

Cross-domain Adaptation of Vision-Language Foundation Models for Medical Applications

Abstract

Vision-language models that understand interactions between medical images and text reports could enable various useful applications in healthcare. However, developing such medical models is challenging due to insufficient training data, unlike natural images. The study examines different pretrained VL models, including those trained on general image-text data (coarse-grained) and those finetuned on object detection (fine-grained), as starting points. It then explores adapting these natural models to the medical domain using self-supervised pre-training on medical image-text data and finetuning on medical report generation. The adapted models are evaluated on retrieval and report generation tasks. Results show both pre-training and finetuning assist adaptation, with coarse-grained natural models demonstrating the best transferability. The investigation provides guidance for harnessing VL models across domains with limited data.

Keywords: Vision-language understanding, cross-domain adaptation, medical report generation, foundation models

1. Introduction

Healthcare in the era of precision medicine and personalized treatment increasingly relies on harnessing complex, multi-modal medical data, including medical images, radiology reports, and electronic health records (Acosta et al., 2022; Moor et al., 2023a; Zhang et al., 2022; Moor et al., 2023b; Tu et al., 2023). Understanding the interactions within and across these data modalities is critical for improving patient outcomes through applications such as automated report generation (Huang et al., 2021), image-text retrieval for clinical decision support (Jeong et al., 2023; Boecking et al., 2022), and zero-shot disease detection (Tiu et al., 2022; Wang et al., 2022; Mishra et al., 2023).

Recent years have seen promising advances in self-supervised representation learning for medical images

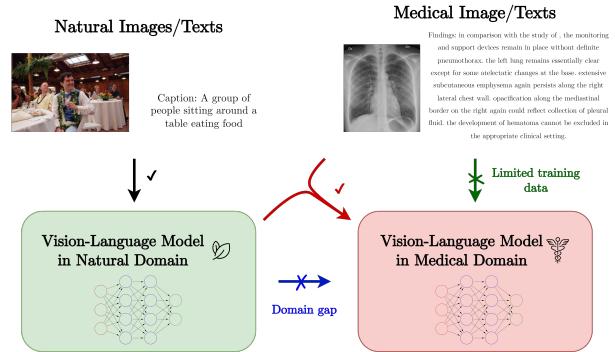


Figure 1: The framework for our study, which aims to answer a critical question: Can vision-language models, initially trained on large-scale natural image and text datasets, be effectively adapted for medical applications where data is often limited? We approach this question by exploring various types of representations, both coarse-grained and fine-grained, and employing adaptation strategies such as medical pre-training and finetuning.

and text, particularly through pre-training on paired chest X-rays and radiology reports using contrastive learning objectives (Tiu et al., 2022; Zhang et al., 2022). The pretrained models are then finetuned on downstream tasks like disease classification, report generation, and retrieval (Jeong et al., 2023; Huang et al., 2021). However, most prior works have focused on contrastive loss for pretraining, while some recent works have started examining the effects of other objectives such as masked language modeling loss (Boecking et al., 2022) or multi-granularity contrastive loss (Wang et al., 2022).

In the natural image and text domains, an explosion of research on vision-language (VL) modeling has demonstrated the remarkable adaptability of these models to a diverse array of downstream tasks (Li et al., 2022; Dou et al., 2022). By pretraining on massive datasets using varied self-supervised losses, VL models have shown an impressive ability to generalize. However, it remains an open question whether

and how such natural domain VL models, which harness orders of magnitude more data, can transfer to medical settings where limited datasets are the norm (Zhang et al., 2023; Xu et al., 2023a,b).

