BaMCo: Balanced Multimodal Contrastive Learning for Knowledge-Driven Medical VQA

Ziya Ata Yazıcı[1] (✉) and Hazım Kemal Ekenel[1,2]

[1] Istanbul Technical University, Dept. of Computer Engineering, Istanbul, Türkiye
yaziciz21@itu.edu.tr
[2] NYU Abu Dhabi, Division of Engineering, Abu Dhabi, UAE

Abstract. Medical Visual Question Answering enables large language models to answer questions related to clinical images. While domain-specific LLMs are capable of strong reasoning, their development can be costly. In contrast, general-purpose models are more efficient, but often lack deep understanding. Previous research has shown that integrating external knowledge enhances the performance of general-purpose LLMs, particularly for questions that involve complex medical terminology. To improve the utilization of external knowledge, we introduce a novel multimodal knowledge space pretraining method trained with the proposed Balanced Multimodal Contrastive Learning Loss. Our approach optimizes knowledge spaces through balanced contrastive learning across modalities, together with the auxiliary classification task. Additionally, we developed a novel framework to improve knowledge-driven Medical VQA for LLMs by integrating the pretrained knowledge space. Experiments on the Slake, VQA-RAD, and PathVQA datasets demonstrate that our approach outperforms state-of-the-art Medical VQA methods with an average accuracy of 85.8%, 76.7%, and 60.0%, respectively. The source code is available at https://github.com/yaziciz/BaMCo.

Keywords: Medical VQA, Multimodal LLMs, Knowledge Space

1 Introduction

The Medical Visual Question Answering (Medical VQA) task focuses on accurately responding to a question related to a medical image [8]. Existing works employ visual encoders pretrained in the medical domain and a large language model (LLM) that merges visual features with question tokens to generate an answer [20,13]. These approaches typically depend on large pretrained LLMs in the medical field, which can effectively respond to questions incorporating domain-specific terminology. However, the high computational costs of developing these models have led researchers to enhance the visual question answering capabilities of general-purpose LLMs [25,28]. Generally, two mainstream approaches have been studied to incorporate domain knowledge: Contrastive learning [28,10,30] and retrieval-augmented generation (RAG) [21,23]. The first approach involves creating a knowledge space by grouping similar features, whether uni- or multimodal, while separating those from different classes. The second approach uses

user queries to extract relevant information from a database to generate comprehensive answers. However, both methods have their own limitations. The typical approach to contrastive learning primarily emphasizes textual data, meaning that visual aspects of the knowledge are often underexplored [25,28]. Similarly, the multimodal approaches do not extensively represent the inter-modality relations, and introduce the visual features during the VQA task only, leading to limited representation of knowledge [11,10,30]. In contrast, RAG-based methods for visual question answering depend on comprehensive knowledge databases. The effectiveness of these methods is heavily influenced by the database's coverage and the efficiency of the retrieval systems, which increases the model's computational cost [23].

To address the above limitations, this study proposes a knowledge-driven Medical VQA framework (Fig. 1) in conjunction with a multimodal knowledge space pretraining approach (Fig. 2). The proposed **Ba**lanced **M**ultimodal **Co**ntrastive (**BaMCo**) Loss focuses on optimizing a latent space by integrating visual and textual features, specifically using images associated with terminologies along with their textual definitions and relationships. We utilize a multimodal knowledge space to overcome limitations in modality representation and enhance optimization through inter-modality connections. We also implement a long-tailed classification task using cross-entropy loss, guided by the features extracted from the reference and intra-class images. As illustrated in Fig. 1, we leverage the pretrained *knowledge encoder* and *intra-class image encoder* to incorporate supplementary knowledge from both questions and images for knowledge-driven Medical VQA. This approach allows the Vision Language Models (VLMs) to take advantage of the pretrained multimodal latent space without the need for large-scale databases.

