Harnessing CLIP for Evidence Identification in Scientific Literature: A Multimodal Approach to the Context24 Shared Task

Anukriti Kumar **Lucy Lu Wang**
University of Washington
{anukumar, lucylw}@uw.edu

Abstract

Knowing whether scientific claims are supported by evidence is fundamental to scholarly communication and evidence-based decision-making. We present our approach to Task 1 of the Context24 Shared Task—Contextualizing Scientific Figures and Tables (SDP@ACL2024), which focuses on identifying multimodal evidence from scientific publications that support claims. We finetune CLIP, a state-of-the-art model for image-text similarity tasks, to identify and rank figures and tables in papers that substantiate specific claims. Our methods focus on text and image preprocessing techniques and augmenting the organizer-provided training data with labeled examples from the SciMMIR and MedICaT datasets. Our best-performing model achieved NDCG@5 and NDCG@10 values of 0.26 and 0.30, respectively, on the Context24 test split. Our findings underscore the effectiveness of data augmentation and preprocessing in improving the model's ability in evidence matching.

1 Introduction

Scientific claims, essential for communicating findings in literature reviews, research problem formulation, and interpreting conflicting data, require careful contextualization and verification with supporting empirical evidence and methodological details. Figures and tables in papers presenting key results, measures, and sample characteristics are critical for understanding and validating claims. However, retrieving such contextual information from scientific papers is challenging and time-consuming, particularly when researchers and practitioners encounter claims without direct access to the source materials.

Recent advancements in multimodal AI models, which excel at image-text similarity tasks, present an opportunity to automate the retrieval and contextualization of evidence. However, their application for evidence identification in scientific literature remains underexplored. Existing research primarily focuses on text summarization (Zhu et al., 2021; Li et al., 2020; Rafi and Das, 2023), entity recognition (Liu et al., 2023; Chen and Feng, 2023), and information retrieval (Jin et al., 2023; Imhof and Braschler, 2018), with limited attention to integrating visual evidence from figures and tables.

This paper presents our approach to Task 1 of the Context24 Shared Task,[1] focusing on evidence identification for scientific claims. The task involves predicting a ranked list of figures or tables from a relevant research paper that provides supporting evidence for a given scientific claim. We leverage the Contrastive Language–Image Pretraining (CLIP) model's image-text similarity capabilities, augment training data with diverse datasets and augmentation techniques, and employ text-image pre-processing methods to retrieve and rank relevant figures and tables from scientific papers.

The contributions of our work are as follows:

- We introduce a comprehensive dataset that combines augmented training data from the shared task with additional data from SciMMIR (Wu et al., 2024) and MedICaT (Subramanian et al., 2020), significantly expanding the scope and diversity of the training samples (§3).
- We finetune the CLIP model on this combined dataset and demonstrate the impact of more training data, as well as text and image pre-processing techniques on enhancing the model's relevancy matching capabilities (§4; §5).

2 Related Work

Scientific Evidence Identification Early work on scientific evidence identification focused on text analysis and argument mining techniques, like those of Guo et al. (2011), who investigated weakly-supervised approaches for detecting argu-

[1] https://sdproc.org/2024/sharedtasks.html#context24

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 307–313
August 16, 2024 ©2024 Association for Computational Linguistics

ment zones in scientific abstracts using features such as location, word bi-grams, and verb cues. Building on this, Houngbo and Mercer (2014) further explored the identification of rhetorical structures in biomedical articles; their analysis of the IMRaD structure (Introduction, Methods, Results, and Discussion) clarifies how different sections of a scientific paper contribute to the overall argumentative flow. Later, Faiz and Mercer (2014) developed techniques to extract "higher order relations" between biomedical entities, useful in the identification of causal links essential for evidence-based conclusions.