Specifically, our contributions are as follows:

- We examine the generalization capabilities of different VL models originally designed for natural domains when adapted to medical tasks. Our experiments use the MIMIC-CXR dataset, which contains a large number of chest X-ray images and associated radiology reports, to evaluate the models on zero-shot image-report retrieval and medical report generation tasks.
- We assess the comparative effectiveness of adaptation approaches, including continued pretraining and finetuning, for enabling these cross-domain transfers. Our results indicate that both medical pre-training and fine-tuning significantly improve performance over models trained from scratch. Notably, coarse-grained pre-trained models outperform fine-grained models in medical tasks, likely due to their ability to capture more general global representations.
- We evaluate the performance of adapted models on critical medical applications using both similarity-based metrics like BLEU and clinically-focused metrics like RadCliQ (Yu et al., 2022). Our findings reveal that adapted models outperform those trained from scratch in medical report generation. However, the performance gaps are smaller for report generation, suggesting that fine-tuning during this supervised task may lead to a loss of some pre-trained representations.
- Our study also uncovers nuanced findings in the performance of coarse-grained versus fine-grained models. Coarse-grained models excel in zero-shot retrieval tasks and medical report generation, indicating their broader applicability in the medical domain. Fine-grained models, although less effective in our medical tasks, offer insights into the limitations of current adaptation strategies.

By rigorously addressing these questions on model adaptability, adaptation strategies, and evaluation on priority medical tasks, this study generates valuable insights and guidance for deploying data-intensive VL models in limited-data domains like healthcare. Our findings offer a pathway towards harnessing state-of-the-art VL methods to create automated solutions that can enhance clinical workflows and improve patient care.

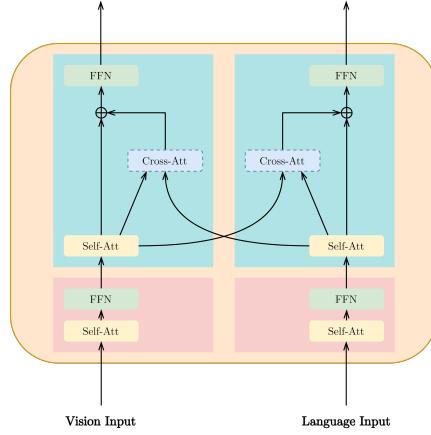


Figure 2: A simplified representation of the FIBER architecture Dou et al. (2022) consisting of vision and language backbones with injected cross attention modules for deep multimodal fusion.

2. Methods

Figure 3 outlines our approach to exploring the issue of cross-domain adaptation in vision-language models. Initially, we examine a range of vision-language models that have been pre-trained on extensive datasets of natural images and text. These models are designed to learn various types of representations, both at coarse-grained and fine-grained levels (Dou et al., 2022). “Coarse-grained” refers to global, image-level tasks and representations, whereas “fine-grained” refers to more localized region-level image-text tasks and representations. We refer to these models as **natural domain vision-language models**. Their focus on different aspects of representation learning, as defined by these granularity levels, may affect how well they generalize to the medical domain.

For the **medical adaptation** phase, we employ different strategies for cross-domain adaptation. These include: (1) *medical pre-training*, where we use a variety of self-supervised loss functions to adapt both the image and text encoders using medical images and reports; and (2) *medical fine-tuning*, where the image encoder and text decoder are adapted through a generative task that employs an image captioning loss function.

In this section, we delve into the technical specifics of both the natural domain vision-language models (section 2.2) and the methods used for medical adaptation. We begin by detailing the backbone architecture used for learning vision-language representa-