Our contributions are: **(1)** We propose a new method for knowledge source generation on Slake [14], VQA-RAD [12], and PathVQA [9] datasets for radiology and pathology. **(2)** We introduce a novel multimodal knowledge encoder pretraining technique through our novel balanced multimodal contrastive learning method. **(3)** We extensively experiment with the zero-shot and few-shot matching performance of the proposed knowledge encoder. **(4)** We show that our knowledge-driven Medical VQA approach outperforms the state-of-the-art methods.

2 Methods

Our primary goal is to enhance the visual question answering performance of multimodal LLMs by leveraging knowledge embeddings from our pretrained multimodal space. To achieve this, we develop a knowledge source tailored to the separate datasets and pretrain embedding spaces. This is achieved by integrating reference images and intra-class image features, along with textual features that include terms, definitions, and their relationships with other terms.

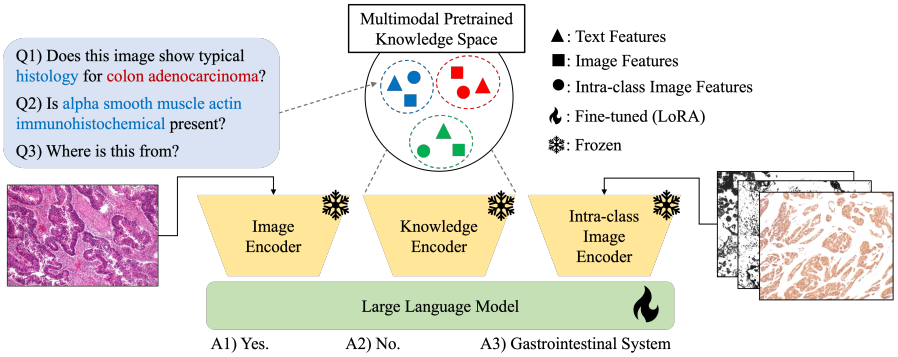

Fig. 1. Overview of the proposed framework for Medical VQA. The model integrates an image encoder, a knowledge encoder, pretrained with BaMCo, and an intra-class image encoder to enhance knowledge-driven question answering performance in medical imaging domains.

Table 1. The statistics of the created knowledge sources for each dataset. Each column represents the unique number of items in the knowledge set.

Datasets	# Images	# Head Entities	# Tail Entities	# Relations	All
Slake [14]	383	28	164	24	8387
VQA-RAD [12]	534	285	1672	73	4008
PathVQA [9]	3687	1450	8438	113	64599

Knowledge Source. In our work, we experiment with three different datasets: Slake [14], VQA-RAD [12], and PathVQA [9], which contain images, questions, and answers related to the fields of radiology and pathology. To gather knowledge items from these datasets, we employ a structured approach to extract domain-specific information from the training sets. For each question/answer couple, if the question is closed-ended and the answer is "yes", we use ScispaCy [15], a parser tool specialized in biomedical entities, to extract relevant domain terms from the question. In cases of open-ended questions, we retrieve the entities from the answers instead. Then, for each identified entity, we leverage the Unified Medical Language System (UMLS) [2] to obtain term definitions and establish relationships between terms, which are already registered in the system. Consequently, we created knowledge items in the format of "*Head Entity [Relation] Tail Entity*". Finally, we group the images according to the head terms included in their respective textual samples. After de-duplication and noise reduction, we obtained 8387, 4008, and 64599 items for Slake, VQA-RAD, and PathVQA, respectively. The unique head entities, tail entities, and relations are summarized in Table 1. The "All" column indicates the total count of knowledge items derived from these unique samples.

