BioD2C: A Dual-level Semantic Consistency Constraint Framework for Biomedical VQA

Zhengyang Ji[1,2], Shang Gao[2], Li Liu[2], Yifan Jia[1,2], and Yutao Yue[2,3†]

[1] Shandong University, Qingdao, China
[2] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
yutaoyue@hkust-gz.edu.cn
[3] Institute of Deep Perception Technology, JITRI, Wuxi, China

Abstract. Biomedical visual question answering (VQA) has been widely studied and has demonstrated significant application value and potential in fields such as assistive medical diagnosis. Despite their success, current biomedical VQA models perform multimodal information interaction only at the model level within large language models (LLMs), leading to suboptimal multimodal semantic alignment when dealing with complex tasks. To address this issue, we propose **BioD2C**: a novel **D**ual-level Semantic **C**onsistency **C**onstraint Framework for **Bio**medical VQA, which achieves dual-level semantic interaction alignment at both the model and feature levels, enabling the model to adaptively learn visual features based on the question. Specifically, we firstly integrate textual features into visual features via an image-text fusion mechanism as feature-level semantic interaction, obtaining visual features conditioned on the given text; and then introduce a text-queue-based cross-modal soft semantic loss function to further align the image semantics with the question semantics. Specifically, in this work, we establish a new dataset, BioVGQ, to address inherent biases in prior datasets by filtering manually-altered images and aligning question-answer pairs with multimodal context, and train our model on this dataset. Extensive experimental results demonstrate that BioD2C achieves state-of-the-art (SOTA) performance across multiple downstream datasets, showcasing its robustness, generalizability, and potential to advance biomedical VQA research. The source code of this work and the BioVGQ dataset can be accessed through code and dataset.

Keywords: Biomedical VQA · Dual Interaction · Semantic Alignment

1 Introduction

Biomedical VQA aims to design and develop systems capable of understanding biomedical images and generating relevant answers based on given textual instructions.

† Corresponding author.

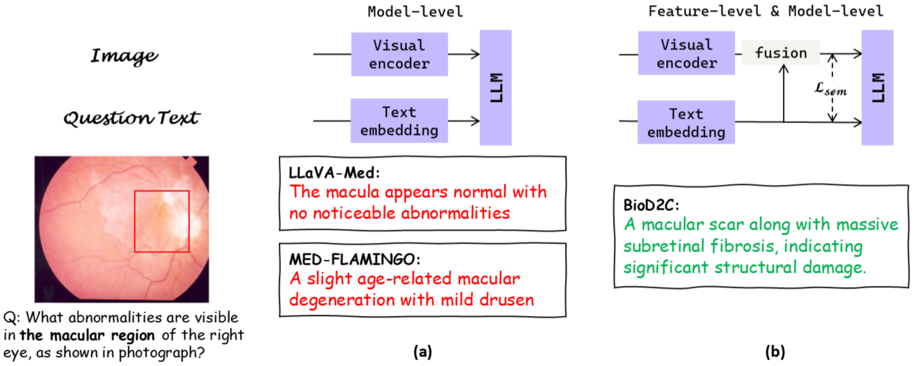


Image
Question Text
Q: What abnormalities are visible in the macular region of the right eye, as shown in photograph?
Model-level
Visual encoder
Text embedding
LLM
LLaVA-Med:
The macula appears normal with no noticeable abnormalities
MED-FLAMINGO:
A slight age-related macular degeneration with mild drusen
(a)
Feature-level & Model-level
Visual encoder
fusion
$\mathcal{L}_{sem}$
Text embedding
LLM
BioD2C:
A macular scar along with massive subretinal fibrosis, indicating significant structural damage.
(b)

Fig. 1: (a) and (b) illustrate the performance of the model-level interaction framework and BioD2C under image-related questions, respectively. Red text represents incorrect answers, while green text represents correct answers.

In real clinical scenarios, question texts often refer to specific elements within an image. Therefore, achieving optimal semantic alignment between the image and the text instruction, i.e., the model should focus on the image regions corresponding to the textual query, becomes the key to the success of biomedical question answering models. However, existing biomedical models [26,12,18,24] extract visual and textual features independently using separate visual encoders and text embedding layers, relying solely on LLMs for model-level multimodal semantic interaction, and lacking semantic alignment at the feature level.