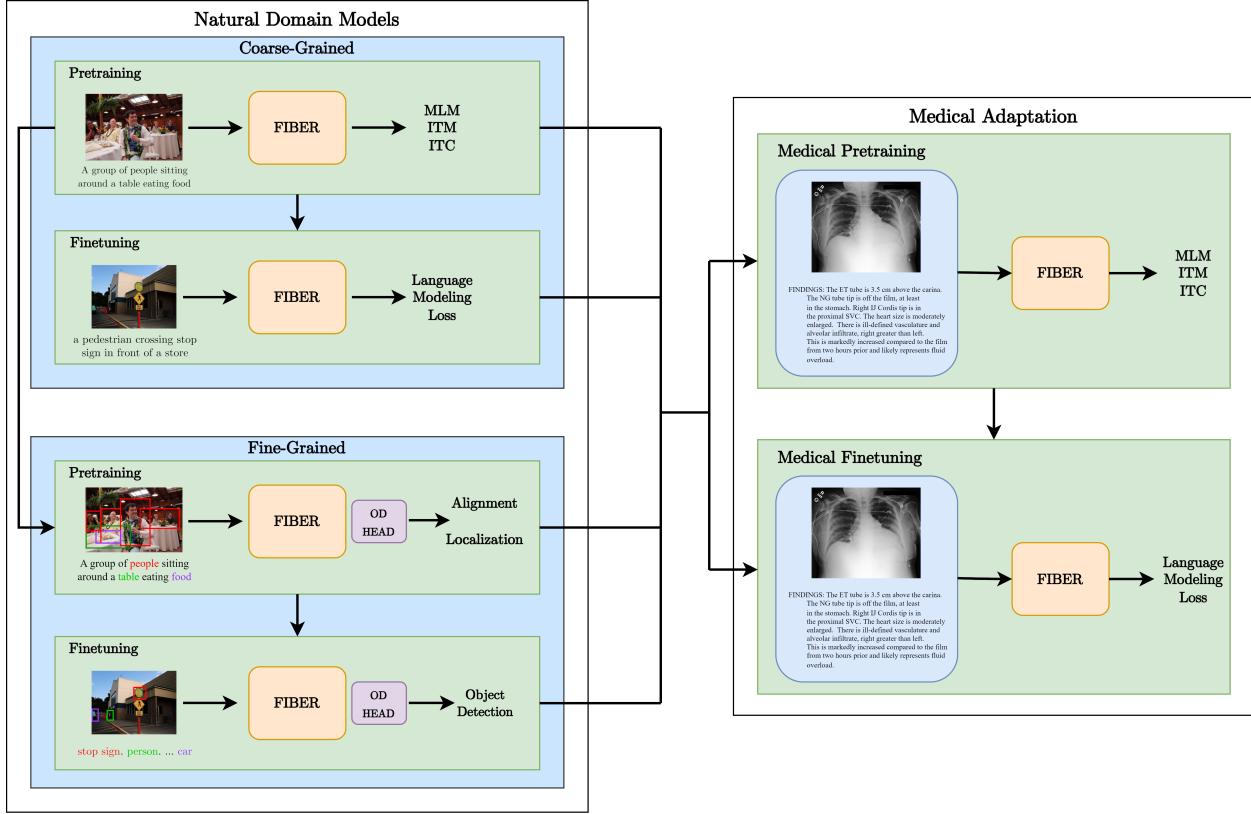


Figure 3: Framework of cross-domain adaption process. For **natural domain vision-language models**, there are multiple training stages where the models can be categorized as four classes including: (1) coarse-grained representation learning through *pre-training* via self-supervised learning and then *fine-tuning* on specific dataset (COCO dataset) and task (image captioning) with corresponding labels. (2) fine-grained representation learning through *pre-training* via self-supervised learning and then *fine-tuning* on specific dataset (COCO dataset) and task (object detection) with corresponding labels. For **medical adaption methods**, we consider two adaptation tasks, that is, (1) *medical pre-training*: utilize various self-supervised losses to learn from the pairs of medical images and reports to adapt image encoder and text encoder; (2) *medical fine-tuning*: learn to adapt the image encoder and text decoder through a generative task with the image captioning loss function.

tions (section 2.1). Following that, we discuss the two methods we employ for cross-domain adaptation: medical pre-training using self-supervised tasks (section 2.3) and medical fine-tuning through a generative task (section 2.4).

2.1. Model architecture

We adopt FIBER as our primary model, as depicted in Figure 2 (Dou et al., 2022). FIBER integrates both an image and a text backbone and employs deep multimodal fusion through direct cross-attention modules between these backbones.