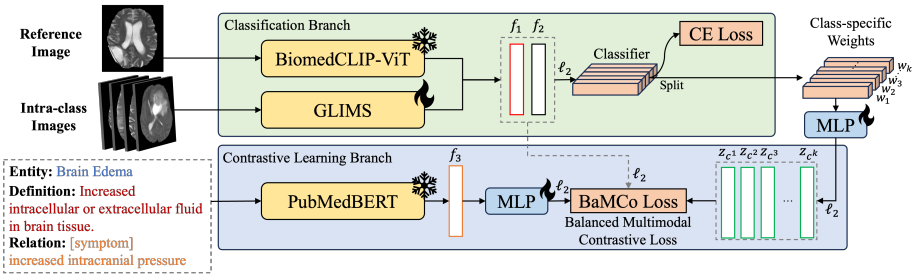

Fig. 2. The Balanced Multimodal Contrastive Learning (BaMCo) framework. BaMCo features a classification branch that employs reference and intra-class images for class-specific weight retrieval. Simultaneously, its contrastive learning branch uses image and textual input embeddings and incorporates class-specific weights from the classification branch to establish a balanced multimodal contrastive loss.

Balanced Multimodal Knowledge Space Pretraining. By utilizing the collected knowledge sources, we pretrain distinct knowledge spaces that incorporate multimodal encoders. In our approach, given a collection of images, entities, definitions, and relationships, we aim to optimize the knowledge space by converging the positive pair modalities, while diverging the negative pairs.

Fig. 2 illustrates the proposed framework's pipeline. This approach consists of two branches: Classification and contrastive learning, which are trained simultaneously. For the classification branch, we process a reference image using the Vision Transformer (ViT) [5] model from the frozen BiomedCLIP [27], known for its strong performance in image-text matching within the medical domain, to produce the feature f_1. Additionally, we randomly sample and concatenate intra-class images to retrieve the common features, f_2, which guide the knowledge space. To process the concatenated intra-images, we employ GLIMS [24], a 3D multi-scale hybrid image encoder for its effective hybrid feature extraction capabilities. Subsequently, f_1 and f_2 are concatenated and fed into a classifier that performs cross-entropy logit compensation via the cross-entropy loss, $\mathcal{L}_{CE}$. The image distribution for the unique entities is observed to represent a long-tailed distribution. To address the dataset imbalance in pretraining, this step aims to remove bias by adjusting the decision boundaries [31].

$$\mathcal{L}_{BaMCo} = -\sum_{i \in B_y} \frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_p \cdot z_i)}{\sum_{j \in Y} \frac{1}{|B_j|} \sum_{k \in B_j} \exp(z_k \cdot z_i)} \tag{1}$$

For the contrastive learning branch, we utilized the frozen PubMedBERT [7] model, which serves as a text encoder optimized within BiomedCLIP. In every iteration, we randomly choose among head entities from the questions, their definitions, and the relations between the head and tail entities to produce the output f_3. Subsequently, the textual features are fed into a multi-layer perceptron (MLP) layer. Following the work [31], we utilize class prototypes since they are

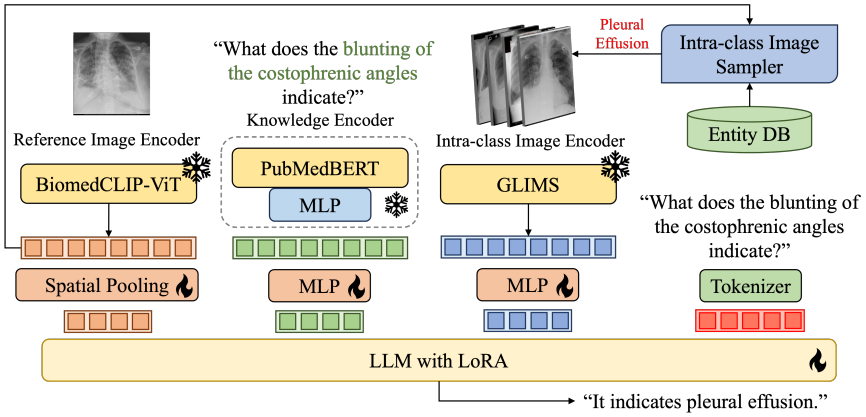

Fig. 3. The architecture of the proposed knowledge-driven Medical VQA model. The model takes in an image, knowledge, and intra-class image embeddings, which are then combined with question tokens to generate an answer. When a reference image is provided, the model identifies the most similar term from the knowledge source and randomly samples intra-class images associated with the selected entity.