To address these challenges, we propose BioD2C, a novel dual-level semantic consistency constraint framework for biomedical VQA, as illustrated in Fig. 2. Compared to existing VQA models, BioD2C employs a novel image-text fusion mechanism for feature-level multimodal semantic interaction after extracting image and text features, obtaining image features conditioned on the given text. Furthermore, we introduce a text queue mechanism to project image and text semantics from high-dimensional vector spaces into corresponding probability distributions. By minimizing the divergence between these distributions, we achieve quantized alignment of cross-modal semantic representations. Fig. 1 illustrates this behavior. When faced with complex questions, both baseline models rely solely on semantic interaction within LLMs, resulting in biased answers, while BioD2C benefits from semantic alignment at the feature level and produces the correct answer.

Due to the scarcity of real biomedical data, existing biomedical vision-language datasets such as PMC-OA [15] and PMC-VQA [26] rely on biomedical papers publicly available from the PubMedCentral (PMC)'s OpenAccess subset [21], some of which contain images that differ significantly from real-world biomedical images. Additionally, when generating question-answer pairs, existing biomedical visual question-answering datasets often rely solely on image captions or captions supplemented with visual information, which may cause a misalignment between the generated question-answer pairs and the information contained in the images. As a result, visual question-answering models trained on these datasets may have an inherent limitation in understanding the images.

To this end, we establish a new dataset, BioVGQ, to address the issues present in existing datasets. It is based on the existing PMC-based dataset and integrates multiple public datasets, filtering out images that have undergone significant manual manipulation. When generating question-answer pairs, both the images and their corresponding captions are utilized to ensure a strong correlation between the question-answer pairs and the images.

Overall, the main contributions of our work are as follows: i) We propose BioD2C, a dual-level framework that enforces semantic consistency both through model-level interactions within LLMs and specifically through feature-level image-text fusion mechanisms, while further optimizing visual-textual alignment via a cross-modal semantic loss function. ii) BioVGQ, a biomedical VQA dataset with cleaner images that incorporate contextual information, has been established, and our model is trained on this dataset. iii) Extensive comparative and ablation experiments demonstrate the superiority of BioD2C over current SOTA biomedical VQA models in terms of performance and the effectiveness of each component.

2 The BioVGQ Dataset

Most of the image data in BioVGQ comes from PMC-VQA. To remove images with low informational content or obvious noise, we manually annotated 3,000 images. We label an image as “polluted” if it has i) more than six sub-figures, which dilutes useful information, or ii) clearly man-made content such as tables or hand sketches. All others are “clean”. Using these labeled data, we trained an image classifier to automate the classification process, ultimately obtaining 77K clean biomedical images. Specifically, we added an MLP classification head to the pre-trained image encoder of PMC-CLIP [15] to serve as the image classifier.

For generating biomedical question-answer pairs, we used the ChatGPT-4o API [10], providing both images and their corresponding captions to ensure that the generated pairs accurately reflect the image content without deviation. Further generation details and prompts are as follows:

> You are an AI assistant specialized in biomedical topics. Generate 2-3 clinically meaningful open-ended question-and-answer pairs based on the provided medical image and caption.
> Requirements: - Each question must be a single, clear sentence; each answer should directly address it. - Cover overall understanding and specific details, without copying the caption.

> - Answers must require examining the image, not just medical background knowledge. - Ensure clinical relevance, professionalism, and conciseness.
> - Format: ["question": "xxx", "answer": "xxx", ...]
> caption: {caption}, image-url: {image-url}

To enrich the dataset, we incorporated various modalities of biomedical images, closed-ended questions, and short dialogues, integrating the training splits of SLAKE [16], Path-VQA [8], and RAD-VQA [11]. As a result, BioVGQ comprises 81K medical images and 188K question-answer pairs.

