# MVCM: Enhancing Multi-View and Cross-Modality Alignment for Medical Visual Question Answering and Medical Image-Text Retrieval

Supplementary Material

## 6. Impacts of Coefficient of Loss Functions

We also conduct ablative experiments to optimize the coefficients of the loss functions associated with each pretext task in our model. Given the multitude of potential combinations of the pretext tasks, and considering previous studies for Vision-Language Pre-training like [33, 35] that have used balanced coefficients for ITM, ITG, and ITGC, our focus was primarily on assessing the coefficients for ITLC and IMVC.

The results of these ablative experiments are summarized in Tab. 5. Notably, entry #3, where the coefficients balance the influence of each task, achieves the best performance on the CheXpert 5×200 dataset. This suggests that despite the importance of ITLC and IMVC (#1, #2), overly emphasizing ITLC and IMVC may detract from the model's ability to learn effectively from the other tasks, potentially due to competition for model capacity.

## 7. Failure Cases and Disscussion

Despite the performance we achieved in various downstream tasks, there are still some limitations regarding the model performance. We summarize the failure cases and discuss the potential direction for improving them.

For the Image-Text Retrieval task, as shown in Fig. 5, our model can differentiate the representations of the first 3 labels better (bottom) compared with using all 5 labels (top). This difference in cluster indicates the two disease labels (*i.e.*, *Edema* and *Pleural Effusion*) are similar for our model, which is also proved by the X-ray images (Fig. 6). Therefore, we believe that there are still some limitations for current models to differentiate radiology images if the symptoms or visual features are very similar.

Additionally, for the VQA task, Fig. 7 highlights various representative failure cases within the PathVQA dataset [23]. Our pre-training datasets (*i.e.*, MIMIC-CXR, ROCO, MedICaT) predominantly consist of radiological images, such as X-rays, which may limit our model's performance on non-radiological images. Specifically, when confronted with images in the form of abstract illustrations (Fig. 7(a)) or photographs from cameras(Fig. 7(b)), our model struggles to fully comprehend the cross-modality information from these distinct domains, particularly for *Open* questions. Additionally, more complex cases (Fig. 7(c)) require the understanding of not only abstract illustrations but also embedded textual content within the images—a task that remains challenging for current models. Meanwhile, PathVQA also presents confusing samples without curating, such as those with multiple valid answers for the same question (Fig. 7(d), underlined questions), which further complicates the fine-tuning process. These aspects collectively contribute to the observed lower performance on *Open* questions in the PathVQA dataset.

| # | Coefficient | | Metric | | |
|---|---|---|---|---|---|
| | $\lambda_2$ (ITLC) | $\lambda_3$ (IMVC) | Prec@1 (I2T) | Prec@1 (T2I) | Acc |
| 1 | 0.5 | 0.5 | 33.28 | 40.37 | 41.22 |
| 2 | 0.5 | 1 | 35.73 | 42.03 | 43.81 |
| 3 | 1 | 1 | **36.52** | **42.70** | **44.40** |
| 4 | 1 | 2 | 33.67 | 40.21 | 42.39 |
| 5 | 2 | 2 | 32.40 | 39.04 | 40.82 |

Table 5. **Impact of Loss Function Coefficients.** We evaluate the effects of $\lambda_2$ for Image-Text Local Contrastive (ITLC) and $\lambda_3$ for Image Multi-View Contrastive (IMVC). Retrieval results on the CheXpert 5×200 Dataset are assessed using Precision (Prec) at the first candidate for Image-to-Text (I2T) and Text-to-Image (T2I), as well as Accuracy (Acc) for zero-shot classification of disease labels.

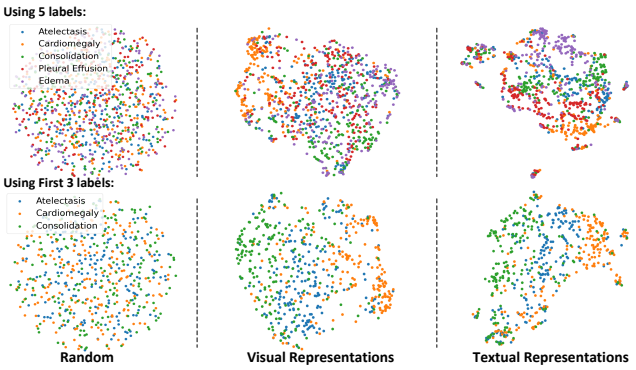


Figure 5. UMAP visualization of MVCM's representations on the CheXpert 5×200 dataset.

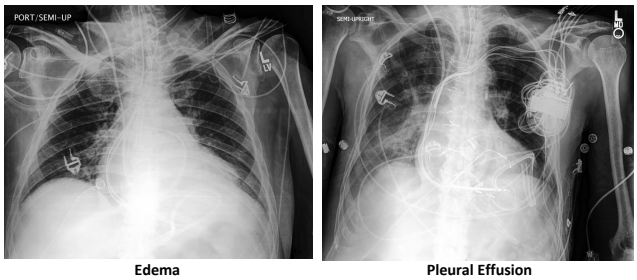


Figure 6. Example images of *Edema* and *Pleural Effusion* on the CheXpert 5×200 Dataset.