shown to be also co-linear with the weights of the linear classifier to which the classes collapse. This enables the retrieval of class-specific weights, $w_1, w_2, ..., w_k$. Consequently, they are processed with an MLP layer to retrieve the class-specific centers, $z_{c^1}, z_{c^2}, ..., z_{c^k}$. Finally, the class-specific embeddings and the multimodal features are combined to compute the BaMCo Loss, denoted as $\mathcal{L}_{BaMCo}$, as illustrated in Eq. 1, where B_y denotes the subset of B that includes all samples of class y, $|\cdot|$ indicates the number of samples in a specific set, and z refers to multimodal encoded features.

Overall, the loss function $\mathcal{L} = \lambda\mathcal{L}_{CE} + \mu\mathcal{L}_{BaMCo}$ is used for training the knowledge space with $\lambda = \mu = 0.5$. This approach optimizes the multimodal features of similar entities, enabling them to cluster closely while balancing the contributions of long-tailed samples during contrastive learning.

Knowledge-driven Medical VQA. Our proposed model, shown in Fig. 3, uses an encoder-decoder architecture with three branches for processing image, knowledge, and intra-class image modalities. We experimented with the Llama 3.2 (1B/3B) [19] and GPT-2 XL (1.5B) [18] models for generating answers. To encode images, we use the pretrained ViT model from BiomedCLIP to extract visual features $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{n\times\ell}$, where ℓ is the length of the embeddings. To leverage the pretrained knowledge space, we use PubMedBERT and the MLP model to encode domain terms into embeddings $K = \{k_1, k_2, ..., k_n\} \in \mathbb{R}^{n\times\ell}$ for the LLM model. Additionally, to benefit the common features among images in the same entity class, we sample $K = 36$ intra-class images. To determine the entity class, we use image features from the pretrained ViT model (the *reference image*) and term embeddings from the pretrained PubMedBERT

model to find the entity i^* with the highest cosine similarity. Finally, by utilizing the GLIMS model, we extract intra-class features $V = \{v_1, v_2, ..., v_n\} \in \mathbb{R}^{n \times \ell}$. Since the encoder outputs possess high dimensions, we utilize spatial pooling perceiver [1] for the image encoder and MLP layers for knowledge and intra-class image encoders to reduce the dimensionality.

To determine the structure of the prompt, we referred to previous work on Medical VQA [13]. Given the question tokens, we sequence the image, knowledge, and intra-class features. For example, the prompt takes the following structure: ***p=[X, K, V, What does the blunting of the costophrenic angles indicate?]***, which is then given as input to the language model.

3 Experimental Setup

Datasets. We trained and tested our proposed method using three different datasets: Slake, VQA-RAD, and PathVQA. These datasets are often cited in research for their variety of open and closed questions and answers. In our study, we utilized the official training, validation, and test splits available for each dataset. Regarding domain specifics, PathVQA contains data related to pathology, VQA-RAD focuses on radiology, and Slake includes a combination of both domains.

Implementation Details. For knowledge space pretraining, we resized the reference images to 224 × 224 pixels and intra-class images to 84 × 84 pixels to manage encoder complexity. The text encoder had a context length of 256 tokens. For image augmentation, we used AutoAugment [3], as per [31]. Both image and text embeddings were set to 512 dimensions. The BaMCo framework was trained with a batch size of 48 for 300 epochs at a learning rate of 1e-3, utilizing a cosine scheduler. For Medical VQA, we adjusted the BaMCo pretrained encoders' outputs via projection layers to match LLM token size and fine-tuned for 10 epochs using the same augmentation techniques as BaMCo pretraining. Fine-tuning of the LLMs employed the Low-Rank Adaptation (LoRA) method with $r = 64$ and $\alpha = 16$, using a learning rate of 1e-4 and a cosine scheduler. The AdamW optimizer was used for both training stages. All experiments were conducted on a single NVIDIA A100 GPU.

