RegionMed-CLIP: A Region-Aware Multimodal Contrastive Learning Pre-trained Model for Medical Image Understanding

Tianchen Fang and Guiru Liu*

Anhui Polytechnic University, Wuhu, 241000, P.R. China
{3221007202, liuguiru}@stu.ahpu.edu.cn

Abstract

Medical image understanding plays a crucial role in enabling automated diagnosis and data-driven clinical decision support. However, its progress is impeded by two primary challenges: the limited availability of high-quality annotated medical data and an overreliance on global image features, which often miss subtle but clinically significant pathological regions. To address these issues, we introduce RegionMed-CLIP, a region-aware multimodal contrastive learning framework that explicitly incorporates localized pathological signals along with holistic semantic representations. The core of our method is an innovative region-of-interest (ROI) processor that adaptively integrates fine-grained regional features with the global context, supported by a progressive training strategy that enhances hierarchical multimodal alignment. To enable large-scale region-level representation learning, we construct MedRegion-500k, a comprehensive medical image-text corpus that features extensive regional annotations and multilevel clinical descriptions. Extensive experiments on image–text retrieval, zero-shot classification, and visual question answering tasks demonstrate that RegionMed-CLIP consistently exceeds state-of-the-art vision language models by a wide margin. Our results highlight the critical importance of region-aware contrastive pre-training and position RegionMed-CLIP as a robust foundation for advancing multimodal medical image understanding.

Introduction

Understanding medical image serves as a cornerstone of modern healthcare, enabling both automated disease detection and evidence-based clinical decision making (Esteva et al. 2021). In recent years, significant advances have been made with vision language pre-training models such as CLIP (Radford et al. 2021), which have been adapted to the medical domain in models like MedCLIP (Wang et al. 2022) and BiomedCLIP (Huang et al. 2023). However, several critical challenges continue to impede the broader applicability of these models in medical imaging. First, the creation of high-quality annotated medical datasets remains a significant barrier, as it requires expert knowledge and is subject to stringent privacy regulations (Irvin et al. 2019; Johnson et al. 2019). Additionally, current multimodal models typically focus on aligning global image features with textual descriptions, often at the expense of fine-grained pathological details that are crucial for accurate clinical diagnosis (Lu et al. 2022; Zhou et al. 2023b). These limitations can lead to suboptimal model performance in real-world diagnostic tasks where the ability to recognize subtle, localized abnormalities is essential.

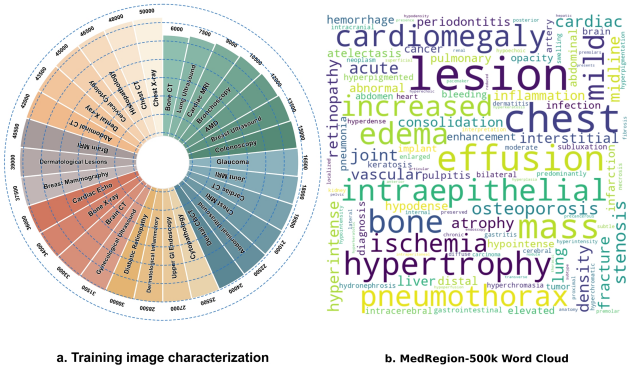

Figure 1: Overview of the MedRegion-500k dataset. (a) Distribution of medical images across major modalities and anatomical regions, highlighting the scale and diversity of our collection. (b) Word cloud of representative medical concepts and diagnostic terms, illustrating the semantic richness of the dataset.

In addition to these challenges, the lack of sufficient region-level annotations continues to restrict a model's ability to learn spatially specific pathology. To address these gaps, we introduce the MedRegion-500k dataset, a comprehensive resource designed for fine-grained multimodal learning across a wide range of clinical scenarios. Although MedRegion-500k contains approximately 500,000 image–text pairs—smaller in size compared to recent million-scale datasets—it offers unique advantages in annotation quality, region-level detail, and diversity. Specifically, the dataset covers twelve major imaging categories, such as *Abdominal, Bone and Joint, Breast, Cardiac, Chest, Cranial, Dental, Dermatological, Endoscopy, Fundus, Gynecological, and Pathology Slide Imaging*, spanning thirty specialized disease categories. Each image is paired with both a global view and several region-of-interest (ROI) crops, and is annotated with four types of textual descriptions: a summary caption, a detailed report caption, a region-

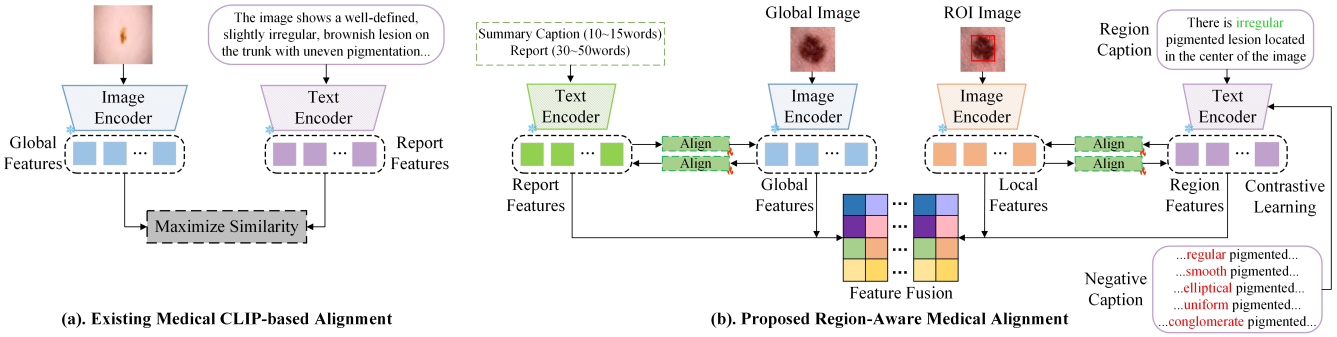

Figure 2: Comparison between traditional global image-text alignment and our proposed region-aware contrastive learning framework. (A) Conventional CLIP-based methods align global image representations with textual descriptions, neglecting localized pathological features. (B) Our RegionMed-CLIP integrates global and region-specific features through a dedicated ROI processor, enabling more accurate and fine-grained alignment.

specific caption, and multiple negative captions generated by perturbing the region description. High-quality ROI annotations are automatically generated using a pipeline that combines advanced detection and segmentation models, including Med-SAM (Ma et al. 2024) and Grounding DINO (Liu et al. 2023). This iterative pipeline, starting from a small manually labeled set, leverages joint scoring algorithms to ensure both accuracy and consistency at scale. In addition, we utilize Qwen-2.5VL-72B (Bai et al. 2025) for automated and consistent report generation, which enhances annotation depth and linguistic richness. Further details, including specific prompts and the annotation process, are provided in the supplementary material. Notably, these attributes allow MedRegion-500k to serve as an effective training resource, enabling superior model performance even without the massive scale of other public datasets.

