MedGround-R1: Advancing Medical Image Grounding via Spatial-Semantic Rewarded Group Relative Policy Optimization

Huihui Xu[1,3 *], Yuanpeng Nie[4 *], Hualiang Wang[3], Ying Chen[1], Wei Li[1], Junzhi Ning[1], Lihao Liu[1], Hongqiu Wang[5], Lei Zhu[3,5], Jiyao Liu[1,6], Xiaomeng Li[3 †], and Junjun He[1,2 †]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
hejunjun@pjlab.org.cn
[2] Shanghai Innovation Institute, Shanghai, China
[3] The Hong Kong University of Science and Technology, Hong Kong SAR, China
[4] Department of Nephrology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen, China
[5] The Hong Kong University of Science and Technology (Guangzhou), China
[6] Fudan University, Shanghai, China

Abstract. Medical Image Grounding (MIG), which involves localizing specific regions in medical images based on textual descriptions, requires models to not only perceive regions but also deduce spatial relationships of these regions. Existing Vision-Language Models (VLMs) for MIG often rely on Supervised Fine-Tuning (SFT) with large amounts of Chain-of-Thought (CoT) reasoning annotations, which are expensive and time-consuming to acquire. Recently, DeepSeek-R1 demonstrated that Large Language Models (LLMs) can acquire reasoning abilities through Group Relative Policy Optimization (GRPO) without requiring CoT annotations. In this paper, we adapt the GRPO reinforcement learning framework to VLMs for Medical Image Grounding. We propose the **Spatial-Semantic Rewarded Group Relative Policy Optimization** to train the model without CoT reasoning annotations. Specifically, we introduce **Spatial-Semantic Rewards**, which combine spatial accuracy reward and semantic consistency reward to provide nuanced feedback for both spatially positive and negative completions. Additionally, we propose to use the **Chain-of-Box** template, which integrates visual information of referring bounding boxes into the `<think>` reasoning process, enabling the model to explicitly reason about spatial regions during intermediate steps. Experiments on three datasets MS-CXR, ChestX-ray8, and M3D-RefSeg—demonstrate that our method achieves state-of-the-art performance in Medical Image Grounding. Ablation studies further validate the effectiveness of each component in our approach. Code, checkpoints, and datasets are available at https://github.com/bio-mlhui/MedGround-R1

[1] * Equal Contribution
[2] † Corresponding Authors

Keywords: Medical Image Grounding · Spatial-Semantic Rewarded Group Relative Policy Optimization · DeepSeek-R1 · Referring Expression Comprehension

1 Introduction

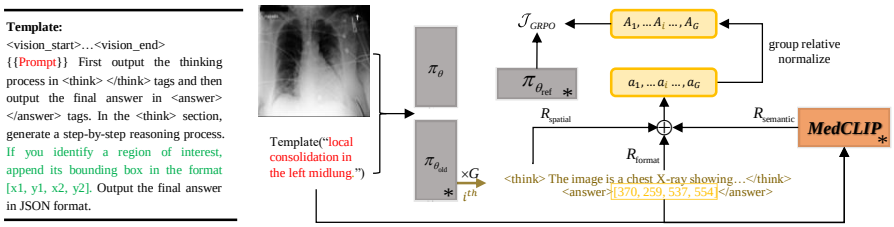
Medical image grounding [6,11,3] is a critical task in computer-aided diagnosis, requiring models to localize specific regions in medical images based on textual descriptions. Unlike general image grounding [21], this task in the medical domain often requires complex reasoning about anatomical structures, pathological features, and their relationships, making it essential for models to not only localize accurately but also reason effectively and meaningfully. For instance, grounding a phrase like "enlarged lymph node near the left lung" requires understanding spatial relationships and medical context, which goes beyond simple pattern recognition in medical images. This reasoning capability is crucial for accurate and interpretable predictions, especially in complex medical scenarios where regions of interest may overlap or have ambiguous boundaries. However, existing methods [6,11,5] often struggle to incorporate such reasoning capabilities, limiting their performance in complex medical tasks.