Neural models and machine learning further transformed capabilities in this domain. Techniques such as Bi-LSTM (Lauscher et al., 2018a,b) and BiLSTM-CRF (Li et al., 2021b; Achakulvisut et al., 2019) were adapted to parse and analyze argumentative structures in scientific texts for sentence-level classification and multi-task learning. These works enhanced the ability to distinguish between different types of evidence based on linguistic cues and discourse patterns. Further, Pinto et al. (2019) introduced learning-to-rank techniques combined with metadata features, such as journal impact factor and citation metrics, to optimize the retrieval process. Wadden et al. (2020) introduced the task of scientific claim verification, to select abstracts from research literature that either support or refute a given scientific claim; the authors introduced the SciFact dataset and baseline models, demonstrating that domain adaptation techniques significantly improve performance compared to models trained on Wikipedia or political news. The authors later extended the SciFact dataset into the open domain by integrating retrieval in the data curation process (Wadden et al., 2022).

Image-Text Matching Recent advancements in vision-language pretraining (VLP) have significantly reshaped the landscape of evidence identification. VLP aims to learn universal and transferable features from images and text, applicable across image-text retrieval tasks (Tan and Bansal, 2019; Zhang et al., 2020; Chen et al., 2019; Zhang et al., 2021; Kim et al., 2021; Jia et al., 2021; Li et al., 2021a; Wang et al., 2021; Li et al., 2022). Multimodal models like CLIP (Radford et al., 2021), BLIP-2 (Li et al., 2023), and Stable Diffusion (Rombach et al., 2021) that have undergone significant multimodal pretraining, have the potential to improve the alignment of text-based claims with image-based evidence.

Our work builds on these advancements by fine-tuning a model based on the CLIP architecture (Huang et al., 2023) on large-scale scientific datasets with an image-text matching objective. Extending beyond the typical focus on biomedical or computer science domains in prior works (Wang et al., 2022; Lin et al., 2023; Eslami et al., 2023), our approach covers a broader range of scientific disciplines. We leverage these image-text embeddings to retrieve and match scientific claims with their supporting evidence.

3 Datasets

This section describes the Context24 dataset (§3.1) and outlines our strategies for data enrichment (§3.2) and external data augmentation (§3.3).

3.1 Context24 Dataset

The Context24 dataset[2] comprises 474 scientific claims used in lab notes and discussions for synthesis and research planning, across domains of biology, computer science, and the social sciences. These claims are accompanied by "gold" annotations identifying figures and tables in the full text of research papers that provide key supporting evidence. We randomly sample 156 claims as the validation set, representing 34% of the dataset.

3.2 Data Enrichment

We enrich the training dataset with additional context. We leverage GPT-4 (OpenAI, 2023) to extract sub-captions and references for each of the associated figures and tables to provide a deeper contextual understanding for each claim.

Data cleaning Before enrichment, we perform the following steps:

- We eliminate unnecessary spaces, punctuation, and non-standard characters that could interfere with text processing;
- We convert all claims to lowercase to maintain consistency across the dataset, and correct formatting inconsistencies that could affect subsequent text parsing;
- We address common textual errors, such as excessive repetition of characters or patterns, non-standard characters (e.g., control characters or invalid Unicode), and executable commands,

[2] Training dataset: https://github.com/oasisresearchlab/context24/blob/main/task1-train-dev-2024-04-25-update.json

which throw errors when input into the GPT-4 API, by removing or heuristically fixing them.

Sub-caption Extraction Using GPT-4, we parse figure captions to extract descriptions for each sub-figure. This extraction is guided by the prompt: "This is a caption for a figure consisting of multiple sub-figures. Extract image descriptions for all sub-figures." Each extracted sub-caption is aligned with its corresponding sub-figure.

Reference Extraction Similarly, we extract textual references to figures and tables from the full texts of papers using the prompt: "This is the full text of a scientific paper. Extract only the relevant details mentioned regarding the figures and tables." These references were mapped to the corresponding claims and captions to strengthen the linkage between textual claims and the paper contexts.