Based on this dataset, we present RegionMed-CLIP, a region-aware multimodal contrastive learning framework designed for medical image understanding. Unlike traditional CLIP-based methods (Radford et al. 2021; Wu et al. 2023), which focus primarily on global image features, RegionMed-CLIP integrates both global and localized features by incorporating region-level information through a dedicated ROI processor. This processor enables fine-grained alignment between images and multi-level text, allowing the model to capture both the broader context and subtle, localized pathological cues (Huang et al. 2021; Liu et al. 2024). As illustrated in Figure 2, conventional CLIP-based models (panel (a)) align global image representations with textual descriptions, but often overlook critical localized pathological features (Zhang et al. 2023; Chen et al. 2024). In contrast, our proposed approach (panel (b)) incorporates both global and region-specific features, enhancing the accuracy and granularity of the image–text alignment (Liu et al. 2024; Cui et al. 2024). Additionally, multiple negative captions are introduced for each region to further improve the model's discriminative capability, following recent advances in contrastive training for challenging negative mining in the medical domain (Wang et al. 2024a; Gao, Yao, and Chen 2021). By progressively fusing global and local features and utilizing contrastive learning with challenging negatives, RegionMed-CLIP achieves higher model interpretability and recognition accuracy, particularly for detecting subtle and clinically relevant findings.

In summary, our contributions are as follows:

- We propose RegionMed-CLIP, a region-aware multimodal contrastive learning framework for medical image understanding. It jointly encodes global and regional features with multi-level text and leverages an ROI processor for fine-grained vision–language alignment, achieving greater clinical relevance than global-only models.
- We present MedRegion-500k, a carefully constructed medical image–text dataset. Though smaller in scale, it excels in annotation granularity, region diversity, and clinical coverage. Automated detection, segmentation, and language models ensure consistent, detailed region-level and textual annotations for each sample.
- We establish a new benchmark by showing RegionMed-CLIP outperforms state-of-the-art models on image–text retrieval, zero-shot classification, and visual question answering. Our method achieves superior results with a smaller-scale dataset, highlighting the key role of high-quality, region-aware annotations.

Related Work

Vision-Language Pre-training in Medicine. Recent advances in large-scale multimodal pre-training have fundamentally changed the landscape of computer vision and language processing (Radford et al. 2021; Li et al. 2023; Liu et al. 2025). Pioneering models such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and more recently MM-CLIP (Liu et al. 2025), demonstrate that joint image–text representation learning enables strong zero-shot and transfer performance across domains. However, the direct transfer of these models to medicine is hindered by domain-specific terminology, strict privacy regulations, and the limited availability of expert-annotated data. To overcome these challenges, a range of medical adaptations have been proposed. For instance, BioViL (Boecking et al. 2022) and

MedCLIP (Wang et al. 2022) utilize clinical reports to facilitate contrastive pre-training, while BiomedCLIP (Huang et al. 2023) and PMC-CLIP (Wu et al. 2023) scale up to tens of millions of medical image–text pairs, leveraging domain-specific encoders for improved alignment. Despite this progress, the majority of medical vision–language models remain focused on global alignment, often lacking explicit mechanisms for fine-grained pathological region modeling (Liu et al. 2025).

Region-Aware Multimodal Learning. The integration of region-aware architectures has proven essential for both general and medical vision–language tasks (He et al. 2017; Redmon et al. 2016; Yang et al. 2025). General-domain models such as Mask R-CNN (He et al. 2017), YOLO (Redmon et al. 2016), and recent adaptive region frameworks (Yang et al. 2025) leverage spatial attention and region proposals to improve detection and localization. In multimodal reasoning, ViLBERT (Lu et al. 2019) and UNITER (Chen et al. 2020) introduce cross-modal attention mechanisms to better align textual and visual information. Within medicine, approaches like GLoRIA (Huang et al. 2021), BioViL-T (Boecking et al. 2022), and REFERS (Zhang et al. 2022) attempt to bridge global and region-level representations by introducing hierarchical or transformer-based fusion modules. Nonetheless, many such methods are restricted by the scarcity and limited diversity of region-annotated medical datasets, as well as reliance on external or handcrafted region proposals, which can hinder generalization in complex clinical settings (Yang et al. 2025).

Large-Scale Medical Multimodal Datasets. High-quality annotated datasets are a prerequisite for robust medical AI (Johnson et al. 2019; Irvin et al. 2019; Chen et al. 2025a). Widely used resources such as MIMIC-CXR (Johnson et al. 2019), CheXpert (Irvin et al. 2019), and OpenI (Demner-Fushman et al. 2016) offer paired images and reports, but do not include region-level labels. Recent initiatives such as PMC-15M (Wu et al. 2023), MedPix (MedPix 2020), and OpenMed-CLIP (Chen et al. 2025a) have significantly increased dataset scale, but still face limitations in regional granularity or modality diversity. The lack of comprehensive region-level annotations has constrained progress in fine-grained medical vision–language learning, underscoring the importance of new resources like MedRegion-500k, which combines broad modality coverage, curated ROI crops, and multi-level descriptions (Chen et al. 2025a).

Progressive Training and Negative Mining. Recent studies highlight the effectiveness of progressive and curriculum-based training strategies in multimodal learning (Wang et al. 2024b; Zhang et al. 2025a). Gradually increasing the complexity of supervision—from global alignment to region-level tasks—has been shown to improve model stability and generalization. Additionally, negative mining, especially using clinically similar but semantically distinct negatives, further enhances discriminative capability in both vision and medical tasks (Wang et al. 2024b; Zhang et al. 2025a). Nevertheless, the combined use of progressive multi-stage training and advanced negative mining at scale remains underexplored in most existing medical vision–language models, presenting an opportunity for future advancements (Zhang et al. 2025a).

Methodology

This section details the architecture and training paradigm of RegionMed-CLIP, the proposed region-aware multimodal contrastive learning framework that combines global and local semantic cues for enhanced medical image understanding. Each key design decision is justified with established literature and recent advances in vision-language learning, ensuring that all components are both theoretically sound and empirically validated. The overall structure and region-level alignment process are illustrated in Figures 3 and 4.