Evaluation Metrics. In our comparisons with existing literature, we focus primarily on the Accuracy metric, as it is the most commonly reported. Additionally, for our ablation studies, we include BLEU-1 and ROUGE-1 to assess lexical similarity, as well as the BERTScore to evaluate semantic similarity.

Table 2. Performance comparisons of existing Medical VQA models on the Slake, VQA-RAD, and PathVQA datasets. The scores are given as exact match accuracy. The best-performing models are bolded, while the second-best ones are underlined.

Models	Slake			VQA-RAD			PathVQA		
	Open Ended	Close Ended	Avg.	Open Ended	Close Ended	Avg.	Open Ended	Close Ended	Avg.
AMAM [17]	-	-	-	-	-	-	18.2	84.4	51.3
VQAMix [6]	-	-	-	-	-	-	13.4	83.5	48.5
MMQ [4]	-	-	-	-	-	-	11.8	82.1	47.0
MEVF [16]	-	-	-	-	-	-	8.1	81.4	44.8
BiomedGPT-B [26]	**89.9**	84.3	**87.1**	60.9	81.3	71.1	28.0	88.0	58.0
CLIP-ViT GPT2-XL [20]	82.1	84.3	83.2	40.0	**87.0**	63.6	-	-	-
LLAVA-Med [13]	83.2	84.2	83.7	61.4	81.3	71.3	-	-	-
LLaVA-Med++ [22]	79.3	84.0	81.7	64.6	77.0	70.8	-	-	-
MedVInT-TD-S [29]	79.7	85.1	82.4	55.3	85.4	70.4	-	-	-
MedVInT-TE-S [29]	84.0	85.1	84.6	53.6	76.5	65.1	-	-	-
VG_CALF [11]	81.4	83.8	82.6	67.0	<u>85.5</u>	76.3	-	-	-
Ours (Llama 3.2/1B)	83.1	<u>85.9</u>	84.5	62.0	83.7	72.8	<u>28.9</u>	<u>89.8</u>	<u>59.4</u>
Ours (Llama 3.2/3B)	<u>84.2</u>	**87.3**	<u>85.8</u>	<u>67.5</u>	85.3	<u>76.4</u>	**29.1**	**90.8**	**60.0**
Ours (GPT2-XL)	83.0	84.8	83.9	**68.5**	84.9	**76.7**	27.2	87.6	57.4

4 Results

Medical VQA Performance Comparison. To show the effectiveness of our proposed balanced multimodal contrastive learning method, BaMCo, on Medical VQA, we compared state-of-the-art methods and several ablated designs across three datasets. The results presented in Table 2 indicate that our model significantly improves performance on Medical VQA tasks. Our approach, using Llama 3.2 (3B) for the Slake dataset, achieved the second-best average exact match accuracy of 85.8%. This is particularly notable compared to BiomedGPT, as our LLM model does not rely on domain-specific pretrained weights. For the VQA-RAD and PathVQA datasets, our models surpass state-of-the-art performance, reaching average accuracies of 76.7% and 60.0% with multimodal knowledge-driven GPT2-XL and Llama 3.2 (3B), respectively. Moreover, the smaller 1B version of Llama 3.2 performs impressively, achieving a 59.0% accuracy for PathVQA, making it the second-best method. This demonstrates the effectiveness of the pretrained knowledge space, optimized with BaMCo. Importantly, these results are accomplished without pretraining the LLM on a large-scale, domain-specific dataset; instead, we utilize a knowledge encoder specifically optimized for image-and-text tasks pertinent to the domain.