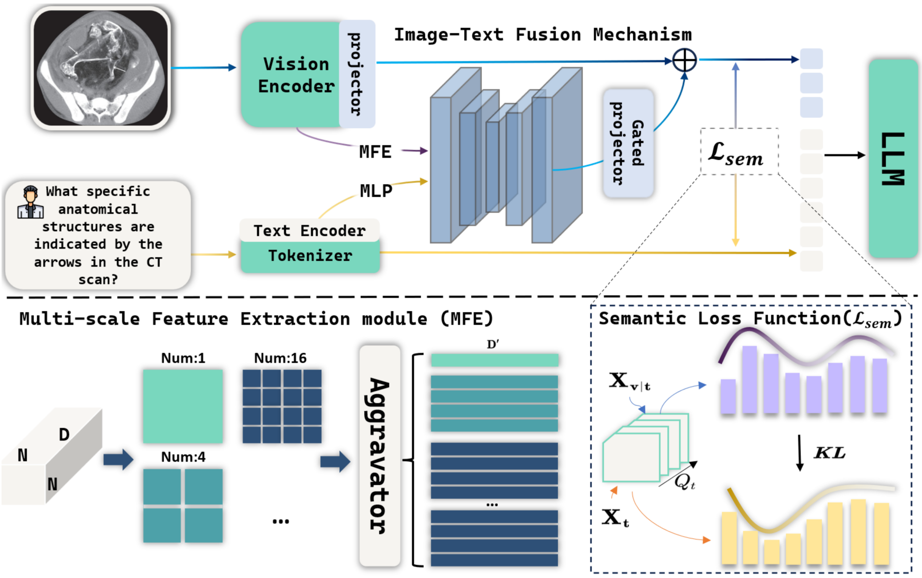

Vision Encoder
projector
Image-Text Fusion Mechanism
Gated projector
MFE
MLP
$\mathcal{L}_{sem}$
LLM
What specific anatomical structures are indicated by the arrows in the CT scan?
Text Encoder
Tokenizer
Multi-scale Feature Extraction module (MFE)
N
D
N
Num:1
Num:16
Num:4
...
Aggravator
D'
...
Semantic Loss Function($\mathcal{L}_{sem}$)
$\mathbf{X}_{v|t}$
Q_t
$\mathbf{X}_t$
KL

Fig. 2: **BioD2C Architecture.** *Feature-level Interaction:* Medical images and text questions are encoded into features X_v and X_t. A multi-scale enhanced X_v is fused with X_t via a Transformer decoder, generating X_{vt}, which is then combined with X_v through a gating mechanism to produce text-conditioned features $X_{v|t}$. *Semantic Loss:* A text-queue loss guides $X_{v|t}$ to align with X_t.

3 Method

3.1 Feature Level Semantic Interaction Mechanism

In this section, we will introduce the technical details of feature-level semantic interaction through the image-text fusion mechanism.

Text Encoding. Before fusing textual features with image features, we preprocess the text in two steps. First, the tokenized text sequence $\mathcal{T}$ is encoded into a meaningful representation $X_t^{'}$, defined as $X_t^{'} = \varepsilon\left(\mathcal{T}\right)$, where ε represents the encoding function. Here, we directly use the text embedding layer of the LLM as the encoding function ε. Next, we use a MLP to map $X_t^{'}$ into the image feature space, resulting in the final text representation X_t. The overall text encoding process is expressed as: $X_t = MLP\left(\varepsilon\left(\mathcal{T}\right)\right)$.

Image Processing. The visual features output by the vision encoder, $X_v \in \mathbb{R}^{N \times N \times D}$, contain only a single level of granularity, where N represents the number of image patches and D denotes the feature dimension. To extract multi-granularity image features, we introduce a multi-scale feature extraction module (MFE) that utilizes a divide-and-aggregate strategy, as illustrated in Fig. 2. Specifically, MFE consists of S different scales. At each scale $s \in \{1, 2, ..., S\}$, the feature map is divided into 4^{s-1} blocks, resulting in a total of $M = \sum_{s=1}^{S} 4^{s-1}$ blocks across all scales. Each 3D feature block undergoes global pooling to produce a 1D feature. For a block $f_{s,t}$ at scale s, where $t \in \left\{1, 2, ..., 4^{s-1}\right\}$ indicates the block index, the global pooling is defined as $f_{s,t}^{'} = maxpool\left(f_{s,t}\right) + avgpool\left(f_{s,t}\right)$, where $maxpool$ and $avgpool$ represent global max pooling and global average pooling, respectively. Finally, the pooled features $f_{s,t}^{'}$ from all scales are concatenated to form the multi-scale image feature $X_v^{'} \in \mathbb{R}^{M \times D}$. In the implementation, the number of scales is set to $S = 6$.