Verification of GPT-4 Responses We manually review a random sample of extracted sub-captions and references (N=20) by comparing them with the original text (captions and full texts) to identify hallucinations or inaccuracies.

We found that GPT-4 accurately extracted sub-captions and references in 75% of cases (N=15). However, hallucinations occurred in 10% of cases (N=2), where GPT-4 fabricated content not present in the original text. The remaining 15% (N=3) contained minor inaccuracies such as incorrect sub-figure associations or slight misinterpretations of the text. As a result, we implement a series of post-processing steps:

- We filter out responses containing phrases like "Sorry..." or similar, as these often indicate GPT-4's inability to provide meaningful content;
- We discard responses with less than three words to ensure sufficient descriptive content;
- We compare extracted sub-captions and references against standard mention formats in scientific papers, e.g., "In Figure X, we...", to flag potential extraction errors. By cross-referencing GPT-4's outputs with these patterns, we discard text that deviates from our expectations.

This hybrid approach of automated extraction followed by heuristic curation ensures a higher degree of accuracy and reliability in the enrichment data. We refer to the resulting data with sub-captions and references as the augmented training data (ATD).

Dataset	**Size**
`Context24 Training Data`	456
`Context24 References Data`	609
`Context24 Captions Data`	438
`Context24 Augmented Cleaned Data`	1466
`SciMMIR`	530975
`MedICaT`	2118
Total (Combined)	534559

Table 1: Dataset Statistics

3.3 External Data Augmentation

To further augment and diversify our training data, we integrate two external datasets: SciMMIR (Wu et al., 2024) and MedICAT (Subramanian et al., 2020). Statistics for the augmented dataset are provided in Table 1.

SciMMIR Dataset SciMMIR (Wu et al., 2024) is a large-scale collection of scientific image-text pairs extracted from papers published on arXiv between May-October 2023. This dataset comprises approximately 530K image-text pairs, categorized into training (N=498K), validation (N=16.4K), and test splits (N=16.3K).

MedICaT Dataset MedICaT (Subramanian et al., 2020) is a collection of 217K medical images and their captions and references extracted from 131K open-access biomedical articles in PubMed Central. The dataset includes inline references for 74% of figures, and manually annotated subfigures and subcaptions for a subset of 2118 figures. For this study, we use only the subfigure-subcaption pairs in MedICaT for training.

4 Experiments

We detail the model selection and training process, and experiments to assess the impact of training data variation and preprocessing techniques.

4.1 Baseline Model

We select OpenAI's CLIP model, specifically the `clip-vit-base-patch32` variant from Hugging Face,[3] as our base image and text encoder. CLIP utilizes a vision transformer architecture (ViT-B/32) as an image encoder and a masked self-attention transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via contrastive loss, making them a suitable choice for this multimodal retrieval task.

[3] HuggingFace CLIP base model: https://huggingface.co/openai/clip-vit-base-patch32

4.2 Training Details

We train the CLIP model by minimizing the contrastive loss between the embeddings for images and corresponding text (claims, captions, and references) on the Context24 and augmented datasets. We finetune all models on Nvidia A100 GPUs, with 100 epochs and early stopping. We conduct hyperparameter tuning to select learning rates (0.00001 to 0.001), batch sizes (16, 32, and 64), and the margin parameter in the contrastive loss function. The final training was over 10 epochs, with a batch size of 32, a learning rate of 0.0001, and a decay rate of 0.01. We chose the Adam optimizer for its adaptive learning rate capabilities. Additionally, we employed 5-fold cross-validation for better model robustness and generalizability.

4.3 Variation in Training Data

We assess the impact of different dataset configurations on the performance of the CLIP model. Initially, we train the model using only the Context24 training dataset (TD) to establish baseline performance. To evaluate the effects of data augmentation, we compare three augmented datasets, one enriched with additional text snippets such as sub-captions and references (ATD), adding the SciMMIR dataset, and adding the SciMMIR and MedICaT datasets.