Ablation Analysis. We also emphasized how the suggested multimodal knowledge space training affects zero-shot image-text and few-shot text-text matching. Moreover, we demonstrated the improvements in answer generation achieved by the knowledge-driven Medical VQA approach through our ablation experiments. As shown in Fig. 4, our approach, which utilizes BaMCo, outperforms

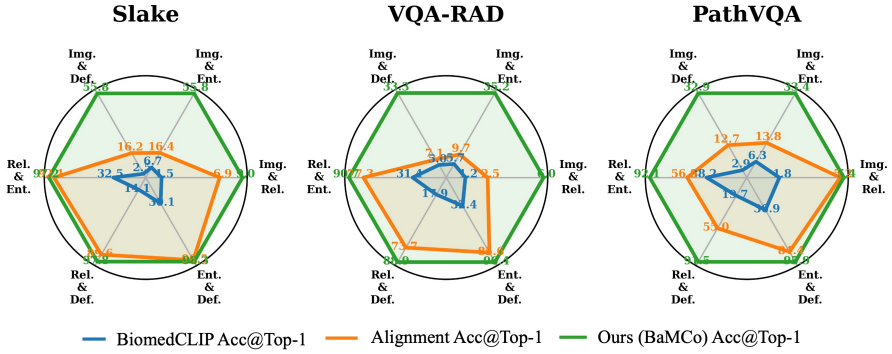

Fig. 4. The zero-shot image-text and few-shot text-text matching results of the experimented knowledge spaces. BiomedCLIP is selected as the baseline. Alignment represents the fine-tuned BiomedCLIP model. *Def.*: Definition, *Ent.*: Entity, *Rel.*:Relation.

Table 3. The averaged results of the ablation experiments. Each row displays the average results of the LLM models across all datasets and models to illustrate the overall performance contribution of the proposed approach.

Model Settings	Metrics			
	BLEU-1	**ROUGE-1**	**BERTScore**	**Avg. Accuracy**
BiomedCLIP	55.96	43.54	93.37	71.39
BiomedCLIP + Alignment	56.89 (+1.68%)	44.08 (+1.24%)	93.46 (+0.10%)	72.20 (+1.13%)
BiomedCLIP + BaMCo	**57.70 (+3.13%)**	**45.13 (+3.67%)**	**93.52 (+0.15%)**	**72.93 (+2.17%)**

the BiomedCLIP [27] model in both zero-shot and few-shot matching on the official test splits of the experimented datasets in terms of top-1 accuracy for image-text and text-text matching. Furthermore, as we utilized the knowledge encoder on the LLMs, we experienced an improvement in the open and closed-ended question answering performance as given in Table 3. In terms of overall performance across the three datasets and models, our method surpasses the baseline model using BiomedCLIP by 3.13% in BLEU-1, 3.67% in ROUGE-1, 0.15% in BERTScore, and 2.17% in accuracy.

Limitations. During our development, we identified several limitations in the optimized model. For computational reasons, we fixed the number of sampled intra-class images at $K = 36$, but this number should be adjusted based on performance outcomes. Additionally, while we utilized training sets of the selected datasets to create our knowledge source, incorporating additional datasets could improve the Medical VQA performance by introducing diverse domain knowledge.

5 Conclusion

In this study, we introduced BaMCo, a Balanced Multimodal Contrastive Learning approach for knowledge-driven Medical VQA, designed to enhance the question answering capabilities of LLMs through a novel multimodal knowledge space pretraining strategy. By integrating balanced contrastive learning across image-image, image-text, and text-text relationships alongside a long-tailed classification task, our method effectively captures medical knowledge and significantly improves the ability to answer complex medical questions. Additionally, incorporating intra-class image features strengthens the alignment between anatomical landmarks and medical terms. Experimental results on Slake, VQA-RAD, and PathVQA datasets demonstrate that our approach outperforms existing state-of-the-art models in the visual question answering task.