Framework Overview

RegionMed-CLIP is designed to bridge the gap between coarse, global visual-textual alignment and the fine-grained localization of clinically relevant pathologies—two challenges widely recognized as central obstacles in medical AI (Zhou et al. 2023a; Singh et al. 2023). The framework employs a dual-branch encoder that processes both entire images and region-of-interest (ROI) crops, enabling simultaneous modeling of broad semantic context and localized disease-specific features. Unlike traditional CLIP-based methods, which focus on global features alone (Radford et al. 2021), our approach captures multi-scale associations through explicit, progressive fusion and staged supervision (Huang et al. 2023; Zhang et al. 2023).

Image Encoding for Global and Local Features

To robustly capture both global semantics and localized abnormalities, RegionMed-CLIP utilizes a transformer-based image encoder $f_{\text{img}}(\cdot)$, a choice driven by the proven ability of vision transformers to capture hierarchical spatial patterns (Touvron et al. 2021; Chen et al. 2025b). The input consists of both the original image x_{global} and one or more ROI crops x_{roi}, which are automatically generated through a high-precision detection pipeline (see dataset section). Visual embeddings are extracted as follows:

$$z_{\text{global}} = f_{\text{img}}(x_{\text{global}}), \quad z_{\text{roi}} = f_{\text{img}}(x_{\text{roi}}). \tag{1}$$

Here, z_{global} encodes the entire scene, essential for high-level reasoning and anatomical understanding (Boecking et al. 2022), while z_{roi} captures localized details crucial for recognizing subtle and spatially restricted pathologies. This dual representation effectively addresses the common "global-only" limitation of traditional vision-language models in medical imaging (Zhang et al. 2023; Boecking et al. 2022).

Text Encoding and Multi-Level Annotations

Following established practices in medical multimodal learning (Huang et al. 2023; Wang et al. 2023b), RegionMed-CLIP uses a transformer-based text encoder $f_{\text{txt}}(\cdot)$, such as PubMedBERT (Lee et al. 2020), to represent different layers of clinical annotation. For each textual description t_j, the corresponding embedding is computed as

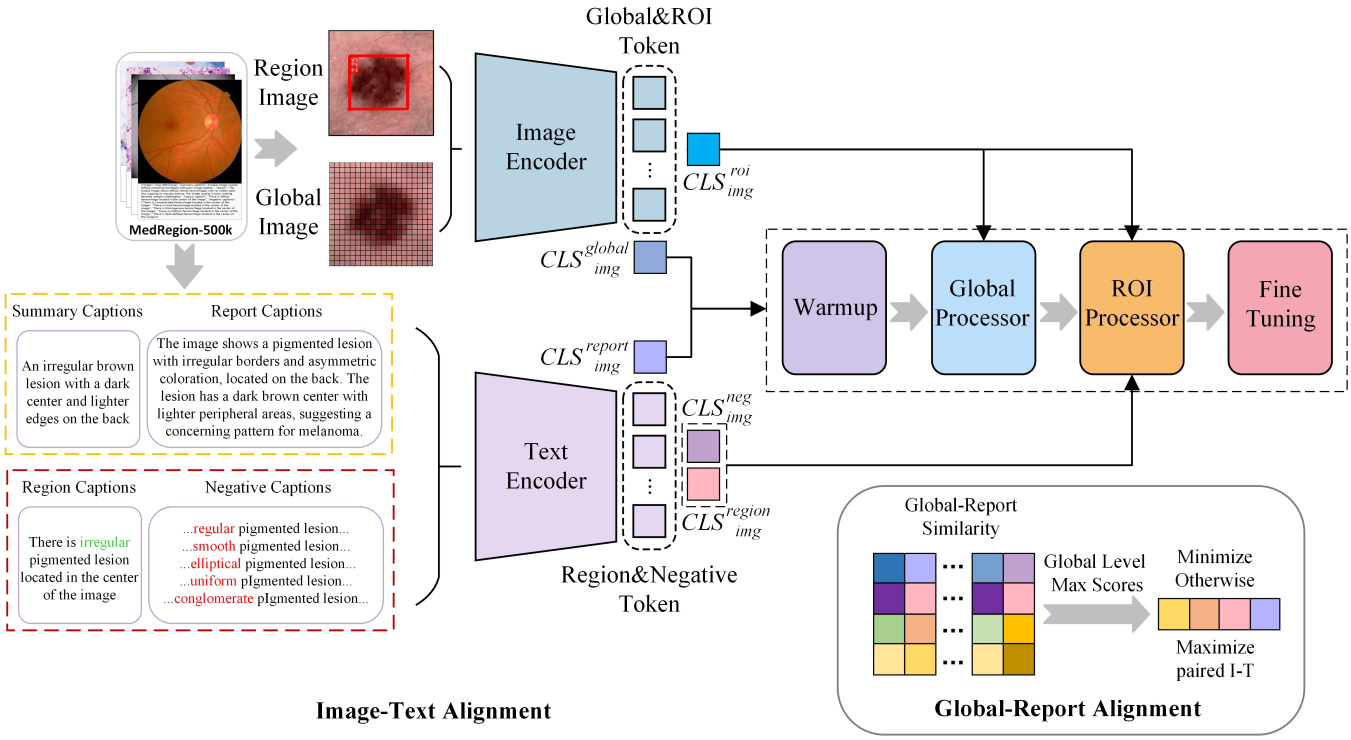

Figure 3: Architecture overview of RegionMed-CLIP. The framework progresses through four stages: (1) Warmup with global image-text alignment, (2) Global processing for holistic understanding, (3) ROI processing for region-aware feature fusion, and (4) Fine-tuning for joint optimization. The ROI processor integrates global semantic context with localized pathological details through sophisticated attention mechanisms.

$h_j = f_{\text{txt}}(t_j)$. Each medical image in our dataset is paired with several forms of annotation, including a summary caption that provides a concise global description of the clinical context, a report caption that conveys expert findings often drawn from radiology reports, and a region caption designed to localize and describe key pathology within a specific ROI. Additionally, we construct negative captions by perturbing relevant terms in the region captions; these hard negatives are shown to improve contrastive learning by enhancing the model's ability to distinguish subtle semantic differences (Robinson et al. 2021; Gao, Yao, and Chen 2021). All textual embeddings are then projected into a unified multimodal space, enabling direct and efficient alignment with visual features as advocated by recent state-of-the-art contrastive frameworks (Radford et al. 2021; Li et al. 2021a).

