance on the PadChest-GR dataset for external validation.

Model	Ro/L	Ro/S	F@0.5	P@0.5	R@0.5
MedRPG [3]	31.44	24.00	2.10	2.34	1.91
ChEX [18]	39.73	30.98	3.09	2.97	3.22
GEMeX [14]	27.93	16.10	9.36	15.15	6.77
MedGemma	38.66	27.01	2.03	2.27	1.84
K2Sight [13]	49.81	25.61	8.68	9.46	8.02
Ours *w/* ICL	**63.55**	**48.00**	**25.34**	**25.98**	**24.72**

Fig. 2: Qualitative results with ground truth (dashed) and predictions (solid).



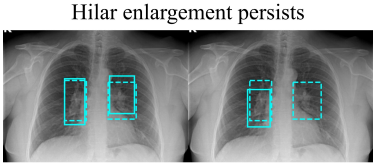
(a) Ours *w/* ICL (b) K2Sight

Table 4: Phrase grounding performance on the PadChest-GR dataset for internal validation.

Model	Ro/L	Ro/S	F@0.5	P@0.5	R@0.5
CheXagent [4]	27.19	14.93	11.30	20.81	7.76
MAIRA-2 [1]	53.57	35.87	32.93	**40.09**	27.94
RadVLM [5]*	70.21	44.85	29.27	27.70	31.03
FT *w/o* Ours	65.54	46.60	28.39	27.81	28.99
FT ***w/*** **Ours**	**70.42**	**47.97**	**33.44**	33.76	**33.14**

* indicates test data seen during training.

Fig. 3: Qualitative results with ground truth (dashed) and predictions (solid).



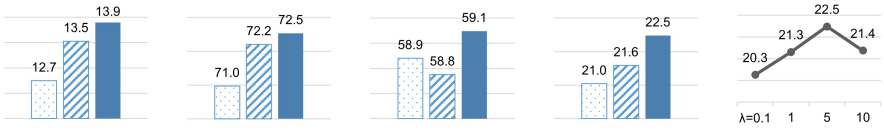
(a) FT *w/* Ours (b) MAIRA-2

against existing chest X-ray grounding models, including MedRPG [3], ChEX [18], GEMeX [14], K2Sight [13], and MedGemma [19].

Table 2 shows the phrase grounding performance on the PadChest-GR dataset for external validation using retrieval-based ICL. Compared to MedGemma's performance without ICL in Table 3, incorporating retrieval-based ICL with contrastive models leads to a significant performance improvement, and additional gains are observed as the number of demonstrations increases. Among these models, our method achieves the best performance, demonstrating its effectiveness in capturing fine-grained representations for real-world settings.

Table 3 shows the phrase grounding performance of chest X-ray grounding models. Although MedRPG and ChEX are designed for grounding abnormal findings, they exhibit limited performance due to inter-observer variability. Similarly, models trained on anatomical-level MIMIC-CXR annotations, including GEMeX and MedGemma (*w/o* ICL), show degraded performance. K2Sight, designed to improve reliability on PadChest-GR through interpretable visual attribute decomposition, similarly shows limited performance (see Fig. 2(b)). Overall, our method with retrieval-based ICL (k=40) outperforms all baselines, highlighting its potential for real-world chest X-ray grounding.

Internal Validation We compared our method (FT *w/* Ours) against chest X-ray grounding LVLMs trained on PadChest-GR, including CheXagent [4],



(a) R@1 (b) R@40 (c) Ro/L (d) F@0.5 (e) F@0.5 (λ)

Fig. 4: (a)-(d) Ablation results of our method using all losses (filled), without $\mathcal{L}_{g,d}$ (hatched), and without both $\mathcal{L}_c$ and $\mathcal{L}_{g,d}$ (dotted). (e) Ablation results for different λ values (line graph).

MAIRA-2 [1], and RadVLM [5]. The performance of RadVLM may be inflated, as the model was trained on the full PadChest-GR dataset. Additionally, we evaluated a variant of our model without MIMIC-CXR pre-training (FT *w/o* Ours) to quantify its impact.

Table 4 shows the phrase grounding performance on the PadChest-GR dataset for internal validation. MAIRA-2 shows relatively higher P@0.5 but lower R@0.5 and Ro/L performance, likely due to its cautious bounding box predictions (see Fig. 3(b)). FT *w/o* Ours outperforms MAIRA-2 on the Ro/L metric but achieves lower performance on F@0.5. By leveraging pretrained parameters, FT *w/* Ours improves F@0.5 by 5% compared to FT *w/o* Ours, surpassing MAIRA-2 on both Ro/L and F@0.5 metrics. Although the performance gains in F@0.5 are marginal compared to MAIRA-2, our model is more parameter-efficient, using 270M parameters, whereas MAIRA-2 uses 7B. In summary, the proposed objective facilitates the learning of pretrained parameters that significantly improve downstream fine-grained grounding performance.