Image-Text Fusion. A 12-layer Transformer decoder achieves cross-modal fusion by treating text encoding X_t as query and multi-scale image features $X_v^{'}$ as key and value, producing the text-contextualized image representation X_{vt}:

$$X_{vt} = Fusion(X_t, X_v', X_v'). \tag{1}$$

To compensate for the potential loss of original image features during modality fusion and achieve complementary information across modalities, we introduce a learnable gating mechanism [1,3,6,9]. This mechanism combines the original image features X_v with the fused features X_{vt} by processing them through a projection layer and an additional projection layer, resulting in the conditioned image features $X_{v|t}$. The gating mechanism ensures a gradual fusion of modality features, avoiding significant feature alteration and overall performance degradation [5]. It is implemented by multiplying the output of the additional projection layer with $tanh\left(\beta\right)$, where β is a learnable parameter initialized to a small positive value. We initialize β to 0.2 to balance initial feature bias and fusion effectiveness. Mathematically, the fusion module is implemented as:

$$X_{v|t} = Proj\left(X_v\right) + Proj_g\left(X_{vt}\right) \cdot tanh\left(\beta\right), \tag{2}$$

then $X_{v|t}$ is input into the LLM together with the text token sequence $\mathcal{T}$.

3.2 Text-queue-based Cross-modal Semantic Loss function

Through the above procedure, we obtain the visual features conditioned on the text, but lack an optimization objective to guide the model toward optimal multimodal semantic alignment at the feature level. Inspired by ALBEF [13] and MoCo [7], we propose a text-queue-based cross-modal semantic loss function, which applies a soft constraint to align visual semantics with text semantics. The core idea is to map the semantics from the high-dimensional vector space to a probability distribution through similarity computation. Specifically, we extract k text samples semantically related to the image either from its corresponding caption or by retrieving from an existing knowledge base, where we set k to 30. These texts undergo the text encoding process in Section 3.1 to obtain the text queue $\mathcal{Q}_t = \{t_i\}_{i=1}^{k}$, where t_i denotes the i_{th} text sample. By calculating the cosine similarity between $X_{v|t}$, X_t, and the elements in $\mathcal{Q}_t$, we derive the semantic distributions of fused image and text features, denoted as $p\left(v\right)$ and $p\left(t\right)$, respectively. The semantic distribution $p\left(v\right)$ is computed as:

$$p\left(v\right) = \left\{ \frac{\exp\left(\left\langle X_{v|t}\,, t_i\right\rangle/\tau\right)}{\sum_{j=1}^{k} \exp\left(\left\langle X_{v|t}\,, t_j\right\rangle/\tau\right)} \right\}_{i=1}^{k}, \tag{3}$$

where $\langle\cdot,\cdot\rangle$ represents the calculation of cosine similarity, and τ is the temperature coefficient. Similarly, $p\left(t\right)$ can be computed. Using $p\left(v\right)$ and $p\left(t\right)$, we minimize the Kullback-Leibler (KL) divergence between these two distributions to align the semantics of image and text features, expressed as:

$$\mathcal{L}_{sem} = D_{KL}\left(p\left(v\right)||p\left(t\right)\right), \tag{4}$$

where $\mathcal{L}_{sem}$ represents the semantic loss between images and text, and D_{KL} denotes the KL divergence. During training, the semantic loss $\mathcal{L}_{sem}$ is combined with the commonly used sequence negative log-likelihood loss $\mathcal{L}_{nll}$ [22] to jointly optimize the model for the best performance. The final loss function used for training the model is defined as:

$$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{sem} + \mathcal{L}_{nll}, \tag{5}$$

where λ is a hyperparameter that controls the weight of the semantic loss.

4 Experiments

4.1 Implementation Details

In this work, we employ a two-stage training strategy to train our model, enabling it to adapt to biomedical VQA tasks. **Stage 1:** Projectors are independently trained to align visual features with language embeddings using 467k image-caption pairs from the LLaVA-Med dataset [12]. During this stage, $\lambda = 0$ in Eq. 5, disabling semantic loss due to limited textual diversity. **Stage 2:** The LORA adapters are fine-tuned on the BioVGQ dataset to improve BioD2C's multimodal understanding, with $\lambda = 1$ in Eq. 5, fully incorporating semantic loss to optimize performance.

