Alignment, Mining and Fusion: Representation Alignment with Hard Negative Mining and Selective Knowledge Fusion for Medical Visual Question Answering

Supplementary Material

6. Implementation Details

For the pre-training phase, we apply CLIP-ViT/B-16 [46] as the initialized visual encoder, and ClinicalBERT [21] as the initialized textual encoder and auto-regressive decoder. All images are resized to 288×288. The dimension of embeddings d is set to 768. The Modality Co-Attention Module uses $L = 6$ layers. The total trainable model size is 326M. The temperature parameters τ_1, τ_2, τ_3 are set to 0.07. The flexibility weight of the positive pair is cross-validated among different value choices and set $\lambda = 0.01$. Our pre-training phase contains 20 epochs for stage one, and 10 epochs for stage two, with 256 as batch size. The model uses AdamW optimizer with a learning rate of 4e-6.

For the fine-tuning phase, the parameters for the Gated Cross-Attention Module are 14.8M, while 122M for the auto-regressive decoder. We fine-tune the model for 60 epochs with a learning rate of 2e-5 and a batch size of 64.

Our computing resources are two NVIDIA A100 GPUs, and the training uses 16-mixed precision. The pre-training costs roughly 24 hours, while fine-tuning time is around 8 hours. The inference time for one image-question pair is around $0.27s$. During inference, to fairly compare with other multi-label classification-based methods, we apply the test set answers as the candidate answers, compare the generated open-form answer from our auto-regressive decoder with these candidates, and choose the one with the lowest language modeling loss.

7. Additive Ablation Studies

Besides the ablation study we conduct in Sec. 4.4, we also evaluate the impacts of different methods we use to generate the soft labels, and how the λ (Eq. (5)) will affect the learning of these soft labels.

#	Soft Label Generation Method	RAD-VQA	SLAKE	Path-VQA	VQA-2019	Avg
1	-	75.64	81.09	60.92	78.14	73.95
2	CLIP[46]	76.35	81.32	61.38	78.42	74.37
3	MedCLIP [54]	78.80	84.82	62.83	81.03	76.87
4	PubMedCLIP[15]	80.02	85.32	63.12	80.93	77.35
5	PMC-CLIP[32]	79.63	84.98	63.01	80.44	77.02
6	BioMedCLIP [57]	80.26	85.37	63.96	81.26	**77.71**

Table 5. Ablation study on different choices of soft label generation methods for Med-VQA tasks.

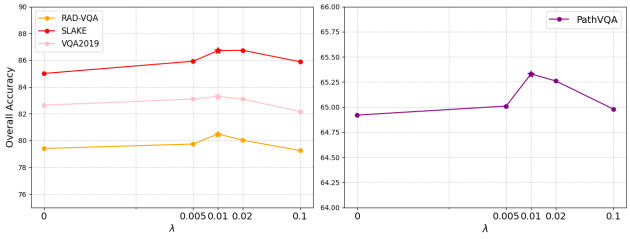

Figure 6. Ablation study on different choices of flexibility weight λ for Med-VQA tasks. Datasets are split based on the range of accuracy values.

7.1. Impacts of Different Soft Label Generation Methods

During our experiment for exploring soft label generation, we attempt different CLIP-based models, evaluating their performances in depicting the similarity between image-text pairs. Our ablative experiment focuses on several representative CLIP-based models [15, 32, 46, 54, 57]. We use the `base` version of these models to keep all model sizes in a comparative range.

As shown in Tab. 5, when we focus on the average accuracy, the comparison between #2 (CLIP [46]) and the best one #6 (BioMedCLIP [57]) indicates that soft labels generated from the medical-specific model have better instructional impacts than the model from natural domain. While #4, #5, #6 are all models pre-trained on medical datasets, the average results are close to each other, which demonstrates that the usage of medical knowledge has a nearly equal effect on improving the model representation ability for Med-VQA tasks. In conclusion, we believe a well-trained CLIP-based model in medical domain will benefit the soft label supervision, and thus improving the performance of downstream tasks (*e.g.*, Med-VQA).