Existing Vision-Language Models (VLMs)for MIG [8,15] primarily rely on supervised finetuning (SFT) with large amounts of CoT reasoning annotations, which are expensive and time-consuming to acquire. Moreover, these methods typically focus on end-to-end localization without explicitly modelling the reasoning process, leading to suboptimal performance in complex scenarios. For example, they may fail to distinguish between semantically similar pathological regions or misinterpret spatial relationships, especially when dealing with ambiguous or overlapping structures. While recent works [23] attempt to integrate reasoning through multi-stage pipelines, they often require extensive labelled data which are expensive to acquire.

Recently, DeepSeek-R1 [7] and DeepSeek-Math [14] introduced Group Relative Policy Optimization (GRPO), a reinforcement learning framework that enables models to develop reasoning capabilities without the need for Chain-of-Thought (CoT) annotations, which achieves comparable and even better performance than OpenAI o1 series. By leveraging a self-evolution strategy through RL, GRPO allows models to autonomously discover reasoning patterns, such as self-verification and reflection, while significantly reducing the dependency on supervised data. This breakthrough opens up new possibilities for image grounding in the medical context, where reasoning is crucial but labelled data is scarce.

To this end, in this paper, 1) we propose the first RL-based framework for Medical Image Grounding, adapting GRPO technique to train VLMs without CoT annotations. 2) We introduce a combination of **Spatial-Semantic Rewards** to provide nuanced relative evaluations for spatially negative completions. 3) We propose to use the novel **Chain-of-Box** reasoning template, which explicitly integrates visual information of referring bounding boxes into the `<think>` process, allowing the VLMs to reason with both visual and textual context dur-

Fig. 1. A Training Step of the Proposed Spatial-Semantic Group Relative Policy Optimization Framework for MIG. $*$ stands for the model is frozen during training.



ing intermediate steps. 4) We conduct comprehensive experiments on three publicly available medical datasets, including MS-CXR [6], ChestX-ray8 [17], and M3D-RefSeg [1], demonstrating that our method achieves state-of-the-art performance on MIG both quantitatively and qualitatively. Ablation studies further validate the effectiveness of each component.

2 Method

As shown in Fig.1, at each training step, given a medical image and a referring expression, an old policy model π_{old}, saved k steps before, is used to randomly sample a group of G different completions. For each completion, we compute three types of rewards. The Format Reward R_{format} uses regular expression to evaluate whether the completion adheres to the format predefined in the template. The Spatial Reward R_{spatial} calculates the Intersection over Union (IoU) between the predicted bounding box and the ground truth box, which can be seen as the spatial accuracy. The Semantic Reward R_{semantic} leverages a frozen MedCLIP [19] to assess the semantic similarity of the bounding box region of interest (ROI) and the referring expression for each completion. The sum of these three rewards are then normalized to compute each of their *relative advantage* A_i. To preserve the original pretrained knowledge, a frozen reference model π_{ref} is devised. The optimization of the current model is performed using the Group Relative Policy Optimization (GRPO) objective $\mathcal{J}_{\text{GRPO}}$ [7,14]. The green part of the template corresponds to the Chain-of-Box prompt.

It should be noted that unlink previous VLM-based methods [8], our framework does not requires any CoT annotation or reasoning data during training. The ground truth reference box is only required.

2.1 Preliminaries of Group Relative Policy Optimization.

DeepSeek-R1 [7] is the first open-sourced LLM that achieves comparable or even superior performance to closed models like OpenAI's o1 series. The core advance-

ment behind DeepSeek-R1 is Group Relative Policy Optimization (GRPO) [14,7], a reinforcement learning algorithm designed to enhance the reasoning capabilities of LLMs without the requirement of large amounts of CoT annotations.

Unlike approaches that rely on supervised fine-tuning (SFT) or extensive Chain-of-Thought (CoT) annotations [4,18], GRPO leverages the inherent potential of LLMs to self-evolve through a pure reinforcement learning process. By directly applying RL to the base model, GRPO enables the model to autonomously develop sophisticated reasoning behaviors, such as self-verification and reflection.