4.4 Variation in Preprocessing Techniques

We additionally explore how preprocessing techniques affect the CLIP model's performance. We apply text transformation techniques such as Named Entity Recognition (NER) and semantic role labeling (SRL) to enhance the model's understanding of grammatical structures and relationships between entities within the text. Specifically, we tag named entities and include SRL labels alongside the original tokens in the input, providing richer context and improving the text representation for the model. Additionally, we implement various image data augmentation techniques—including random cropping, rotation, resizing, and color jittering—during training to improve the model's robustness to visual variations.

5 Results

Following the Context24 shared task, we report Normalized Discounted Cumulative Gain (NDCG) at 5 and 10 to assess the ranking of the retrieved figures and tables. Additionally, we report precision and recall for the top 5 and 10 retrieved images (i.e., P@5, R@5, P@10, and R@10).

5.1 Results on Validation Split

Table 2 presents the performance differences achieved through augmenting the training dataset. No preprocessing is conducted on any of these model variants. Notably, the addition of data shows steady improvements in NDCG@5 (0.19 to 0.27), NDCG@10 (0.21 to 0.29), and P/R@k. P@5 increased from 0.49 with initial training data to 0.76 with the combined dataset (ATD + SciMMIR + MedICaT). Similarly, R@5 showed substantial improvement, from 0.46 to 0.75.

Table 3 highlights the impact of preprocessing on model performance. Using the model finetuned on the combined dataset without any preprocessing techniques as a baseline, the addition of image-text preprocessing, while leading to minimal gains in NDCG, led to substantial gains in P/R@k: P@5 increased to 0.86 and R@5 to 0.82.

The results from our experiments on the validation dataset underscore the significance of dataset enhancement and image-text preprocessing on the performance of image-text similarity models like CLIP in identifying scientific evidence.

5.2 Results on Test Split

We apply our finetuned CLIP model, trained with the combined dataset and enhanced with image-text pre-processing techniques, to a test set comprising 111 scientific claims. Before inference, we perform data cleaning and context enrichment, as detailed in Section 3.2. The model computes similarity scores between the textual claims and the associated visual evidence (figures or tables). These scores were used to rank the evidence, and final rankings were submitted to the eval.ai platform.[4] Our team secured third place on the leaderboard, with NDCG@5 of 0.26 and NDCG@10 of 0.30 on the test split.

6 Future Work

Our current approach leverages the entire MedICaT subcaption subset and SciMMIR datasets for training the CLIP model. However, we can potentially improve performance by selecting subsets of these datasets that more closely align with the scientific domains or types of claims in the Context24 dataset. This approach might enhance the

[4] Evaluation platform for Context24: https://eval.ai/web/challenges/challenge-page/2306/overview

Training Data Used	NDCG@5	NDCG@10	P@5	P@10	R@5	R@10
Training Data (TD)	0.19	0.21	0.49	0.48	0.46	0.46
Augmented TD (ATD)	0.23	0.26	0.56	0.67	0.54	0.61
ATD + MedICAT	0.24	0.27	0.58	0.67	0.55	0.61
ATD + SciMMIR	0.26	0.28	0.75	0.73	0.74	0.65
ATD + SciMMIR + MedICaT	0.27	0.29	0.76	0.73	0.75	0.65

Table 2: Performance metrics for CLIP model finetuned with different datasets (without pre-processing)

Preprocessing Techniques	NDCG@5	NDCG@10	P@5	P@10	R@5	R@10
No pre-processing	0.27	0.29	0.76	0.73	0.75	0.65
Image pre-processing	0.27	0.31	0.77	0.75	0.76	0.71
Text-Image pre-processing	0.28	0.32	0.86	0.81	0.82	0.77

Table 3: Performance metrics for CLIP model finetuned on the combined dataset (ATD + SciMMIR + MedICaT) with different pre-processing techniques

relevance and quality of the training data to the task. In future work, we will explore using text and image embedding similarity methods to filter the augmentation data by domain relevance.