FIBER’s architecture is highly versatile, allowing it to adapt to various tasks and loss functions by tog-

gling the cross-attention modules and incorporating an object detection head.

We use a Swin Transformer, specifically Swin-Base, for the vision backbone and RoBERTa-Base for the text backbone throughout both pre-training and fine-tuning stages.

2.2. Natural Domain VL Models

We employ a two-stage coarse-to-fine pre-training methodology to establish four interconnected types of natural domain FIBER models. Each of these models serve as starting points for medical domain adaptation and represent different degrees of specialization on natural image-text representations.

	
Ground Truth	findings: No focal consolidation, pleural effusion, pneumothorax, or pulmonary edema is seen. Heart size is normal. There is persistent aortic tortuosity. No rib fracture is detected, although sensitivity is low on routine chest radiography. impression: No acute findings.
Model Output	findings: The heart is normal in size. The mediastinal and hilar contours appear within normal limits. There is no pleural effusion or pneumothorax. The lungs appear clear.impression: No evidence of acute cardiopulmonary disease.

Figure 4: Example for medical report generation from the given X-ray image. Sentences highlighted in green show the correct generation in the model outputs while sentences in red indicate the missing information in the generated report.

Coarse-grained Pre-training We initiate with a FIBER model pre-trained on natural image datasets. This model employs a combination of Masked Language Modeling (MLM), Image-Text Matching (ITM), and Image-Text Contrastive (ITC) objectives.

Coarse-grained Fine-tuning The coarse-grained pre-trained FIBER model is further fine-tuned using a natural image captioning dataset. This fine-tuning employs a language modeling loss function.

Fine-grained Pre-training We also consider a fine-grained pre-trained FIBER model, which is a coarse-grained pre-trained FIBER model further pre-trained on natural image object detection and grounding datasets.

Fine-grained Fine-tuning Finally, a fine-grained pre-trained FIBER model is additionally fine-tuned using a natural image object detection dataset to obtain a fine-grained finetuned FIBER model.

In this study, we explore two medical adaptation tasks: medical pre-training and medical fine-tuning starting from each of these four types of natural vision-language models. We also investigate a two-stage medical adaptation process which combines these two adaptation tasks.

2.3. Medical Pre-training

Medical pre-training is conceptually similar to the coarse-grained pre-training approach. We apply a combination of MLM, ITC, and ITM objectives to

a medical image-text dataset, specifically MIMIC-CXR.

2.4. Medical Fine-tuning

Medical fine-tuning aims to generate medical reports and is akin to the coarse-grained fine-tuning approach. We use a medical domain image-text dataset for this purpose.

Technical Details Consistent with prior work, the final image representation, as opposed to the intermediate image representations, is fed into the cross-attention modules, resulting in a sequence-to-sequence model structure. Causal masks are used for auto-regressive text decoding.

3. Experiments

3.1. Datasets

We perform our experiments on the MIMIC-CXR dataset (Johnson et al., 2019a,b; Goldberger et al., 2000), which consists of paired chest X-ray images and corresponding radiology reports. The raw dataset contains 377,110 images linked to 227,835 radiographic studies.

Data Splitting and Preprocessing For our experiments, we create training and testing subsets based on the official data split provided by MIMIC-CXR. Specifically, we consider the target report to be a concatenation of both the “findings” and “impression” sections from the raw radiology reports.

Models			Metrics					
Natural-domain Models	Medical Adaptation		Medical Image-to-Report Retrieval			Medical Report-to-Image Retrieval		
	Pre-training	Fine-tuning	R@5↑	R@10↑	Reciprocal Rank↑	R@5↑	R@10↑	Reciprocal Rank↑
	✓		0.411	0.822	0.603	0.411	0.651	0.593
	✓	✓	0.240	0.445	0.386	0.171	0.308	0.288
✓			0.102	0.377	0.274	0.240	0.514	0.294
✓	✓		27.064	38.472	18.804	26.824	37.239	18.125
✓		✓	0.411	0.754	0.581	0.308	0.514	0.376
✓	✓	✓	3.049	5.721	2.746	0.514	0.856	0.567
Gloria (Huang et al., 2021)			0.102	0.240	0.330	2.672	3.871	2.014

Table 1: Performance of zero-shot retrieval task. (All scores are reported in percent (%))

After preprocessing, our training subset comprises 356,220 images corresponding to 213,501 unique studies, while the testing subset includes 4,665 images and 2,919 unique studies.