Formally, given a problem q, GRPO first samples a set of G different completions $\{s_i\}_{i=1}^{G}$ from an old model $\pi_{\theta_{\text{old}}}$ saved k steps before. A reward model R, usually another LLM or any scalar function, is devised to evaluate the *relative advantage* of each solution:

$$r_i = \mathcal{R}(s_i, q), \quad A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}, \tag{1}$$

By denoting $\pi_\theta(s_i|q)$ as the posterior of s_i predicted by π_θ, the following objective is considered to be maximized:

$$\mathcal{J}_{GRPO}(\theta; q) = \mathbb{E}_{\{s_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(S|q)} \frac{1}{G} \sum_{i=1}^{G} A_i \frac{\pi_\theta(s_i|q)}{\pi_{\theta_{old}}(s_i|q)}, \tag{2}$$

where if a completion s_i has stronger relative advantage A_i, π_θ is expected to be more probable to generate s_i than $\pi_{\theta_{\text{old}}}$. To preserve the original pretrained knowledge, a reference model $\pi_{\theta_{\text{ref}}}$, usually the model before RL, is utilized:

$$\mathcal{J}_{GRPO}(\theta; q) = \mathbb{E}_{\{s_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(S|q)} \frac{1}{G} \sum_{i=1}^{G} A_i \frac{\pi_\theta(s_i|q)}{\pi_{\theta_{old}}(s_i|q)} - \beta \mathbb{D}_{KL}\{\pi_\theta(s_i|q)||\pi_{ref}(s_i|q)\}, \tag{3}$$

where β is a hyper-parameter for KL-divergence. To stabilize training, clipping [13,14] is introduced to formulate the final objective:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\{s_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(S|q)} \\ \frac{1}{G} \sum_{i=1}^{G} A_i \left(\min \left(\frac{\pi_\theta(s_i|q)}{\pi_{\theta_{old}}(s_i|q)}, \text{clip} \left(\frac{\pi_\theta(s_i|q)}{\pi_{\theta_{old}}(s_i|q)}, 1-\epsilon, 1+\epsilon \right) \right) \right. \\ \left. -\beta \mathbb{D}_{KL}\{\pi_\theta(s_i|q)||\pi_{ref}(s_i|q)\} \right) \end{aligned} \tag{4}$$

where ϵ is the clipping hyperparameter. The formulation in Eq. 4 enables GRPO to efficiently optimize the policy model by leveraging group-based relative rewards, eliminating the need for a separate critic model [13] and also significantly reducing computational overhead.

2.2 Format Reward for Medical Image Grounding.

Format Reward. Following DeepSeek-R1 [7], the format reward R_{format} ensures the model's completion adheres to a predefined structure. Specifically, the completion must enclose its reasoning process within `<think>` and `</think>` tags, followed by the final answer within `<answer>` and `</answer>` tags. Moreover, in the grounding task, the `<answer>` tag must additionally contain a bounding box in the format `[`x_1`,` y_1`,` x_2`,` y_2`]`. Both rewards are implemented by regular expressions.[7] [8] If a completion matches both patterns, indicating that it adheres to the required format and contains a valid bounding box, the reward is 1. Otherwise, the reward is 0.

2.3 Spatial-Semantic Consistency Rewards

IoU as Spatial Consistency Reward. To evaluate the spatial accuracy of the predicted bounding box for each completion, Intersection over Union (IoU) R_{IoU} between the ground truth box is calculated. If the IoU exceeds a threshold of 0.5, the reward is 1; otherwise, the reward is 0. This reward R_{spatial} can be seen as measuring the spatial consistency of each completion with respect to the ground truth.

Although R_{spatial} is sufficient to evaluate the *spatial* accuracy, it can not evaluate their *semantic* consistency with the ground truth. Our insight is that completions that are spatially-negative maybe semantically-positive, in the sense that some of them may belong to the same semantic class with the ground truth and should therefore be assigned larger advantage than those neither spatially nor semantically positive.

To address this limitation, we introduce the **Semantic Consistency Reward** R_{semantic}. We leverage a frozen MedCLIP [19], to compute the cosine semantic similarity between the cropped bounding box ROI feature of the completion and the referring expression feature. The R_{semantic} reward provides a continuous reward signal that captures the semantic alignment between the predicted region and the referring text, even when the predicted box does not spatially overlap with the ground truth.