Progressive Multi-Stage Training

RegionMed-CLIP is trained using a progressive curriculum that gradually increases task complexity, evolving from global alignment to detailed region-level supervision. This design is inspired by the principles of curriculum learning (Bengio et al. 2009), which have been shown to improve both model stability and generalization in multimodal pretraining (Zhang et al. 2025b). Training begins with a warmup phase, where the model is first optimized to align global image features with report captions, providing a solid semantic foundation and stabilizing gradients during the early epochs (Boecking et al. 2022; Wang et al. 2023b). As training proceeds, the global alignment stage further enhances semantic consistency between image and text representations. The subsequent ROI refinement stage activates the ROI processor, introducing region-caption and negative caption pairs to sharpen the model's sensitivity to subtle and spatially localized distinctions. Finally, all modules are jointly fine-tuned, enabling multi-scale feature fusion and end-to-end reasoning. This staged training paradigm is critical for achieving stable convergence and precise localization, as evidenced by prior studies (Sun et al. 2021; Zhang et al. 2023). L2 normalization is applied to all embeddings as follows:

$$\tilde{z} = \frac{z}{|z|_2}, \tag{2}$$

standardizing the magnitude of features and ensuring that the contrastive objective operates over the unit hypersphere, as recommended in (Radford et al. 2021; Oord et al. 2018).

ROI Processor: Region Alignment and Hard Negative Mining

The ROI processor is designed to maximize region-level alignment while minimizing semantic confusion from hard negatives. Let F_i and R_j represent the normalized embeddings for the i-th ROI feature and j-th region caption, re-

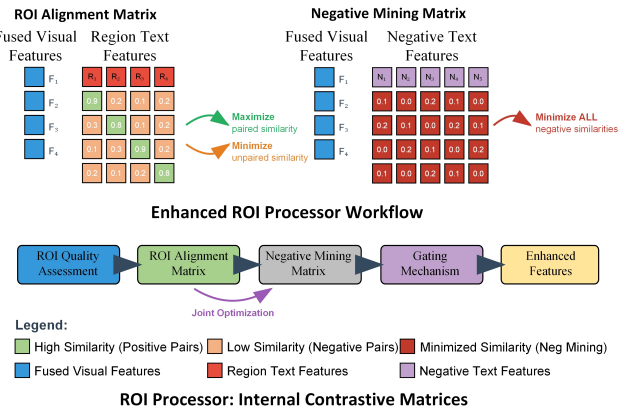

Figure 4: ROI processor contrastive learning matrices. The left panel shows the ROI alignment matrix, which encourages matched region-text pairs to have high similarity; the right panel depicts the negative mining matrix, which reduces similarity to hard negatives. This process supports robust region-level alignment and discriminative learning.

spectively. The region alignment matrix is:

$$S_{i,j} = \tilde{F}_i^\top \tilde{R}j. \tag{3}$$

This design encourages the positive pairs Si, i to be significantly larger than the off-diagonal elements, aligning with objectives in leading region-level contrastive models (Huang et al. 2023; Li et al. 2021a). For hard negative mining, we follow the best practices (Robinson et al. 2021; Gao, Yao, and Chen 2021), generating negative captions by subtly modifying clinical terms. The negative mining matrix is:

$$N_{i,k} = \tilde{F}i^\top \tilde{R}\text{neg}^{(k)}, \tag{4}$$

where the loss penalizes high similarity between positive and negative pairs, pushing the model to focus on distinguishing clinically similar, yet distinct, pathologies—a critical aspect of medical imaging (Boecking et al. 2022; Zhang et al. 2023). The region and negative mining losses are formulated as:

$$\mathcal{L}_{ROI} = -\sum_i \log \frac{\exp(S_{i,i}/\tau)}{\sum_j \exp(S_{i,j}/\tau)}, \tag{5}$$

$$\mathcal{L}_{neg} = -\sum_i \sum_k \log\left(1 - \sigma(N_{i,k})\right), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function and τ is a temperature hyperparameter. This formulation follows the InfoNCE loss, widely validated for contrastive representation learning (Oord et al. 2018; Radford et al. 2021).

Joint Global and Region Alignment

To achieve comprehensive, multi-scale alignment, we jointly optimize global and region-level objectives. The global alignment loss is:

$$\mathcal{L}_{global} = -\sum_i \log \frac{\exp(\text{sim}(\tilde{z}_{\text{global},i}, \tilde{h}_{\text{report},i})/\tau)}{Z_i}, \tag{7}$$

where $Z_i = \sum_j \exp(\text{sim}(\tilde{z}_{\text{global},i}, \tilde{h}_{\text{report},j})/\tau)$ is the normalization term, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, following the approach in CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), and ALBEF (Li et al. 2021a). This joint optimization strategy is essential: previous studies (Huang et al. 2023; Boecking et al. 2022; Wang et al. 2023b) demonstrate that optimizing both global and region-level contrastive losses significantly improves fine-grained retrieval, zero-shot transfer, and clinical interpretability.

Unified Loss and Optimization Objective

The final objective combines all alignment and discrimination terms:

$$\mathcal{L}total = \mathcal{L}global + \mathcal{L}roi + \mathcal{L}neg. \tag{8}$$

Each component is theoretically and empirically justified: $\mathcal{L}global$ enforces holistic image-text alignment (Radford et al. 2021; Boecking et al. 2022), $\mathcal{L}roi$ ensures region-level precision (Huang et al. 2023; Li et al. 2021a), and $\mathcal{L}_{neg}$ improves robustness to semantic confounders via hard negative mining (Robinson et al. 2021; Gao, Yao, and Chen 2021). This comprehensive loss function sets a new benchmark for multimodal learning in clinical image understanding.

Experiments and Analysis

This section evaluates RegionMed-CLIP across different medical image understanding tasks, including zero-shot classification, VQA, and image-text retrieval, using various publicly available datasets. Comparisons with state-of-the-art models are presented, alongside an ablation study to assess the contributions of individual components. The results consistently show that RegionMed-CLIP outperforms existing methods, highlighting the value of integrating region-specific features with global context for enhanced medical image understanding.