We also acknowledge the challenge presented by complex figures with multiple sub-figures, which are not fully described in the captions. While our method of using GPT-4 to extract corresponding subcaptions was somewhat effective, the observed hallucinations and inaccuracies highlight the need for continuous refinement and validation of automated methods. To address this, future work might explore alternative models or techniques for subcaption-subfigure alignment.

Currently, we use references and captions from figures and tables as additional image-text pairs to train the CLIP model. Another potential enhancement involves augmenting the claims with this contextual information *before* computing similarity scores with the images. This method could provide a more holistic understanding of the claim and its context by incorporating essential details directly alongside the claim text in claim verification.

7 Conclusion

In this paper, we presented our approach to Task 1 of the Context24 Shared Task, which focuses on identifying multimodal evidence for scientific claims. We enhanced our training data by integrating sub-captions and inline references as well as additional data from SciMMIR and MedICaT. We finetuned the CLIP model on this enriched dataset, applying image-text preprocessing and augmentation techniques to effectively align scientific claims with their corresponding figures and tables. We secured third place on the leaderboard, with NDCG@5 and NDCG@10 scores of 0.26 and 0.30, respectively on the test dataset. Our findings underscore the significance of expanding training data and employing preprocessing techniques on the model's performance.

Acknowledgments

This work is supported in part by the University of Washington eScience Insitute's Azure Cloud Credits for Research and Teaching.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Ernesto Acuna, and Konrad Paul Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *ArXiv*, abs/1907.00962.

F. Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings*.

Syeed Ibn Faiz and Robert E. Mercer. 2014. Extracting higher order relations from biomedical text. In *ArgMining@ACL*.

Yufan Guo, Anna Korhonen, and T. Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Conference on Empirical Methods in Natural Language Processing*.

Hospice Houngbo and Robert E. Mercer. 2014. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *ArgMining@ACL*.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Y Zou. 2023. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*.

Melanie Imhof and Martin Braschler. 2018. A study of untrained models for multimodal information retrieval. *Information Retrieval Journal*, 21:81–106.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *ArXiv*, abs/2102.05918.

Zhenkun Jin, Xingshi Wan, Xin Nie, Xinlei Zhou, Yuanyuan Yi, and Gefei Zhou. 2023. Ranking on heterogeneous manifold for multimodal information retrieval. *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 989–996.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*.

Anne Lauscher, Goran Glavas, and K. Eckert. 2018a. Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *ArgMining@EMNLP*.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and K. Eckert. 2018b. Investigating the role of argumentation in the rhetorical analysis of scientific publications with neural multi-task learning models. In *Conference on Empirical Methods in Natural Language Processing*.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020. Multimodal sentence summarization via multimodal selective encoding. In *International Conference on Computational Linguistics*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*.

Xiangci Li, Gully A. Burns, and Nanyun Peng. 2021b. Scientific discourse tagging for evidence extraction. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Weide Liu, Xiaoyang Zhong, Jingwen Hou, Shaohua Li, Haozhe Huang, and Yuming Fang. 2023. Integrating large pre-trained models into multimodal named entity recognition with evidential fusion. *ArXiv*, abs/2306.16991.

OpenAI. 2023. Gpt-4 technical report.

José María González Pinto, Serkan Çelik, and Wolf-Tilo Balke. 2019. Learning to rank claim-evidence pairs to assist scientific-based argumentation. In *International Conference on Theory and Practice of Digital Libraries*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Shaik Rafi and Ranjita Das. 2023. Abstractive text summarization using multimodal information. *2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 141–145.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *ArXiv*, abs/2010.06000.

Hao Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, N. A. Moubayed, Jie Fu, and Chenghua Lin. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. *ArXiv*, abs/2401.13478.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and C. Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning in Health Care*.

Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Graph-based multimodal ranking models for multimodal summarization. *Transactions on Asian and Low-Resource Language Information Processing*, 20:1 – 21.