2.4 Chain-of-Box Template for Medical Image Grounding

In contrast to DeepSeek-R1 [7] where the reasoning process within `<think>` tags is purely textual information, we propose the **Chain-of-Box** prompt designed for VLMs in medical image grounding task. Our key insight is that the medical image grounding task requires the model to reason about spatial local regions and their relationships in the image during its thinking steps. The reasoning process essentially involves shifting attention to different regions of interest (ROIs).

As illustrated in Fig. 1, whenever the model references a ROI region in the image, it explicitly appends the corresponding bounding box coordinates `[`x_1`,`

[7] `r"<think>.*?</think>\s*<answer>.*?\{.*\[\d+,\s*\d+,\s*\d+,\s*\d+\].*\}.*?</answer>"`

[8] `r"\[(\d+),\s*(\d+),\s*(\d+),\s*(\d+)\]"`

y_1, x_2, y_2] after the region text. This **Chain-of-Box** approach ensures the visual information is seamlessly integrated into the reasoning context, enabling VLMs to perform multimodal reasoning effectively.

3 Experiments

Datasets. We evaluate our method on three publicly available datasets. The first dataset is **MS-CXR** [3], derived from MIMIC-CXR [10], which contains 1,153 image-phrase-box triples. Following MedRPG [6], we preprocess the data to ensure that each phrase query corresponds to a single bounding box, resulting in 890 samples. The second dataset is **ChestX-ray8** [17], a large-scale dataset for diagnosing eight common chest diseases. It includes 984 pathology images with hand-labeled bounding boxes. We adopt the preprocessing approach from MedRPG [6], using category labels as phrase queries to construct image-text-box triplets. The third dataset is **M3D-RefSeg** [1], which consists of 2,778 mask-text-volume triplets extracted from the Totalsegmentator [20] dataset. To adapt for 2D grounding tasks, we extract the frontal slice from each 3D volume and convert the segmentation mask into the bounding box.
Compared Methods. We compare our method against seven recent state-of-the-art approaches, including **LViT** [11], **MedRPG** [6], **CausalCLIPSeg** [5], **ChEX** [12], **GuideDecoder** [22], **RecLMIS** [9], and one VLM-based method **BiRD** [8]. For segmentation-based methods [11,5,22,9], we convert their output masks into bounding boxes to allow comparison. Each method adopts the hyperparameter settings from their original papers. Notably, BiRD [8] is fine-tuned based on Qwen2-VL [16] in its original paper. To ensure a fair comparison, we retrain BiRD based on Qwen2.5-VL [2] same as our setting.
Evaluation Metrics. We follow the protocol established in **MedRPG** [6] to evaluate all methods. Specifically, we report **Accuracy (Acc)**, where a predicted region is considered correct if its Intersection over Union (IoU) with the ground-truth bounding box exceeds 0.5. Additionally, we report the **mean IoU (mIoU)** metric to provide a more comprehensive comparison of localization precision.
Implementation Details. Our model is based on finetuning **Qwen2.5-VL** [2] under the proposed reinforcement learning framework. We use a batch size of 1 per GPU with gradient accumulation over 2 steps. The model is trained for 5K steps with an initial learning rate of 1×10^{-6}. We employ mixed precision training with bfloat16, gradient checkpointing, and Flash Attention to optimize computation efficiency. We set $G = 4$, $\beta = 0.04$ in our experiments. The maximum completion length is set to 256. More ablation studies on hyperparameters are demonstrated in Sec. 3.2.

3.1 Comparisons with SOTA methods

Our method achieves state-of-the-art performance across all three datasets, demonstrating significant improvements over existing approaches. On MS-CXR, we achieve a mean Intersection over Union (mIoU) of **79.02** and an accuracy (Acc)

Table 1. Quantitative comparisons of all methods on three datasets.