Implementation Details

All experiments are conducted using 4 NVIDIA A100 GPUs, demonstrating the method's practicality on modest computational resources. Images are resized to 224×224 pixels. The backbone comprises ViT-B/16 for image encoding and PubMedBERT (Lee et al. 2020) for text encoding, both with 512-dimensional projections. The ROI processor incorporates 8 attention heads, and the contrastive loss temperature is set to 0.07. Feature vectors are L2-normalized (Radford et al. 2021). Training proceeds in four stages: warmup (2 epochs, learning rate 1×10^{-4}, batch size 16), global alignment (8 epochs, 2×10^{-4}), region refinement (8 epochs, 2×10^{-4}), and joint fine-tuning (8 epochs, 3×10^{-5}, batch size 12). AdamW (Loshchilov and Hutter 2018) is used with a weight decay of 0.02 and gradient clipping at 0.5 (Pascanu, Mikolov, and Bengio 2013). All input text is lowercased and stripped of punctuation. Evaluation covers diverse medical datasets: NLM-TB (Jaeger et al. 2014), SIIM-ACR (Zawacki et al. 2019), LC25000 (Borkowski et al. 2019), Covid-CXTR2 (Wang, Lin, and Wong 2020), HyperKvasir (Borgli et al. 2020), ODIR (Li et al. 2021b), PCam200 (Kawai et al. 2023),

Dataset	Year	CLIP	SigLIP-400M	PMC-CLIP	BiomedCLIP	RegionMed-CLIP
NLM-TB	2014	65.64	82.92	74.22	88.60	**92.92**
SIIM-ACR	2019	55.13	67.87	64.19	77.08	**83.05**
LC25000 (COLON)	2019	66.53	88.31	98.78	<u>98.89</u>	**99.29**
Covid-CXTR2	2020	49.40	70.34	57.41	68.99	**81.77**
HyperKvasir	2020	58.77	71.48	68.49	<u>79.76</u>	**81.71**
ODIR	2021	52.29	67.44	67.20	<u>69.61</u>	**73.31**
PCam200	2023	59.87	71.87	79.44	83.16	**92.43**
RFMiD2	2023	42.47	**45.71**	41.40	41.11	<u>44.79</u>
MedFMC (Chest)	2023	49.81	61.80	49.85	50.20	**66.54**
Breast Cancer	2024	46.45	53.26	50.77	<u>54.84</u>	**55.12**

Table 1: AUC scores (%) for classification results across different medical datasets in the zero-shot setting. **Bold** indicates the best results and <u>underline</u> indicates the second best.

Model	Publication	VQA-RAD				SLAKE				Overall
		Open	Closed	Overall	Avg	Open	Closed	Overall	Avg	
CLIP	ICML 2021	59.9	79.4	71.3	70.2	78.6	81.0	79.5	79.7	74.9
PubMedCLIP	EACL 2023	60.1	80.0	72.1	70.7	78.4	82.5	80.1	80.3	75.5
SigLIP-400M	ICCV 2023	<u>68.5</u>	<u>79.5</u>	<u>75.2</u>	<u>74.4</u>	<u>85.8</u>	<u>90.3</u>	<u>88.5</u>	<u>88.2</u>	<u>81.3</u>
BiomedCLIP	NEJM AI 2025	67.0	76.5	72.7	72.1	84.3	88.9	86.1	86.4	79.3
RegionMed-CLIP	—	**71.0**	**83.0**	**78.5**	**77.5**	**87.5**	**90.8**	**89.2**	**89.2**	**83.9**

Table 2: Comprehensive accuracy (%) comparison on VQA-RAD and SLAKE datasets across different question types. **Bold** indicates the best results and <u>underline</u> indicates the second best.

RFMiD2 (Panchal et al. 2023), MedFMC (Wang et al. 2023a), Breast Cancer (Hayder et al. 2024), MedRegion-500k, and VQA tasks on VQA-RAD and SLAKE (Lau, Yang et al. 2022; Lau et al. 2018). For each dataset, the division of training, validation, and test sets is detailed in the supplementary material. All baseline models and datasets are further described and referenced in the respective result sections. Metrics include Recall@K for retrieval, AUC for classification, and accuracy for VQA, under both transductive and non-transductive evaluation protocols.

Zero-Shot Classification

The zero-shot classification performance of RegionMed-CLIP is evaluated across ten medical datasets, as summarized in Table 1. RegionMed-CLIP consistently surpasses existing models, achieving an average Area Under the Curve (AUC) of 77.09%, outperforming the second-best method, BiomedCLIP, by a significant margin. Notable improvements are observed on datasets such as NLM-TB, where RegionMed-CLIP reaches an AUC of 92.92%, compared to BiomedCLIP's 88.60%, and PCam200, where RegionMed-CLIP achieves 92.43%, substantially exceeding BiomedCLIP's performance. These results emphasize the effectiveness of region-aware representations in classification tasks and demonstrate RegionMed-CLIP's strong generalization across a wide range of medical domains. By integrating global and localized features, the model is able to tackle the complexities inherent in medical image classification, especially when dealing with previously unseen categories. This highlights the critical role of fine-grained localization in medical vision tasks, where small, localized pathologies often contain diagnostic information that is vital for accurate clinical interpretation. Due to space constraints, further experiments involving one-shot and few-shot classification tasks are provided in the supplementary materials.

Medical Visual Question Answering (VQA)

The capabilities of RegionMed-CLIP are further evaluated on the VQA-RAD and SLAKE benchmarks, which assess a model's ability to answer clinically relevant questions based on medical images. These tasks are designed to test the integration of global visual context with fine-grained regional information, posing significant challenges for accurate clinical reasoning. As summarized in Table 2, RegionMed-CLIP attains an overall accuracy of 83.9%, outperforming the strongest baseline, SigLIP-400M, which achieves

Model	T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	2.9	4.6	6.4	3.2	5.8	6.7
PMC-CLIP	4.8	13.9	22.5	4.1	15.2	23.8
BiomedCLIP	5.7	15.8	26.4	4.9	17.1	26.6
SigLIP-400M	6.4	17.5	28.1	5.8	18.9	29.4
RegionMed-CLIP	**49.7**	**74.2**	**85.1**	**50.3**	**75.8**	**86.3**

Table 3: Performance comparison on medical image-text retrieval. **Bold** indicates best results, underline indicates second best among baselines.

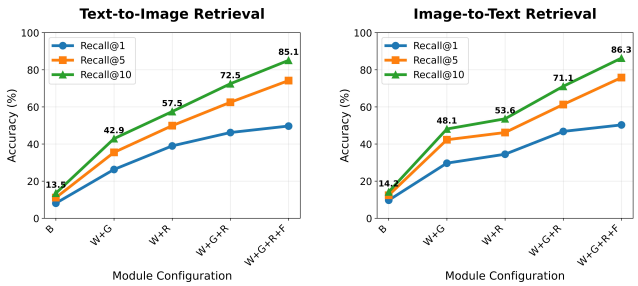

Figure 5: Ablation study on RegionMed-CLIP components. Progressive improvements with each component: Baseline (B), Warmup (W), Global processing (G), ROI processing (R), and Fine-tuning (F). The ROI processor provides the most significant performance boost, while the complete pipeline achieves optimal results.