We train our models using the AdamW [17] optimizer. To accelerate training, we employ the Deepspeed strategy along with Automatic Mixed Precision

Dataset	Metric	BioMedGPT	LLaVA-Med-1.5	MedVInT-TD	RadFM	BiMediX2-8B	BioD2C
SLAKE	closed ACC ↑	0.248	0.536	0.498	0.752	**0.831**	0.763
	opened ACC ↑	0.259	0.334	0.338	0.725	0.729	**0.742**
	BLEU-1 ↑	0.175	0.002	0.213	0.746	**0.778**	0.766
	ROUGE-1 ↑	0.26	0.413	0.351	0.695	0.786	**0.810**
RAD-VQA	closed ACC ↑	0.545	0.547	0.475	0.577	0.725	**0.734**
	opened ACC ↑	0.14	0.276	0.195	**0.335**	0.305	0.310
	BLEU-1 ↑	0.033	0.021	0.125	0.475	**0.552**	0.520
	ROUGE-1 ↑	0.372	0.342	0.235	0.438	0.565	**0.588**
Path-VQA	closed ACC ↑	0.512	0.621	0.454	0.505	0.872	**0.918**
	opened ACC ↑	0.053	0.036	0.022	0.005	0.282	**0.291**
	BLEU-1 ↑	0.021	0.011	0.013	0.257	0.587	**0.620**
	ROUGE-1 ↑	0.287	0.116	0.034	0.221	0.593	**0.628**
Average		0.242	0.271	0.246	0.453	0.616	**0.641**

Table 1: Comparison of performance with SOTA models on different benchmarks. The best performance is highlighted in **bold**, while the second-best is underlined.

(AMP) [4] and gradient checkpointing. Set the learning rates for the first and second training stages to $\{5e-5,\ 2e-5\}$, and train for 1 epoch and 5 epochs, respectively. For more details on hyperparameter settings, please refer to the BioD2C GitHub page. All models are implemented in PyTorch and trained on four NVIDIA 4090 GPUs with 24 GB of memory each. In terms of model construction, PMC-CLIP and PMC-LLaMA [23] are selected as the visual encoder and LLM, respectively.

4.2 Datasets and Metrics

The BioVGQ dataset is split into training, validation, and testing sets in an 8:1:1 ratio for model training and evaluation. To validate the effectiveness of BioD2C, we evaluate it on SLAKE, Path-VQA, and RAD-VQA datasets. For ablation studies and visualization analyses, we primarily use the BioVGQ test set to examine the effectiveness of different modules.

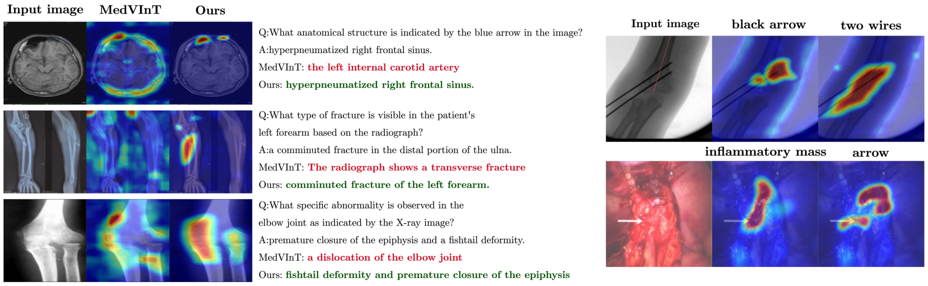
We use closed-ended question accuracy(ACC), open-ended question ACC, BLEU-1 score [20], and ROUGE-1 score [14] to comprehensively evaluate the model's performance on downstream datasets. Additionally, as BioVGQ primarily contains long-text answers, ChatGPT4 [2] is employed to evaluate the reasonableness, accuracy, and similarity of the model's responses compared to the ground truth. A comprehensive score ranging from 0 to 10 is provided, with higher scores indicating better model performance.

4.3 Comparison with SOTAs

We compare the proposed BioD2C with SOTA models in the biomedical visual question answering domain, including BioMedGPT [25], LLaVA-Med-1.5, MedVInT [26], RadFM [24], and BiMediX2-8B [19]. The results are shown in Table 1. BioD2C outperforms current SOTA models on most metrics, achieving the highest average score of 0.641, which is a 4.06% improvement over the second-best model, BiMediX2-8B.

	BLEU-1	ROUGE-1	GPT score/10	Average
BioD2C $_{w/o\ \mathcal{L}_{sem}}$	0.408	0.443	0.623	0.491
BioD2C $_{w/o\ fm}$	0.371	0.428	0.591	0.463
BioD2C $_{w/o\ BioVGQ}$	0.324	0.332	0.483	0.380
BioD2C	0.427	0.494	0.649	0.523

Table 2: The performance of BioD2C and its variants on the BioVGQ test set.