Method	Publication	MS-CXR		ChestX-ray8		M3D-RefSeg	
		mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑
Lvit [11]	TMI'23	52.67	66.94	33.50	35.05	40.22	48.79
GuideDecoder [22]	MICCAI'23	56.21	67.04	32.70	34.68	42.19	51.68
MedRPG [6]	MICCAI'23	59.37	69.86	34.59	38.02	41.25	50.35
RecLMIS [9]	TMI'24	62.04	72.49	33.05	40.46	46.71	57.22
ChEX [12]	ECCV'24	61.23	70.15	38.49	41.60	45.83	56.01
CausalCLIPSeg [5]	MICCAI'24	62.77	73.18	40.12	46.02	47.11	58.36
BiRD [8]	MICCAI'24	73.33	76.05	48.22	50.12	52.02	70.18
Ours	– –	**79.02**	**83.12**	**53.12**	**62.18**	**60.10**	**74.66**

of **83.12**, outperforming the previous best method, BiRD, by 5.69 and 7.07 points, respectively. On ChestX-ray8, our method attains an mIoU of **53.12** and an accuracy of **62.18**, surpassing BiRD by 4.90 and 12.06 points. For the M3D-RefSeg dataset, our method achieves an mIoU of **60.10** and an accuracy of **74.66**, outperforming BiRD by 8.08 and 4.48 points. The consistent performance gains across all datasets validate the effectiveness of our proposed reinforcement learning framework, particularly the integration of Spatial-Semantic Rewards and the Chain-of-Box reasoning template, which together enhance the grounding accuracy.

Visual Comparison As shown in Fig. 2, qualitative results further demonstrate the superiority of our method. Visual comparisons reveal that the bounding boxes predicted by our model are significantly more aligned with the ground truth (GT) compared to other methods. This alignment is also evident in cases with complex anatomical structures or ambiguous referring expressions, such as row 1 and 3, where our model's ability to reason spatially and semantically ensures precise localization. For example, in cases where multiple regions of interest overlap, our method consistently identifies the correct region, while other methods often produce inaccurate or overly broad bounding boxes.

Table 2. Number of Sampled Completions.

G	mIoU↑	Acc↑
2	77.40	81.05
4	79.02	83.12
6	80.16	83.43
8	80.77	84.01

Table 3. Reward Modeling.

R_{format}	R_{spatial}	R_{semantic}	mIoU↑	Acc↑
✓	✓	✓	79.02	83.12
×	✓	✓	61.22	68.74
✓	×	✓	2.66	4.30
✓	✓	×	72.14	76.46

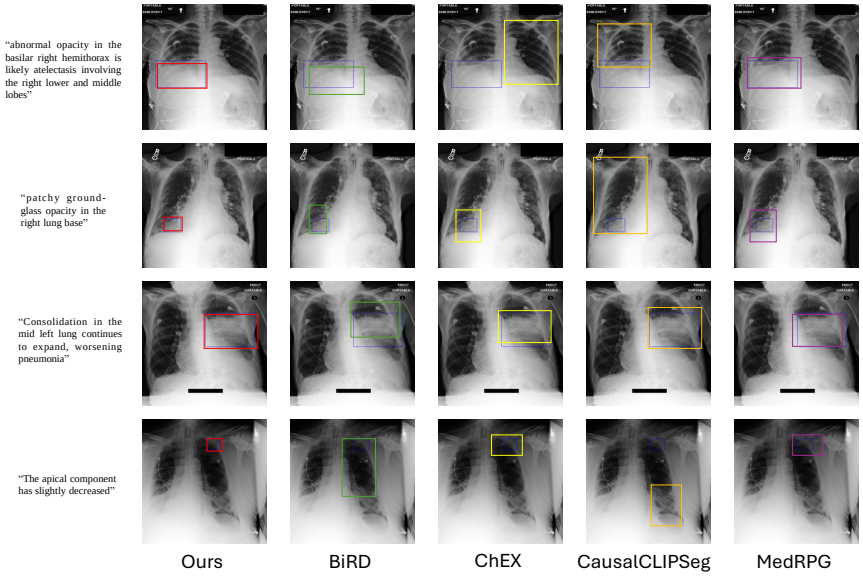
Table 4. Finetuning Regime

Regime	mIoU↑	IoU↑
SFT-1k	68.65	74.18
SFT-3k	56.26	67.50
GRPO-1k	64.25	73.86
GRPO-3k	70.58	75.61
GRPO-5k	77.58	81.09
GRPO-5K w. Chain-of-Box	79.02	83.12