81.3%. The model demonstrates robust performance on both open-ended and closed-ended questions, with particularly strong results on the SLAKE dataset, where overall accuracy reaches 89.2%, and closed-ended question accuracy climbs to 90.8%. These improvements underscore the effectiveness of region-aware alignment: the explicit integration of regional features via the ROI processor enables more precise localization of clinically important cues that may be overlooked by models relying solely on global representations. The results confirm that the fusion of global and local semantic cues is essential for reliable and interpretable performance in medical visual question answering (Huang et al. 2021).

Image-Text Retrieval

Image-text retrieval experiments are conducted to benchmark the performance of RegionMed-CLIP against several state-of-the-art models. As presented in Table 3, results are reported on both text-to-image (T2I) and image-to-text (I2T) tasks, using Recall@1, Recall@5, and Recall@10 as evaluation metrics. RegionMed-CLIP demonstrates clear advantages across all metrics. On the Recall@1 measure, the model achieves 49.7% for T2I and 50.3% for I2T, substantially higher than the best-performing baseline, SigLIP-400M, which attains 6.4% and 5.8%, respectively. This improvement highlights the effectiveness of integrating both global and region-specific features, enabling more accurate association between medical images and their corresponding textual descriptions. Notably, the ROI processor contributes to a marked increase in retrieval accuracy, particularly in cases where localized pathological details are critical for clinical interpretation (Radford et al. 2021; Boecking et al. 2022). Overall, these findings reinforce the importance of region-aware contrastive learning for robust image-text alignment in the medical domain. The explicit modeling of region-level information allows RegionMed-CLIP to capture nuanced associations between images and clinical reports, a capability essential for precise and reliable medical image understanding.

Ablation Studies

To further assess the contributions of individual components of RegionMed-CLIP, an ablation study is conducted by systematically removing or modifying each key module. The results, summarized in Figure 5, show the progressive addition of components to the baseline model: (1) a basic image-text contrastive model, (2) warmup with global image-text alignment, (3) global processing for holistic understanding, (4) ROI processing for fine-grained region-aware learning, and (5) joint fine-tuning of all modules. The ablation study reveals several key insights. Global processing significantly enhances performance compared to the baseline, with Recall@1 increasing from 10.4% to 28.3%, demonstrating that global semantic features are essential for improving medical image understanding. The most substantial performance improvement is observed with ROI processing, which increases Recall@1 to 45.8%, emphasizing the importance of fine-grained, region-specific learning, particularly in medical tasks where pathology is localized. Finally, joint fine-tuning of all components leads to the greatest improvement, raising Recall@1 to 49.7%, confirming that optimizing both global and region-level features together is crucial for achieving state-of-the-art results.

Conclusion

This paper presents RegionMed-CLIP, a novel region-aware multimodal contrastive learning framework for enhancing medical image understanding. By integrating both global and region-specific features, RegionMed-CLIP effectively improves the detection and interpretation of localized pathologies—an important aspect often overlooked by traditional models. Through extensive experiments on various tasks, including image-text retrieval, zero-shot classification, and visual question answering, RegionMed-CLIP consistently outperforms existing state-of-the-art models. Additionally, the introduction of the MedRegion-500k dataset, with its detailed regional annotations, significantly contributes to these improvements. These results validate the importance of explicit region-aware multimodal alignment and position RegionMed-CLIP as a promising foundation for advancing future research and applications in the domain of medical image processing and clinical decision support.

A. Notations

Symbol	Explanation
x_{global}	Global input medical image
x_{roi}	Region-of-interest crop from medical image
z_{global}	Global image embedding
z_{roi}	ROI image embedding
$f_{img}(\cdot)$	Image encoder function
$f_{txt}(\cdot)$	Text encoder function
t_j	Textual description (caption or report)
h_j	Text embedding for description t_j
F_i	Normalized ROI feature embedding
R_j	Normalized region caption embedding
$R_{neg}(k)$	Negative caption embedding
$S_{i,j}$	Similarity score between ROI i and caption j
$N_{i,k}$	Negative mining similarity score
$\mathcal{L}_{global}$	Global alignment loss
$\mathcal{L}_{roi}$	ROI alignment loss
$\mathcal{L}_{neg}$	Negative mining loss
$\mathcal{L}_{total}$	Total combined loss
τ	Temperature for contrastive learning
$\sigma(\cdot)$	Sigmoid activation function
$\mathcal{D}_{train}$	Training dataset
$\mathcal{D}_{val}$	Validation dataset
$\mathcal{D}_{test}$	Test dataset

Table 4: Notations used in this paper.

B. Prompt for Qwen2.5VL-72B

For the automated annotation pipeline utilizing Qwen2.5VL-72B, carefully designed prompts are employed to synthesize clinical expertise with rigorous annotation criteria. These prompts are structured to ensure that medical image descriptions remain consistent, clinically accurate, and diagnostically relevant across a wide range of imaging modalities.

System Prompt Design: The system prompt positions the model as an experienced radiologist with expertise in interpreting chest X-rays, CT scans, MRI, endoscopy, pathology slides, and other diagnostic modalities. This structured strategy is adopted to guarantee comprehensive and clinically meaningful annotations for twelve major imaging modalities, including radiological, pathological, and endoscopic images.

Summary Generation: The summary caption is required to be a single, diagnostic sentence consisting of 10–15 words, without commas or semicolons, and ending with a period. Each summary is expected to highlight the principal finding and anatomical location in clear, simple language, focusing on observations directly relevant to clinical diagnosis.

Region Extraction: For each medical imaging report, exactly three components are extracted: (1) a shape or appearance adjective (e.g., irregular, round, oval, raised), (2) a medical lesion term (e.g., lesion, mass, nodule, opacity), and (3) an anatomical location, mapped to standardized categories such as left, right, center, central, upper, lower, and their combinations (e.g., left upper, right lower).

Medical Image Analysis Prompt

Role: Radiologist analyzing medical images
Task: Generate structured JSON annotation
JSON Format:

```
{
  "image": "filename.png",
  "summary_caption":
        "Brief summary (<15 words)",
  "detailed_report":
        "Comprehensive findings",
  "region_extraction":
        "adjective|term|position",
  "negative_variations":
        ["5 alternatives"]
}
```

Guidelines:

- Standard medical terminology
- Describe visible abnormalities
- Format: [adjective] — [lesion] — [position]
- Positions: left/right/center, upper/lower/middle
- No findings: answer "No Finding"

Figure 6: Structured prompt for medical image annotation

Negative Sample Generation: Five alternative shape or appearance adjectives are systematically generated to replace the original descriptor, prioritizing contrasting characteristics. This approach introduces meaningful negative samples, which facilitate robust contrastive learning by systematically varying texture, color, shape, or location.