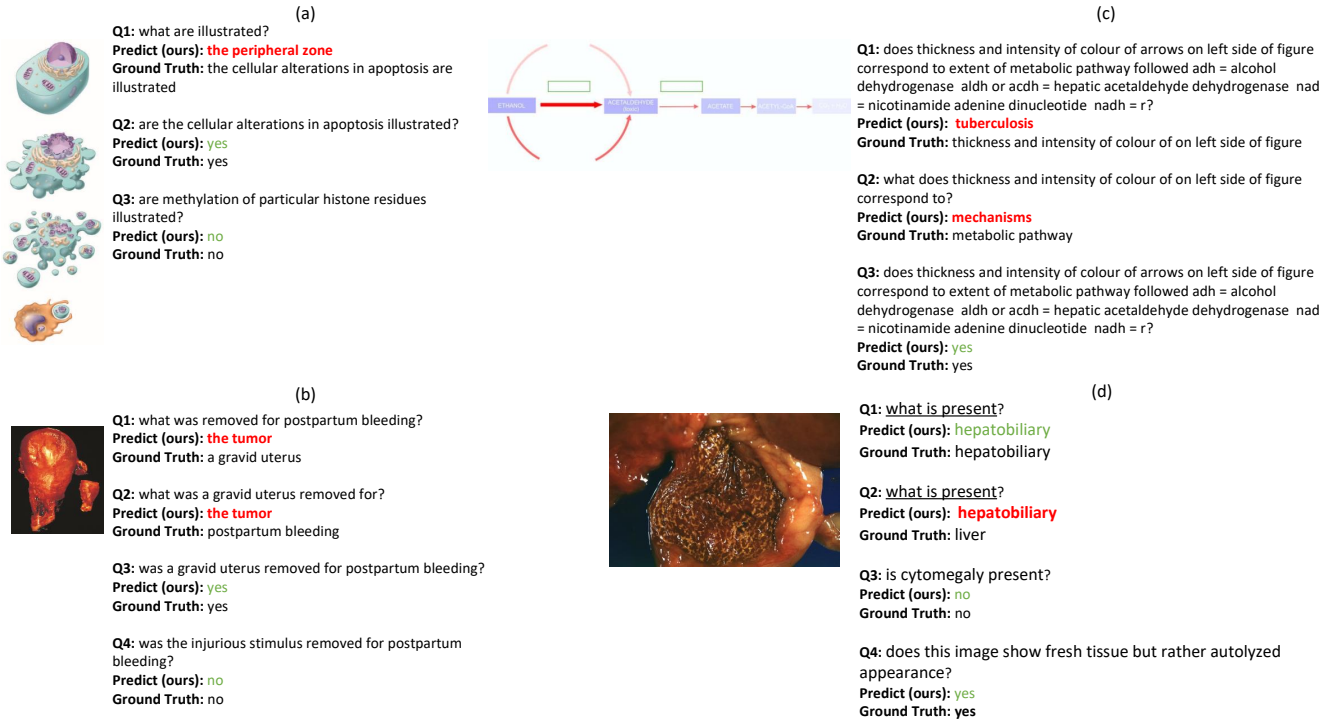


Figure 7. Examples of various failure cases for our MVCM on the PathVQA dataset.

## References


[1] Alan Joseph Bekker, Moran Shalhon, Hayit Greenspan, and Jacob Goldberger. Multi-view probabilistic classification of breast microcalcifications. *IEEE Transactions on medical imaging*, 35(2): 645–653, 2015. 2

[2] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. *Deep learning for medical image analysis*, pages 321–339, 2017. 2

[3] 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. 2

[4] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 1

[5] 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. 2, 6

[6] 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. 6

[7] 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. 2, 3, 6, 7

[8] 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. 2, 3, 7

[9] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21361–21371, 2023. 2, 6, 7

[10] 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

[11] 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. 3, 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 7

[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. 7

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 7

[15] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[16] 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. 7

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1

[18] 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*, pages 4971–4980, 2024. 3, 7

[19] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1

[20] 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

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 6

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[23] 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, 6, 1

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 5

[25] 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. 2, 6, 7, 8

[26] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 2, 6

[27] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017. 3

[28] 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. 2, 6

[29] 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. 7

[30] 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, 6

[31] 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. 1, 2, 3, 5, 6

[32] 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. 1

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[35] 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. 2, 3, 7, 1

[36] 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. 2, 3, 7

[37] 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, 6

[38] 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. 2, 7

[39] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022. 2

[40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1

[41] 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. 2

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[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. 2, 6

[44] 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

[45] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11): e1002686, 2018. 1

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[47] 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. 2, 6

[48] 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. 2, 7, 8

[49] Ruizhi Wang, Xiangtao Wang, Jie Zhou, Thomas Lukasiewicz, and Zhenghua Xu. Cˆ 2m-dot: Cross-modal consistent multi-view medical report generation with domain transfer network. *arXiv preprint arXiv:2310.05355*, 2023. 2, 3

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

[51] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023. 2

[52] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. 2, 3

[53] 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. 2, 7, 8

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

[55] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–6, 2023. 2