7.2. Impacts of flexibility weight λ for soft labels

We also evaluate different settings of λ that control the positive sample's similarity, and we find out that $\lambda = 0.01$ is a locally best choice, as shown in Fig. 6. Our explanation for this performance difference is that when we add less weight, the model might ignore the positive samples because of the high similarity of both hard negative ones and positive ones. Moreover, if we apply too much weight, the model will focus too much on positive samples, ignoring the similarity caused by similar diseases among patients, resulting in the decline of performance. Therefore, we believe our setting

is a trade-off between focus and ignorance.

References

[1] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16430–16441, 2022. 3

[2] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019. 2, 7

[3] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. 2

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 5

[5] Cheng Chen, Aoxiao Zhong, Dufan Wu, Jie Luo, and Quanzheng Li. Contrastive masked image-text modeling for medical visual representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 493–503. Springer, 2023. 1, 2, 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 6

[7] Xiaofei Chen, Yuting He, Cheng Xue, Rongjun Ge, Shuo Li, and Guanyu Yang. Knowledge boosting: Rethinking medical contrastive vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023. 1, 3

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3, 5

[9] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 1, 3, 6, 7

[10] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022. 1, 2, 3, 6, 7

[11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 6

[12] Gefen Dawidowicz, Elad Hirsch, and Ayellet Tal. Limitr: Leveraging local information for medical image-text representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21165–21173, 2023. 1, 3

[13] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer, 2021. 2, 3, 7

[14] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019. 3

[15] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. 1, 2, 3, 7

[16] Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. Lapa: Latent prompt assist model for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 1, 2, 3, 7

[17] Jinlong He, Pengfei Li, Gang Liu, Zixu Zhao, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv preprint arXiv:2401.02797*, 2024. 7

[18] Xuehai He. Towards visual question answering on pathology images. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, 2021. 2, 7

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[20] Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, pages 18298–18306, 2024. 3

[21] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. 7, 1

[22] 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. 1, 2

[23] Jongseong Jang, Daeun Kyung, Seung Hwan Kim, Honglak Lee, Kyunghoon Bae, and Edward Choi. Significantly improving zero-shot x-ray pathology classification via fine-tuning pre-trained image-text encoders. *Scientific Reports*, 14(1):23199, 2024. 3

[24] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 1, 6

[25] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021. 7

[26] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[27] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2, 7

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021. 3, 6

[29] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023. 1, 2, 3, 7

[30] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pre-training for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. 1, 2, 3, 6, 7

[31] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 273–283. Springer, 2021. 2

[32] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023. 1

[33] Zudi Lin, Erhan Bas, Kunwar Yashraj Singh, Gurumurthy Swaminathan, and Rahul Bhotika. Relaxing contrastiveness in multimodal representation learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2227–2236, 2023. 1, 3

[34] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143: 102611, 2023. 1

[35] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer, 2021. 3, 7

[36] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 2, 7

[37] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022. 7

[38] Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. Q2atransformer: Improving medical vqa via an answer querying decoder. In *International Confer-*

ence on Information Processing in Medical Imaging, pages 445–456. Springer, 2023. 3

[39] Sruthy Manmadhan and Binsu C Kovoor. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, 53(8):5705–5745, 2020. 1

[40] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, pages 685–701. Springer, 2022. 1, 3

[41] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019. 2, 3, 7

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[43] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018. 6

[44] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*, 2019. 2

[45] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6967–6977, 2023. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 7

[47] Syed A Rizvi, Ruixiang Tang, Xiaoqian Jiang, Xiaotian Ma, and Xia Hu. Local contrastive learning for medical image recognition. In *AMIA Annual Symposium Proceedings*, page 1236. American Medical Informatics Association, 2023. 2

[48] Prashant Shrestha, Sanskar Amgain, Bidur Khanal, Cristian A Linte, and Binod Bhattarai. Medical vision language pretraining: A survey. *arXiv preprint arXiv:2312.06224*, 2023. 3

[49] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020. 6

[50] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 1

[51] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 3

[52] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 1, 3

[53] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021. 2

[54] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 1, 2, 3

[55] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020. 3, 5

[56] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556, 2023. 7

[57] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 2, 3, 4, 7, 8

[58] 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. 1

[59] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023. 1, 2, 3