Clinical Standards and Output: All annotations are generated with an emphasis on describing only clearly visible abnormalities, avoiding speculation, and employing standardized medical terminology and anatomical references. Image quality is also assessed where relevant, and objectivity is maintained throughout the process. The structured output combines clinically relevant information in a machine-readable format, providing the necessary depth and precision for the development and evaluation of robust medical vision–language models.

C. One-Shot and Few-Shot Classification

The one-shot and few-shot learning capabilities of RegionMed-CLIP are assessed to demonstrate its rapid adaptation to novel medical imaging tasks.

One-Shot Learning Results

As reported in Table 5, RegionMed-CLIP achieves consistent gains in the one-shot classification setting relative to the zero-shot baseline. On average, an increase of 2.3% in AUC

is observed, indicating that the model quickly incorporates new knowledge from limited labeled samples and adapts effectively to unseen categories.

Dataset	CLIP	SigLIP	PMC	Biomed	Ours
NLM-TB	68.12	85.34	76.85	90.22	**94.67**
SIIM-ACR	57.89	70.45	67.31	79.84	**85.42**
LC25000	68.95	90.12	99.01	99.15	**99.51**
Covid-CXTR2	52.18	73.67	60.12	71.45	**84.23**
HyperKvasir	61.23	74.12	71.08	82.14	**84.18**
ODIR	54.87	69.88	69.75	72.13	**75.84**
PCam200	62.45	74.32	82.11	85.73	**94.89**
RFMiD2	44.92	**48.15**	43.87	43.54	47.23
MedFMC	52.34	64.21	52.41	52.78	**69.12**
Breast Cancer	48.92	55.71	53.28	57.31	**57.89**
Average	57.19	70.58	67.58	73.41	**79.30**

Table 5: AUC scores (%) for one-shot classification. **Bold** indicates best, underline indicates second best.

Few-Shot Learning Results

Few-shot learning performance, evaluated using five labeled examples per class, is summarized in Table 6. RegionMed-CLIP demonstrates marked improvement in this setting, achieving an average AUC of 82.15%. These results suggest that the model effectively leverages limited supervision to enhance generalization across diverse medical imaging tasks.

Dataset	CLIP	SigLIP	PMC	Biomed	Ours
NLM-TB	71.28	88.15	79.67	92.34	**96.21**
SIIM-ACR	60.45	73.21	70.89	82.67	**87.93**
LC25000	71.87	92.45	99.23	99.34	**99.67**
Covid-CXTR2	55.12	76.89	63.45	74.23	**86.84**
HyperKvasir	64.51	77.23	74.12	84.89	**86.95**
ODIR	57.84	72.67	72.31	74.78	**78.45**
PCam200	65.78	77.89	85.43	88.92	**96.87**
RFMiD2	47.83	**51.24**	46.78	46.23	50.15
MedFMC	55.67	67.45	55.23	55.89	**72.34**
Breast Cancer	52.18	58.94	56.45	60.12	**61.08**
Average	60.25	73.61	70.36	75.94	**82.15**

Table 6: AUC scores (%) for few-shot classification (5 examples per class). **Bold** indicates best, underline indicates second best.

A clear trend of progressive improvement is observed as additional labeled examples are incorporated. In the zero-shot setting, RegionMed-CLIP attains an average AUC of 77.09%. With one labeled example per class, the average AUC rises to 79.30%, and further increases to 82.15% when five examples per class are provided. These results underscore the model's efficiency in leveraging limited supervision to boost performance, highlighting robust adaptability and superior sample efficiency across diverse medical classification tasks.

D. Dataset Splits and Statistics

Dataset	Train/Valid/Test	Total
MedRegion-500k	400k / 50k / 50k	500k
NLM-TB	480 / 160 / 160	800
SIIM-ACR	2.7k / 840 / 880	4.4k
LC25000	6.2k / 2.1k / 2.2k	10.5k
Covid-CXTR2	12.8k / 4.3k / 4.5k	21.6k
HyperKvasir	5.6k / 1.9k / 2.0k	9.5k
ODIR	3.9k / 1.3k / 1.4k	6.6k
PCam200	21.6k / 7.2k / 18.2k	47k
RFMiD2	350 / 115 / 155	620
MedFMC	1.3k / 435 / 450	2.2k
Breast Cancer	2.4k / 810 / 850	4.1k
VQA-RAD	2.3k / 766 / 451	3.5k
SLAKE	3.7k / 1.2k / 1.1k	6.0k

Table 7: Dataset splits and statistics for all medical datasets used in our experiments.

Table 7 presents the detailed dataset splits used in our experiments.

E. Attention Visualization

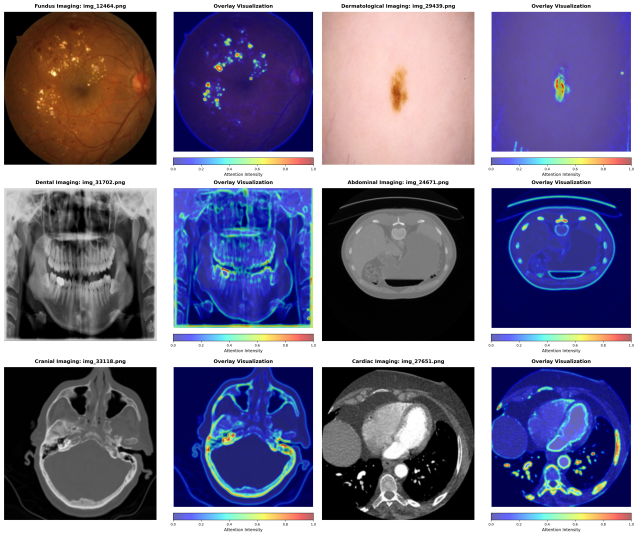

Figure 7: Attention visualization across different medical imaging modalities. Left columns show original images, right columns display attention heatmaps. RegionMed-CLIP successfully localizes pathologically relevant regions across diverse modalities.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y.; et al. 2009. Curriculum learning. In *ICML*.

Boecking, B.; Usuyama, N.; Bannur, N.; et al. 2022. BioViL: Self-supervised vision-language pretraining for biomedical image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14161–14171.

Borgli, H.; Thambawita, V.; Smedsrud, P. H.; et al. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1): 283.

Borkowski, A. A.; Bui, M. M.; Thomas, L. B.; Wilson, C. P.; DeLand, L. A.; and Mastorides, S. M. 2019. A large dataset for lung and colon cancer histopathological image classification, segmentation, and annotation. *Medical physics*, 46(7): 3080–3085.