3.2 Ablation Study

Different Setting of G. G stands for the number of sampled completions at each training step. All models are finetuned with 5K steps on MS-CXR. As shown

Fig. 2. Qualitative comparisons on MS-CXR. The blue dashed boxes are ground truth bounding boxes.



in Tab. 2, the performance improves as G increases. However, larger values of G also lead to higher computational costs at each training step, including the time required for sampling and computing rewards for each completion. This trade-off between performance and computational efficiency must be carefully considered..

Different Rewards Combination. R_{format}, R_{spatial}, and R_{semantic} play different role in policy decision. All models are finetuned with 5K steps on MS-CXR. As shown in Tab. 3, removing the format reward R_{format} causes performance drop. We found some outputs are invalid and do not adhere to the standard box pattern. Without spatial reward, which directly evaluates the accuracy between the prediction and the ground truth box, performance collapses significantly. Removing the semantic reward R_{semantic} results in a moderate drop, as it provides nuanced relative evaluations for spatially-negative completions, distinguishing semantically related predictions from unrelated ones. These results confirm that all three rewards are essential and effective for robust performance.

SFT, GRPO, Chain-of-Box Template. As shown in Tab. 4, we finetune the VLM under two regimes: SFT (maximize the next-token probability of the ground-truth box) and GRPO with varying training steps on MS-CXR. Due to the small training set, SFT tends to overfit, as evidenced by the performance drops from 1k to 3k steps. In contrast, GRPO maintains stable performance. Moreover, compared with the original GRPO template used in DeepSeek-R1-Zero, the Chain-of-box template is shown to be effective in improving performance for multimodal reasoning in MIG.

4 Conclusion

In this work, we present a reinforcement learning (RL) framework for Medical Image Grounding, leveraging Group Relative Policy Optimization (GRPO). To adapt DeepSeek-R1 GRPO to VLMs, we introduce two key innovations: (1) a combination of **Spatial-Semantic Rewards** to provide nuanced relative advantage for the spatially-negative completions, and (2) a **Chain-of-Box** prompt template that explicitly integrates visual information into the intermediate reasoning steps. Extensive experiments on three datasets demonstrate the superiority of our method, achieving state-of-the-art performance. Ablation studies further validate the effectiveness of each component.

Acknowledgments. This work was supported by the National Key R&D Program of China (2022ZD0160101, 2022ZD0160102), the National Natural Science Foundation of China (Grant No.62272450), Shanghai Artificial Intelligence Laboratory, a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC), the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24), as well as a grant from the RGC of the Hong Kong Special Administrative Region, China (Project No. R6005-24).

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

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
4. Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms. arXiv preprint arXiv:2412.18925 (2024)
5. Chen, Y., Wei, M., Zheng, Z., Hu, J., Shi, Y., Xiong, S., Zhu, X.X., Mou, L.: Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87. Springer (2024)
6. 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)
7. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
8. Huang, X., Huang, H., Shen, L., Yang, Y., Shang, F., Liu, J., Liu, J.: A refer-and-ground multimodal large language model for biomedicine. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 399–409. Springer (2024)
9. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging (2024)
10. 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)
11. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging **43**(1), 96–107 (2023)
12. 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)
13. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
14. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
15. Wang, L., Wang, H., Yang, H., Mao, J., Yang, Z., Shen, J., Li, X.: Interpretable bilingual multimodal large language model for diverse biomedical tasks. arXiv preprint arXiv:2410.18387 (2024)

16. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
17. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
18. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
21. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
22. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733. Springer (2023)
23. Zou, K., Bai, Y., Chen, Z., Zhou, Y., Chen, Y., Ren, K., Wang, M., Yuan, X., Shen, X., Fu, H.: Medrg: Medical report grounding with multi-modal large language model. arXiv preprint arXiv:2404.06798 (2024)