Evaluation Subset All tasks are evaluated using this testing subset.

3.2. Task I: Zero-shot Retrieval

To assess the efficacy of our vision-language representation learning, we conduct a zero-shot image-report retrieval task without any model fine-tuning. This task serves as a measure of the quality of visual and text embeddings generated by our various models.

Experimental Setup In this experimental setup, the retrieval task is conducted on the entire testing subset, which contains 2,919 unique pairs of medical images and reports. For image-to-report retrieval, a query medical image is used to retrieve a target report from a database of 2,919 candidate reports, based on cosine similarity between their respective visual and text embeddings. Importantly, all embeddings are extracted using pre-trained encoders without any additional fine-tuning, making these true zero-shot retrieval tasks.

Evaluation Metrics We employ multiple quantitative metrics to evaluate retrieval performance, including recall at top-k ($R@k$) and reciprocal rank. $R@k$ quantifies the likelihood that the true retrieval target appears within the top-k retrieved results. Reciprocal rank is calculated as the inverse of the rank at which the first relevant item appears. We report the average scores across all test samples for these metrics.

3.3. Task II: Medical Report Generation

To assess the utility of our learned multimodal representations in downstream applications, we fine-tune our model for medical report generation. The model takes a medical image as input and generates a free-text report that includes both *findings* and *impressions* related to the image.

Experimental Setup The model receives a single X-ray image and generates a corresponding radiology report through a text generator and beam search.

Evaluation Metrics To gauge the quality of the generated reports, we compare them to ground-truth reports extracted from the raw radiology data associated with each patient study. We employ a variety of metrics to evaluate the generated reports, including *BLEU-2*, *BLEU-3*, and *BLEU-4* scores (Papineni et al., 2002) for linguistic similarity. Additionally, we use semantic similarity metrics such as *BERTScore* (Zhang et al., 2020), *CheXbert* (Smit et al., 2020), and *RadGraph* (Jain et al., 2021).

Clinical Relevance To provide a clinically relevant evaluation, we incorporate *RadCliQ*, a metric introduced in Yu et al. (2022). This metric combines multiple evaluation criteria to align with the assessments made by clinical experts. We report the average scores across all test samples for each of these metrics.

4. Results

4.1. Zero-shot Retrieval

The outcomes of the zero-shot retrieval task are summarized in Table 1. This table includes results for both medical image-to-report and report-to-image retrieval tasks. The first two rows display the performance of models trained solely on medical data, with-

Models						Metrics					
Coarse-grained		Fine-grained		Medical Adaptation		Medical Image-to-Report Retrieval			Medical Report-to-Image Retrieval		
Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	R@5↑	R@10↑	Reciprocal Rank↑	R@5↑	R@10↑	Reciprocal Rank↑
✓				✓		27.064	38.472	18.804	26.824	37.239	18.125
✓	✓	✓		✓		10.380	16.581	8.010	8.702	14.046	7.061
✓		✓	✓	✓	✓	12.264	19.253	8.763	10.106	17.575	7.999
✓				✓		10.963	18.774	8.311	10.414	16.375	7.600
✓					✓	0.411	0.754	0.581	0.308	0.514	0.376
✓	✓	✓				0.206	0.548	0.384	0.171	0.274	0.280
✓		✓	✓			0.274	0.548	0.332	0.171	0.308	0.287
✓		✓	✓	✓		0.171	0.343	0.285	0.103	0.411	0.296
✓				✓	✓	3.049	5.721	2.746	0.514	0.856	0.567
✓	✓	✓		✓	✓	2.295	4.488	2.339	1.370	2.672	1.494
✓		✓	✓	✓	✓	1.370	2.192	1.176	0.171	0.411	0.302
✓		✓	✓	✓	✓	0.548	1.131	0.818	0.171	0.308	0.295