Ablation Studies We performed ablation studies on the retrieval task in MIMIC-CXR and the grounding task in PadChest-GR. We evaluated variants of our method without $\mathcal{L}_{g,d}$ and without both $\mathcal{L}_c$ and $\mathcal{L}_{g,d}$. Additionally, we evaluated our method using different values of λ. Since the dataset contains samples without bounding boxes, we did not consider the variant without $\mathcal{L}_c$. We reported (a) R@1, (b) R@40, (c) Ro/L (k=20), and (d)-(e) F@0.5 (k=20).

Fig. 4(a)-(d) show the ablation results on both tasks. These results suggest that incorporating the location-aware loss is beneficial for fine-grained representation learning. Fig. 4(e) shows performance with varying λ. These results indicate that λ=5 performs best and that assigning a moderately higher weight to the proposed components benefits training. In summary, each loss plays an important role in fine-grained representation learning.

4 Conclusion

We propose LoFi to address the limitations of existing medical contrastive models and LVLMs in chest X-ray retrieval and phrase grounding. By jointly optimizing sigmoid, captioning, and location-aware captioning losses with a lightweight

LLM, our model facilitates learning from long-form text and region-level supervision, enhancing fine-grained representations. Extensive experiments demonstrate that our method consistently achieves superior performance in retrieval on MIMIC-CXR and in phrase grounding on PadChest-GR under both internal and external validation settings. The computational overhead introduced by ICL may limit the scalability of the proposed framework, highlighting the need for future research to alleviate this bottleneck and enable more efficient deployment.

Acknowledgments. The research is supported, in part, by the NSERC Discovery Grant RGPIN-2022-05316, NSERC Alliance Grant ALLRP 602633-24, Tri-Agency Canada; Canada CIFAR AI Chair Awards, and Canada Research Chair Fellowship; Google Gemini Research Awards, IITP grant, the Ministry of Science and ICT (RS-2024-00445087, RS-2025-25464461), funded by the Korea government (MSIT); the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00515536).

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

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
2. de Castro, D.C., Bustos, A., Bannur, S., Hyland, S.L., Bouzid, K., Wetscherek, M.T., Sánchez-Valverde, M.D., Jaques-Pérez, L., Pérez-Rodríguez, L., Takeda, K., et al.: Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. NEJM AI **2**(7), AIdbp2401120 (2025)
3. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G.S.K., Cheng, L.T.E., Thng, C.H., Xu, X., Liu, Y., et al.: Medical phrase grounding with region-phrase context contrastive alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 371–381. Springer (2023)
4. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
5. Deperrois, N., Matsuo, H., Ruipérez-Campillo, S., Vandenhirtz, M., Laguna, S., Ryser, A., Fujimoto, K., Nishio, M., Sutter, T.M., Vogt, J.E., et al.: Radvlm: A multitask conversational vision-language model for radiology. arXiv preprint arXiv:2502.03333 (2025)
6. Gaggion, N., Mosquera, C., Mansilla, L., Saidman, J.M., Aineseder, M., Milone, D.H., Ferrante, E.: Chexmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data **11**(1), 511 (2024)
7. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: Self-supervised learning for medical image analysis: Discriminative, restorative, or adversarial? Medical Image Analysis **94**, 103086 (2024)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)

9. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
10. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25472–25483 (2026)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
12. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11137–11146 (2024)
13. Li, J., Liu, C., Bai, W., Liu, M., Arcucci, R., Bercea, C.I., Schnabel, J.: Knowledge to sight: Reasoning over visual attributes via knowledge decomposition for abnormality grounding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2359–2369 (2026)
14. Liu, B., Zou, K., Zhan, L.M., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.M., Fu, H.: Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21310–21320 (2025)
15. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., et al.: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19724–19735 (2025)
16. Meissen, F., Müller, P., Kaissis, G., Rueckert, D.: Robust detection outcome: A metric for pathology detection in medical images. In: Medical Imaging with Deep Learning. pp. 568–585. PMLR (2024)
17. Müller, P., Jungmann, F., Kaissis, G., Rueckert, D.: A structured, tagged, and localized visual question answering dataset with full sentence answers and scene graphs for chest x-ray images. ICLR (2026)
18. Müller, P., Kaissis, G., Rueckert, D.: Chex: Interactive localization and region description in chest x-rays. In: European Conference on Computer Vision. pp. 92–111. Springer (2024)
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)
21. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
22. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)

23. Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I.M., Wang, X., Susano Pinto, A., Steiner, A., Beyer, L., Zhai, X.: Locca: Visual pretraining with location-aware captioners. Advances in Neural Information Processing Systems **37**, 116355–116387 (2024)
24. Xiao, R., Kim, S., Georgescu, M.I., Akata, Z., Alaniz, S.: Flair: Vlm with fine-grained language-informed image representations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24884–24894 (2025)
25. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3081–3089 (2022)
26. Zambrano Chaves, J.M., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. Nature Communications **16**(1), 3108 (2025)
27. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025)
28. Zhang, T., Zhao, Z., Wu, C., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Radir: A scalable framework for multi-grained medical image retrieval via radiology report mining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 508–518. Springer (2025)