Acknowledgments. Computing resources used in this work were provided by the National Center for High Performance Computing of Türkiye (UHeM) under grant number 4021942025. This study was partially funded by the European Union's Horizon Europe Research and Innovation Program (Grant No. 101135798).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. arXiv preprint arXiv:2404.00578 (2024)
2. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research **32**, D267–D270 (2004)
3. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: AutoAugment: Learning Augmentation Strategies From Data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
4. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple Meta-Model Quantifying for Medical Visual Question Answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 64–74 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021)
6. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: VQAMIX: Conditional Triplet Mixup for Medical Visual Question Answering. IEEE Transactions on Medical Imaging **41**(11), 3332–3343 (2022)
7. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

8. Hartsock, I., Rasool, G.: Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review. Frontiers in Artificial Intelligence **7**, 1430984 (2024)
9. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ Questions for Medical Visual Question Answering. arXiv preprint arXiv:2003.10286 (2020)
10. Hu, X., Gu, L., Kobayashi, K., Liu, L., Zhang, M., Harada, T., Summers, R.M., Zhu, Y.: Interpretable Medical Image Visual Question Answering via Multi-Modal Relationship Graph Learning. Medical Image Analysis **97**, 103279 (2024)
11. Lameesa, A., Silpasuwanchai, C., Alam, M.S.B.: VG-CALF: A Vision-Guided Cross-Attention and Late-Fusion Network for Radiology Images in Medical Visual Question Answering. Neurocomputing **613**, 128730 (2025)
12. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A Dataset of Clinically Generated Visual Questions and Answers about Radiology Images. Scientific Data **5**(1), 1–10 (2018)
13. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
14. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical Visual Question Answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654 (2021)
15. Neumann, M., King, D., Beltagy, I., Ammar, W.: SciSpaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 319–327. Association for Computational Linguistics (2019)
16. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming Data Limitation in Medical Visual Question Answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 522–530 (2019)
17. Pan, H., He, S., Zhang, K., Qu, B., Chen, C., Shi, K.: Amam: An Attention-based Multimodal Alignment Model for Medical Visual Question Answering. Knowledge-Based Systems **255**, 109763 (2022)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. OpenAI Blog **1**(8), 9 (2019)
19. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971 (2023)
20. Van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended Medical Visual Question Answering through Prefix Tuning of Language Models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 726–736 (2023)
21. Xia, P., Zhu, K., Li, H., Zhu, H., Li, Y., Li, G., Zhang, L., Yao, H.: RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1093 (2024)
22. Xie, Y., et al.: MedTrinity-25M: A Large-Scale Multimodal Dataset with Multigranular Annotations for Medicine. arXiv preprint arXiv:2408.02900 (2024)
23. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking Retrieval-Augmented Generation for Medicine. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233–6251. Association for Computational Linguistics (2024)

24. Yazıcı, Z.A., Öksüz, I., Ekenel, H.K.: GLIMS: Attention-Guided Lightweight Multi-Scale Hybrid Network for Volumetric Semantic Segmentation. Image and Vision Computing **146**, 105055 (2024)
25. Zafar, A., Sahoo, S.K., Bhardawaj, H., Das, A., Ekbal, A.: KI-MAG: A Knowledge-Infused Abstractive Question Answering System in Medical Domain. Neurocomputing **571**, 127141 (2024)
26. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A Generalist Vision–Language Foundation Model for Diverse Biomedical Tasks. Nature Medicine pp. 1–13 (2024)
27. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Wang, S., Poon, H.: A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image-Text Pairs. NEJM AI **2**(1), AIoa2400640 (2025)
28. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-Enhanced Visual-Language Pre-Training on Chest Radiology Images. Nature Communications **14**(1) (2023)
29. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv preprint arXiv:2305.10415 (2023)
30. Zhou, X., Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-Enhanced Visual-Language Pretraining for Computational Pathology. In: Computer Vision – ECCV 2024. p. 345–362 (2024)
31. Zhu, J., Wang, Z., Chen, J., Chen, Y.P.P., Jiang, Y.G.: Balanced Contrastive Learning for Long-Tailed Visual Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6908–6917 (2022)