Table 2: Performance of zero-shot retrieval task. (All scores are reported in percent (%))

out leveraging any pre-trained natural-domain models. Rows three to six feature the results of cross-domain adaptation models, which start from pre-trained natural-domain models and undergo either medical pre-training or fine-tuning. For a detailed breakdown of the natural domain models and the various medical adaptation methods, refer to Fig. 3.

Role of Initialization and Adaptation Our results show that both initializing with natural domain vision-language models and further medical adaptation training play key roles in improving representations for medical retrieval through cross-domain transfer. This aligns with our hypothesis, illustrated in Fig. 1, that adapting large-scale natural models can leverage limited medical data while bridging the domain gap. To demonstrate the effectiveness of adapted models, we compare against GLORIA, and a baseline FIBER model trained from scratch on medical image-text pairs via self-supervision.

Differential Impact of Pre-training and Fine-tuning Although both medical pre-training and fine-tuning improve performance, pre-training results in much larger gains on the image-report retrieval task. This significant difference in impact can be attributed to the nature of the objectives used in each adaptation approach.

In pre-training, self-supervised objectives like masked language modeling and image-text matching directly optimize the model to produce effective encodings of medical images and text. In contrast, fine-tuning focuses on the downstream task of medical report generation, which does not directly optimize the text encoder for retrieval tasks. However, it still refines the visual embeddings, albeit less dramatically than direct pre-training.

Cross-Domain Generalization To further probe generalization of natural domain vision-language models, we perform an ablation using variants from different training stages as initialization for adaptation. Specifically, we consider four model types - coarse-grained pre-training, coarse-grained fine-tuning, fine-grained pre-training, and fine-grained fine-tuning.

As Table 2 shows, coarse-grained pre-training models transfer best across domains, with their generic global representations enabling robust adaptation. Fine-grained models learn localized features beneficial for medical tasks, but their specialized representations seem less transferable.

4.2. Medical Report Generation

Impact on Medical Report Generation As shown in Table 3, adapting natural domain vision-language models also assists performance on the medical report generation task, beyond just improving retrieval.

Efficiency of Task-Similar Fine-tuning Natural domain models that were fine-tuned on image captioning also achieve near state-of-the-art performance after adaptation on medical report generation. This suggests that task similarity enables more efficient transfer.

Performance Gaps and Knowledge Transfer However, the performance gaps between different model variants are smaller for medical report generation compared to retrieval tasks. This implies that fine-tuning on the supervised generation task may result in some loss of transferable knowledge from the pre-trained representations.

Models						Metrics						
Coarse-grained		Fine-grained		Medical Adaptation		BLEU-2↑	BLEU-3↑	BLEU-4↑	BERTScore↑	CheXbert↑	RadGraph↑	RadCliQ↓
Pre-trained	Fine-tuned	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned							
✓				✓	✓	0.143	0.089	0.062	0.340	0.304	0.134	3.814
✓	✓			✓	✓	0.155	0.099	0.069	0.352	0.346	0.156	3.671
✓		✓		✓	✓	0.156	0.100	0.069	0.353	0.351	0.159	3.656
✓		✓	✓	✓	✓	0.154	0.098	0.068	0.351	0.344	0.154	3.681
✓		✓	✓	✓	✓	0.153	0.096	0.066	0.347	0.344	0.149	3.700
✓	✓			✓	✓	0.157	0.100	0.069	0.354	0.359	0.159	3.638
✓	✓			✓	✓	0.155	0.100	0.070	0.355	0.356	0.159	3.640
✓		✓		✓	✓	0.155	0.098	0.068	0.350	0.344	0.155	3.682
✓		✓	✓	✓	✓	0.156	0.100	0.069	0.353	0.349	0.156	3.664

Table 3: Performance of medical report generation task. (Study-level)

4.3. Concluding Observations

The overall strong results confirm that leveraging natural domain vision-language models via adaptation techniques is a promising approach to overcome limited training data in medical applications like report generation. Further investigations on model architectures and adaptation methods would help unlock additional performance gains.