Q:What anatomical structure is indicated by the blue arrow in the image?
A:hyperpneumatized right frontal sinus.
MedVInT: **the left internal carotid artery**
Ours: **hyperpneumatized right frontal sinus.**

Q:What type of fracture is visible in the patient's left forearm based on the radiograph?
A:a comminuted fracture in the distal portion of the ulna.
MedVInT: **The radiograph shows a transverse fracture**
Ours: **comminuted fracture of the left forearm.**

Q:What specific abnormality is observed in the elbow joint as indicated by the X-ray image?
A:premature closure of the epiphysis and a fishtail deformity.
MedVInT: **a dislocation of the elbow joint**
Ours: **fishtail deformity and premature closure of the epiphysis**

Fig. 3: Visualization of the attention map of the input image.

4.4 Ablation Study

In this section, we conduct ablation studies to analyze the impact of different model configurations on its performance in biomedical question answering tasks. Specifically, we investigate the following three scenarios, i) $w/o\ \mathcal{L}_{sem}$: not using semantic loss, ii) $w/o\ fm$: directly using the visual encoder's output as the image vector for LLMs multimodal input without the fusion mechanism, and iii) $w/o\ BioVGQ$: using the PMC-VQA dataset instead of BioVGQ in the second training stage. The results of the ablation studies are shown in Table 2.

4.5 Visualization Analysis

In this section, by visualizing the attention maps, we show how the model focuses on specific regions of the image based on textual instructions. The left figure in Fig. 3 compares the performance of BioD2C and the baseline model. While both understand abstract image concepts, BioD2C accurately focuses on image regions for correct answers, while the baseline model's answers deviate due to multimodal alignment issues. The right figure illustrates the dynamic nature of the model's attention: as the text prompt changes, the attention shifts to different image regions. For example, when the question mentions "**black arrow**", the model's attention focuses on the area near the black arrow. When the prompt changes to "**two wires**", the model shifts its attention to the region where the wires are located.

5 Conclusion

In this work, we propose BioD2C, a dual-level semantic interaction biomedical VQA framework, which achieves dynamic alignment of visual features to textual features at the feature level through an image-text fusion mechanism. A cross-modal semantic loss function is employed to further optimize multimodal semantic alignment at the feature level. The framework is trained on BioVGQ, a curated dataset consisting of 81K images and 188K question-answer pairs. Extensive experiments demonstrate that, compared to baselines that independently extract visual and textual features, BioD2C can dynamically focus on specific regions of the image based on the text, achieving SOTA performance. BioD2C shows strong potential for clinical decision support, with future work targeting multi-modal integration and broader medical applications.

Acknowledgments. This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008) and the Education Bureau of Guangzhou Municipality.

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

References

1. Aberdam, A., Bensaïd, D., Golts, A., Ganz, R., Nuriel, O., Tichauer, R., Mazor, S., Litman, R.: Clipter: Looking at the bigger picture in scene text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21706–21717 (2023)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems. pp. 23716–23736 (2022)
4. Feng, J., Huang, D.: Optimal gradient checkpoint search for arbitrary computation graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11433–11442 (2021)
5. Ganz, R., Kittenplon, Y., Aberdam, A., Ben Avraham, E., Nuriel, O., Mazor, S., Litman, R.: Question aware vision transformer for multimodal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13861–13871 (2024)
6. Ganz, R., Nuriel, O., Aberdam, A., Kittenplon, Y., Mazor, S., Litman, R.: Towards models that can see and read. In: 2023 IEEE/CVF International Conference on Computer Vision. pp. 21661–21671 (2023)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)

8. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation. pp. 1735–1780 (1997)
10. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
11. 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. pp. 1–10 (2018)
12. 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. pp. 28541–28564 (2023)
13. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems. pp. 9694–9705 (2021)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
15. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536 (2023)
16. 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: IEEE International Symposium on Biomedical Imaging. pp. 1650–1654 (2021)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
18. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health. pp. 353–367 (2023)
19. Mullappilly, S.S., Kurpath, M.I., Pieri, S., Alseiari, S.Y., Cholakkal, S., Aldahmani, K., Khan, F., Anwer, R., Khan, S., Baldwin, T., et al.: Bimedix2: Bio-medical expert lmm for diverse medical modalities. arXiv preprint arXiv:2412.07769 (2024)
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
21. Roberts, R.J.: Pubmed central: The genbank of the published literature. National Academy of Sciences. pp. 381–382 (2001)
22. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in neural information processing systems. pp. 1735–1780 (2014)
23. Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., Wang, Y.: Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association. pp. 1833–1843 (2024)
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463 (2023)
25. 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)

26. 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)