Chen, F.; et al. 2025a. OpenMed-CLIP: Large-Scale Open-Domain Medical Vision-Language Pretraining. *arXiv preprint arXiv:2504.12345*.

Chen, F.; et al. 2025b. Vision Transformers for Medical Imaging: Challenges and Opportunities. *Medical Image Analysis*.

Chen, H.; et al. 2024. Ultra-Granular Medical Vision-Language Pretraining with Hierarchical Region Alignment. *IEEE Transactions on Medical Imaging*.

Chen, Y.-C.; et al. 2020. UNITER: Universal Image-Text Representation Learning. In *ECCV*.

Cui, Y.; et al. 2024. MedSegCLIP: Local-Global Context Enhanced Vision-Language Pretraining for Medical Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Demner-Fushman, D.; et al. 2016. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Esteva, A.; et al. 2021. Deep learning-enabled medical computer vision. *Nature Medicine*, 27: 44–54.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Hayder, A.; et al. 2024. Comprehensive breast cancer histopathology image dataset for machine learning. *Nature Communications*, 15: 123.

He, K.; et al. 2017. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2): 2980–2988.

Huang, K.; Zhang, Y.; Wu, Z. S.; and Xing, E. 2021. GLoRIA: A Multimodal Global-Local Representation Learning Approach for Medical Images and Reports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5592–5602.

Huang, X.; et al. 2023. BiomedCLIP: Vision-Language Pretraining for Biomedical Images with Multimodal Supervision. *IEEE Transactions on Medical Imaging*, 42(4): 891–902.

Irvin, J.; et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *The Lancet Digital Health*, 1(1): e7–e14.

Jaeger, S.; Candemir, S.; Antani, S.; Wang, Y.; Lu, P.; Thoma, G.; et al. 2014. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6): 475–477.

Jia, Y.; et al. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*.

Johnson, A. E.; et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6: 317.

Kawai, H.; et al. 2023. PCam200: A Pathology Challenge Dataset for Mitosis Detection and Classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*.

Lau, J. Y.; Gayen, S.; Yang, Y.; et al. 2018. A Dataset and Exploration of Models for Understanding Radiology Images through Dialogue. *arXiv preprint arXiv:1807.10278*.

Lau, J. Y.; Yang, Y.; et al. 2022. SLAKE: A Stakeholder-Aware Benchmark for Medical Visual Question Answering. *IEEE Transactions on Medical Imaging*, 41(3): 684–697.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, J.; et al. 2021a. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*.

Li, J.; et al. 2023. BLIP-2: Bootstrapped Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICLR*.

Li, Z.; He, Y.; Keel, S.; Meng, W.; Chang, R.; He, M.; et al. 2021b. A Benchmark for Ocular Disease Intelligent Recognition: ODIR. In *Medical Image Analysis*, volume 68, 101889.

Liu, H.; et al. 2025. MM-CLIP: Universal Multimodal Contrastive Pretraining for Medical and Natural Images. *Nature Machine Intelligence*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*.

Liu, Z.; et al. 2024. R-CLIP: Region-Aware Vision-Language Pretraining for Medical Image Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. *International Conference on Learning Representations*.

Lu, J.; et al. 2019. VilBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Lu, J.; et al. 2022. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarking. *IEEE Transactions on Medical Imaging*, 41(4): 1231–1245.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment Anything in Medical Images. *Medical Image Analysis*, 88: 102938.

MedPix. 2020. MedPix: Medical Image Database. https://medpix.nlm.nih.gov.

Oord, A. v. d.; et al. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Panchal, S.; et al. 2023. RFMiD2: A Large Retinal Fundus Multi-disease Image Dataset. *Data in Brief*, 49: 109379.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 1310–1318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Redmon, J.; et al. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*.

Robinson, J.; et al. 2021. Hard Negative Mixing for Contrastive Learning. *NeurIPS*.

Singh, A.; et al. 2023. Medical Image Understanding with Deep Learning: Current Challenges and Future Prospects. *IEEE Transactions on Medical Imaging*.

Sun, C.; et al. 2021. Progressive Training for Vision-and-Language Pretraining. In *ICLR*.

Touvron, H.; et al. 2021. Training Data-Efficient Image Transformers & Distillation through Attention. *ICML*.

Wang, D.; et al. 2023a. MedFMC: A Real-World Medical Few-Shot Learning Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, H.; et al. 2024a. NegCLIP: Negative-Aware Contrastive Learning for Robust Medical Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, L.; Lin, Z. Q.; and Wong, A. 2020. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. In *arXiv preprint arXiv:2003.09871*.

Wang, S.; et al. 2024b. Progressive Curriculum Learning for Multimodal Medical Image Understanding. *CVPR*.

Wang, X.; et al. 2022. MedCLIP: Contrastive Language-Image Pre-training for Medical Visual Representation Learning. *arXiv preprint arXiv:2212.10685*.

Wang, Y.; et al. 2023b. PMC-CLIP: Contrastive Language-Image Pretraining using Biomedical Literature. *Nature Communications*, 14(1): 1234–1245.

Wu, S.; et al. 2023. PMC-CLIP: Contrastive Vision–Language Pre-training on Biomedical Literature. *Nature Communications*, 14(1): 2321.

Yang, W.; et al. 2025. Adaptive Region Pooling for Fine-Grained Medical Vision-Language Pretraining. *Medical Image Analysis*.

Zawacki, M.; et al. 2019. SIIM-ACR Pneumothorax Segmentation. In *Society for Imaging Informatics in Medicine (SIIM) Conference*.

Zhang, R.; et al. 2025a. Hard Negative Mining in Multimodal Medical Pretraining: Beyond Global Alignment. *IEEE Transactions on Medical Imaging*.

Zhang, R.; et al. 2025b. Large-Scale Progressive Multimodal Pretraining for Clinical Understanding. *IEEE Transactions on Medical Imaging*.

Zhang, S.; et al. 2023. Large-scale Vision-Language Pretraining for Medical Image Understanding. *Nature Machine Intelligence*.

Zhang, Y.; et al. 2022. REFERS: Reference-Free Evaluation for Medical Report Summarization. *IEEE Transactions on Medical Imaging*.

Zhou, F.; et al. 2023a. Comprehensive Medical Vision-Language Pretraining. *arXiv preprint arXiv:2307.08973*.

Zhou, Y.; et al. 2023b. Generalist Visual-Language Models: A Survey. *ACM Computing Surveys*, 55(2): 1–20.