Limitations. While this study offers valuable contributions to the field, it’s important to view its findings as a starting point for further research rather than as definitive conclusions. Firstly, the study focuses on the MIMIC-CXR medical imaging dataset, offering a specialized rather than a generalized perspective. Future research could benefit from incorporating a variety of datasets, including different imaging modalities like MRI and CT scans, as well as diverse clinical applications such as cardiology and dermatology. This would enrich the study’s applicability across different medical domains. Secondly, the study explores a limited range of vision-language model architectures. While this provides a focused analysis, additional research could explore a wider array of model designs and sizes to offer a more comprehensive understanding of how these models can be adapted for medical applications. Thirdly, the study primarily addresses two specific applications: radiology report generation and image-report retrieval. Extending the scope to include other medically relevant tasks, such as disease classification or personalized diagnosis, could provide a more complete picture of the model’s utility in healthcare settings. Fourthly, the study employs a pretraining-finetuning framework for model adaptation, which is just one of many possible approaches. Future work could explore alternative methods like intermediate task fine-tuning or adapter modules to potentially enhance the model’s adaptability across domains. Lastly, the study does not delve into performance variations across different

disease types or image qualities. Subsequent research could conduct more detailed analyses to uncover any model biases or limitations in these areas.

5. Conclusion

In this study, we explore the problem of cross-domain adaptation for vision-language (VL) foundation models, specifically focusing on adapting from natural domains to medical domains. Our approach efficiently leverages large-scale pretrained VL models developed for natural domains, while effectively bridging the domain gap even when only limited data is available in the new medical domain. We conduct a systematic evaluation of various pretrained models and adaptation methods, assessing their performance across multiple medical applications. The insights gained from this study offer valuable guidance for adapting VL models to different domains. Future research directions may include fine-grained medical adaptation using FIBER for tasks such as medical phrase-grounding and object detection, as well as exploring the benefits of fine-grained medical pre-training for report generation tasks.

References

- Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multi-modal biomedical AI. *Nature Medicine*, 28(9):1773–1784, September 2022. doi: 10.1038/s41591-022-01981-2. URL <https://doi.org/10.1038/s41591-022-01981-2>.
- Benedikt Boecking, Naoto Usuyama, Shruthi Banur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve

- biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 1–21. Springer, 2022.
- Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23), June 2000. doi: 10.1161/01.cir.101.23.e215. URL <https://doi.org/10.1161/01.cir.101.23.e215>.
- Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 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*, pages 3942–3951, 2021.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
- Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation, 2023.
- Alistair E. W. Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. The mimic-cxr database, 2019a. URL <https://physionet.org/content/mimic-cxr/>.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih Ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), December 2019b. doi: 10.1038/s41597-019-0322-0. URL <https://doi.org/10.1038/s41597-019-0322-0>.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
- Aakash Mishra, Rajat Mittal, Christy Jestin, Kostas Tingos, and Pranav Rajpurkar. Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models, 2023.
- Michael Moor, Oishi Banerjee, Zahra Shakeri Hosseini Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023a. doi: 10.1038/s41586-023-05881-4. URL <https://doi.org/10.1038/s41586-023-05881-4>.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023b.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.
- Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, September 2022. doi: 10.1038/s41551-022-00936-9. URL <https://doi.org/10.1038/s41551-022-00936-9>.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *arXiv preprint arXiv:2210.06044*, 2022.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023a.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders, 2023b.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL <https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1), July 2023. doi: 10.1038/s41467-023-40260-7. URL <https://doi.org/10.1038/s41467-023-40260-7>.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.